BAB 1. PENDAHULUAN

1.1 Latar Belakang

Indonesia menghadapi sebuah tantangan serius di sektor pertanian yang diakibatkan perubahan iklim yang serius. Untuk menghadapi permasalahan tersebut, penting bagi pemerintah dan petani untuk bekerja sama. Salah satu cara yang dapat dilakukan adalah memaksimalkan potensi pemuliaan tanaman. Pemuliaan tanaman merupakan suatu ilmu yang digunakan untuk meningkatkan sifat tanaman, baik dari segi kualitas maupun kuantitas tanaman, dengan tujuan utama untuk meningkatkan hasil produksi (Putri dkk., 2024). Menurut (Damayanti, 2021), terdapat beberapa teknik pemuliaan tanaman yang umum digunakan yakni, mutasi genetik, induksi, hibridasi, dan introduksi gen. Salah satu metode pemuliaan tanaman yang umum diterapkan untuk mendapatkan suatu varietas unggul adalah mutasi genetik. Mengutip dari (Arta Dana dkk., 2021), pada sektor pemuliaan tanaman, teknik ini bukanlah teknik yang baru dan sudah lazim digunakan sejak lama.

Mutasi genetik atau mutagenesis merupakan suatu proses ilmiah dalam perubahan permanen pada DNA tanaman, baik secara buatan maupun alami, yang bertujuan untuk menghasilkan suatu variasi genetik baru pada sifat tanaman tersebut. Terdapat dua jenis mutagenesis, yang terjadi secara alami dan juga secara buatan seperti induksi menggunakan senyama kimia, radiasi, ataupun unsur biologis lainnya (Reddy & Pinjari, 2023). Salah satu metode yang umum digunakan pada mutasi genetik benih tanaman adalah dengan menggunakan *Ethyl Methane Sulfonate* (EMS). . EMS bekerja dengan mengalkilasi basa guanin pada DNA, sehingga menyebabkan perubahan pasangan basa dari G/C menjadi A/T dan memunculkan berbagai variasi fenotipik pada morfologi, fisiologi, maupun perkembangan tanaman (Arta Dana dkk., 2021).

Popularitas EMS dalam dunia pemuliaan tanaman didukung oleh beberapa keunggulan praktis yang tidak dimiliki agen mutagen lainnya. EMS memiliki laju mutasi yang tinggi, mudah diperoleh, dan yang terpenting, tidak lagi bersifat

mutagenik setelah mengalami proses hidrolisis, sehingga lebih aman untuk ditangani dalam lingkungan laboratorium (Lestari, 2021). Kombinasi keunggulan-keunggulan tersebut menjadikan EMS sebagai alat yang sangat andal dalam upaya menghasilkan populasi mutan yang kaya keragaman genetik dalam waktu relatif singkat. Tidak mengherankan apabila penerapan mutagenesis EMS telah menghasilkan kontribusi nyata bagi dunia pertanian global, sebagaimana tercermin dari tercatatnya sekitar 3.200 varietas dari lebih dari 210 spesies tanaman, termasuk padi, jagung, gandum, barley, dan kedelai yang berhasil dikembangkan melalui teknik mutagenesis kimia (EFSA Panel on Genetically Modified Organisms (GMO) dkk., 2021).

Luasnya penerapan EMS di berbagai spesies tanaman tersebut secara langsung mendorong pertumbuhan pesat publikasi literatur ilmiah yang membahasnya. Penelitian-penelitian mengenai mutagenesis EMS mencakup topik yang sangat beragam, mulai dari optimasi dosis dan konsentrasi EMS pada spesies tertentu, identifikasi gen target yang mengalami mutasi, karakterisasi fenotipe mutan yang dihasilkan, hingga analisis mekanisme molekuler yang mendasarinya. Keragaman topik ini menghasilkan korpus literatur yang sangat heterogen, di mana informasi-informasi penting tersebar di ribuan artikel jurnal internasional dengan format, terminologi, dan struktur pelaporan yang berbeda-beda. Kondisi ini menciptakan tantangan tersendiri bagi peneliti yang ingin memperoleh gambaran komprehensif mengenai parameter-parameter mutagenesis EMS yang telah dilaporkan dalam berbagai studi sebelumnya.

Tantangan pengelolaan literatur tersebut semakin berat apabila diletakkan dalam konteks pertumbuhan publikasi ilmiah global yang tengah mengalami lonjakan luar biasa. Data dari (González-Márquez dkk., 2024), menunjukkan bahwa terdapat lebih dari 1,5 juta literatur ilmiah di bidang hayati dan biomedis yang terdaftar di PubMed, salah satu basis data literatur terbesar di dunia, dan jumlah tersebut terus bertumbuh setiap tahunnya. Laju pertumbuhan yang pesat ini menjadikan penelusuran manual atau pendekatan *cherry-picking* dari literatur yang tersebar tidak lagi memadai, baik dari segi waktu, tenaga, maupun objektivitas (S. Zhao dkk., 2021). Akibatnya, informasi penting yang sebenarnya relevan bagi

peneliti sering kali tidak dapat teridentifikasi atau membutuhkan waktu yang sangat lama untuk ditemukan.

Perkembangan *Machine Learning* (ML) dan *Natural Language Processing* (NLP) menawarkan jalur keluar yang menjanjikan dari tantangan tersebut. ML memungkinkan mesin untuk memproses dan menganalisis data dalam skala besar berdasarkan pola yang dipelajari dari data sebelumnya, tanpa perlu diprogram secara eksplisit untuk setiap skenario (Shaveta, 2023). Salah satu cabang ML yang paling relevan untuk permasalahan pengelolaan literatur adalah *text mining*, yaitu proses eksplorasi dan analisis teks dari kumpulan dokumen untuk mengidentifikasi hubungan, keterkaitan, dan pengelompokan antar informasi secara otomatis (Ompo & Pakereng, 2024). Penerapan *text mining* dan NLP pada literatur ilmiah telah terbukti mampu meningkatkan efisiensi sintesis data, memperbaiki reprodusibilitas tinjauan literatur, serta mendorong penemuan pola-pola pengetahuan baru yang sulit dilakukan secara manual (Farrell dkk., 2022).

Di antara berbagai model NLP yang telah dikembangkan, BERT (*Bidirectional Encoder Representations from Transformers*) terbukti cukup baik dan konsisten untuk tugas ekstraksi informasi dari teks ilmiah biomedis. Keunggulan ini bertumpu pada mekanisme *self-attention* dua arah yang memungkinkan model memahami konteks suatu kata dari kedua sisi secara bersamaan, sesuatu yang tidak mampu dilakukan oleh model tradisional seperti BiLSTM-CRF, SVM, maupun CRF (Ramadhan & Siswoyo, 2024). Studi komparatif menunjukkan bahwa BERT secara konsisten mengungguli arsitektur BiLSTM-CRF tradisional di berbagai tugas NER, dan varian domain-spesifiknya, BioBERT, mencatatkan peningkatan F1-Score sebesar 0,62% pada NER, 2,80% pada *relation extraction*, dan 12,24% MRR pada *question answering* dibandingkan model-model berbasis BiLSTM dan CRF yang digunakan sebelumnya. Efektivitas ini terkonfirmasi oleh beberapa penelitian terdahulu, sepert BioBERT mencapai *F1-Score* 79,98% dalam mengekstraksi asosiasi gen-penyakit, melampaui seluruh pendekatan berbasis aturan dan *machine learning* klasik pada dataset yang sama (Deng dkk., 2021). Metode BERT-LSTM juga terbukti mengungguli *random forest*, *decision tree*, dan CNN-LSTM dalam mengekstraksi asosiasi SNP dari abstrak biomedis, dengan capaian *F1-Score* 82,40%, nilai ini jauh di atas CNN-LSTM

(71,10%), *decision tree* (54,90%), dan *random forest* (32,80%) (Bokharaeian dkk., 2023). Serta arsitektur BERT berhasil mencapai *F1-Score* 86,85% pada klasifikasi interaksi regulasi transkripsi bakteri (Varela-Vega dkk., 2024). Konsistensi keunggulan ini atas model-model tradisional di berbagai tugas dan dataset menjadi dasar utama pemilihan model BERT dalam penelitian ini.

Penelitian ini bertujuan mengembangkan metode ekstraksi informasi berbasis model BERT dari literatur maupun kajian ilmiah mengenai mutagenesis EMS, dengan berfokus pada identifikasi parameter yang umum digunakan pada penelitian tersebut seperti dosis EMS yang digunakan, organisme target, gen yang mengalami mutasi, dan fenotipe. Kemampuan dalam merepresentasikan kata secara kontekstual serta hasil akurasi yang lebih baik daripada model-model NLP tradisional menjadikan peneliti memilih model BERT dalam penilitian ini. Melihat dari hasil penelitian yang dilakukan oleh (Eang & Lee, 2024), menunjukkan bahwa penggunaan *batch size* 32 dengan *learning rate* 2×10^{-5}, 3×10^{-5}, dan 5×10^{-5} menghasilkan akurasi sebesar 90,48%, 90,25%, dan 89,79% secara berurutan, tentu saja nilai tersebut dapat digolongkan sangat baik.

Hasil dari penelitian ini diharapkan dapat mengisi kesenjangan penelitian-penelitian sebelumnya serta dapat memberikan kontribusi penting seperti mengintegrasikan temuan dari berbagai sumber literatur untuk memberikan gambaran komprehensif tentang pemanfaatan EMS dalam studi biologi molekuler dan pemuliaan tanaman. Selain itu, hasilnya diharapkan mempermudah pengelolaan informasi penelitian terkait EMS serta memperkaya pemahaman akan pola penggunaannya yang umum ditemukan di berbagai kajian ilmiah. Studi ini juga dirancang sebagai referensi awal bagi pengembangan aplikasi NLP di bidang biologi dan genetika tanaman di masa depan.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, dapat dihasilkan beberapa rumusan masalah sebagai berikut:

1. Bagaimana pengembangan metode ekstraksi informasi berbasis model BERT dapat mengatasi kesulitan dalam mengidentifikasi parameter kunci (dosis EMS

yang digunakan, organisme target, gen yang mengalami mutasi, dan fenotipe) dari literatur ilmiah?
2. Bagaimana penerapan model NLP berbasis BERT dalam *text mining* dapat meningkatkan efisiensi ekstraksi informasi dari literatur ilmiah mengenai mutagenesis EMS?
3. Apa kontribusi model BERT dalam *text mining* pada literatur ilmiah tentang mutagenesis EMS?

1.3 Tujuan

Dari rumusan masalah yang sebelumnya telah dipaparkan, tujuan penelitian ini adalah sebagai berikut:

1. Mengembangkan metode ekstraksi informasi berbasis model BERT untuk mengidentifikasi berbagai parameter kunci dalam penelitian mutagenesis EMS, seperti dosis EMS yang digunakan, organisme target, gen yang mengalami mutasi, dan fenotipe yang didapatkan dari berbagai literatur ilmiah.
2. Menerapkan metode ekstraksi informasi berbasis model BERT untuk meningkatkan efisiensi ekstraksi literatur ilmiah mengenai mutagenesis EMS, yang berfokus pada kontekstualisasi dan pengelompokan informasi yang relevan.
3. Mengevaluasi kontribusi penggunaan model berbasis BERT dalam *text mining* untuk meningkatkan akurasi dan efektivitas pengelolaan informasi penelitian mengenai mutagenesis EMS, serta memberikan wawasan untuk pengembangan aplikasi NLP di bidang pemuliaan tanaman dan biologi molekuler.

1.4 Manfaat

Manfaat yang diambil dari penelitian ini sebagai berikut:

1. Meningkatkan efisiensi peneliti dalam pengelolaan literatur ilmiah mengenai mutagenesis EMS.
2. Mengisi kesenjangan penelitian terdahulu terutama pada bidang aplikasi NLP berbasis model BERT untuk studi genetika tanaman.
3. Mendukung pengembangan varietas tanaman toleran iklim sebagai upaya mitigasi dampak perubahan iklim di sektor pertanian Indonesia.

1.5 Batasan Masalah

Adapun batasan masalah pada penelitian ini:

1. Penelitian ini hanya difokuskan pada ekstraksi informasi dari publikasi ilmiah. Publikasi yang digunakan meliputi artikel jurnal ilmiah, laporan hasil konferensi, dan makalah lainnya yang dapat diakses secara terbuka (*open access*) berkaitan dengan mutagenesis EMS. Penelitian ini juga hanya berfokus pada parameter kunci pada penelitian mengenai mutagenesis EMS, seperti dosis EMS yang digunakan, organisme target, gen yang mengalami mutasi, dan fenotipe yang didapatkan dari berbagai literatur ilmiah.
2. Analisis dalam penelitian ini hanya akan mencakup informasi yang tersedia dalam format teks (abstrak, artikel, laporan), dan tidak akan mengintegrasikan data berupa gambar, grafik, atau tabel eksperimen yang ada dalam publikasi tersebut. Penelitian ini tidak akan mengkaji atau mengekstrak informasi dari sumber data non-teks seperti *file* mentah hasil eksperimen.