CHAPTER 1
INTRODUCTION

1.1 Background

The existence of global warming has a detrimental effect on many industries, including agriculture, where variations in temperature and rainfall may result in losses in crop yield, particularly in the production of staple food. Ten percent of the currently suitable area for major crops is projected to be climatically unsuitable (*Climate Change Impacts and Adaptation Options in the Agrifood System*, 2022). Plant breeding is one of the various approaches to this problem. Plant breeding is a critical field for developing new crop varieties with improved traits such as disease resistance, higher yield, and better nutritional content. There're many techniques available to do this such as: selective breeding, mutation breeding, transgenic breeding, and genome editing etc. But of all these techniques, mutation breeding, specifically induced mutagenesis is becoming increasingly popular in plant molecular biology as a method for identifying and isolating genes, as well as studying their structure and function. Molecular mutation breeding is ushering in a new era of crop enhancement mutation breeding (Yali & Mitiku, 2022).

Induced mutagenesis is a method used in mutation breeding to purposefully introduce mutations into an organism's DNA, usually a plant, in order to induce genetic variants that can result in desirable traits or attributes. In contrast to natural mutations, which happen randomly, induced mutagenesis is a planned and intentional procedure. There are four common mutagenesis methods (1) physical agents such as UV, X-ray radiation and fast neutron (FN), (2) chemical mutagens such as ethyl methane sulfonate (EMS), N-nitroso-N-methylurea (NMU), ethyl nitrosourea (ENU), diepoxybutane (DEB), (3) biological agents such as T-DNA and transposons, and (4) transgenic technologies such as CRISPR-Cas9, TALENs, gene knockdown using RNAi (Espina et al., 2018). Ethyl methane sulfonate (EMS) is one of the most popular chemical mutagens that induce mutations in plants (Xi-ou et al., 2017). EMS is a substance that acts as an alkylating agent and can cause point mutations in an organism's DNA. EMS is a derivative of methane sulfonic acid, where one of the hydrogen atoms has been replaced by an ethyl group (C2H5). The primary function of EMS is to alkylate the bases in DNA, which can result in base-pair changes. Particularly, it has a propensity to alter DNA's guanine (G) nucleotides, resulting in

G:C to A:T transitions in the genetic code. Therefore, EMS-induced mutations frequently entail the replacement of one base pair with another. It is possible for researchers and breeders to produce a variety of mutants in the lab thanks to EMS-induced mutagenesis, which is utilized to diversify the genetic makeup of populations. These mutants can be researched to learn more about how genes work or to create novel features in organisms, especially in plant breeding to create better crop types.

In this era, advances in computing technology have made a huge impact on how people live. One example is the rise of AI (Artificial Intelligence). In its strictest definition, AI stands for the imitation by computers of the intelligence inherent in humans (Sheikh et al., 2023), which can be categorized as ANI (Artificial Narrow Intelligence) and AGI (Artificial General Intelligence). In simple terms, ANI is an idea where machines perform a single task extremely well, even surpassing humans, while AGI is an idea where machines can be made and function as human minds.

Machine learning (ML) is an umbrella term that refers to a broad range of algorithms that perform intelligent predictions based on a data set (Nichols et al., 2018). Machine learning can be categorized into two types, such as supervised learning and unsupervised learning. Supervised learning is the use of labeled dataset to train algorithms so that they can predict a certain outcome or transform the data into something else, while unsupervised learning is the use of unlabeled dataset to be analyzed so they can produce a pattern.

Concurrently, machine learning has rapidly evolved and is now widely applied in science in general and in plant genotyping and phenotyping in particular (Van Dijk et al., 2021). In this context, the current research project, titled "Machine Learning-Based Predictive Modeling of Outcomes in EMS-Induced Plant Mutagenesis: An Innovative Approach for Genetic Improvement," seeks to leverage the wealth of data generated through EMS-induced mutagenesis. This project takes a pioneering approach by integrating machine learning techniques, especially supervised learning to predict and understand the outcomes of induced mutagenesis in plant. By doing this, it hopes to hasten crop genetic advancement and make it easier to create more robust and productive plant kinds.

1.2 Problem Statement

While induced mutagenesis, specifically the one using EMS, has many advantages for creating new varieties of crops, for example: simple operation, efficient mutation frequency and shortening the traditional breeding years (Lian et al., 2020), it has one big downfall, which is that the result is unpredictable. This is due to the fact that EMS does not target particular genes or areas, but rather randomly distributes point mutations across the genome. In order to find desirable mutants, comprehensive screening procedures are necessary since this unpredictability might lead to several mutations that are unrelated to the desired feature.

When experimenters conduct EMS mutagenesis, they generally refer to published data and make appropriate choices based on the actual situation, instead of blindly conducting experiments according to the published data (Chen et al., 2023). To produce a new crop variety, the researcher or breeder will conduct mutagenesis on multiple seeds. The steps can be explained as of follows: (1) Preparation of seeds: Seeds are soaked in distilled water for 24 hours to ensure uniform hydration. (2) Preparation of EMS solution: A stock solution of EMS is prepared by dissolving it in distilled water. The concentration of the stock solution depends on the plant species and the desired mutation frequency. (3) Treatment of seeds: The hydrated seeds are treated with EMS solution for a specific duration depending on the plant species and the desired mutation frequency. (4) Rinsing and germination: The treated seeds are rinsed with distilled water to remove any residual EMS and then germinated under controlled conditions. (5) Screening for mutants: The progeny of the treated seeds are screened for visible phenotypic variations or traits of interest. These traits can include resistance to pests, improved yield, altered flowering time, or any other characteristic. Of all these processes, a few problems can be mentioned, such as:

1. Because of the nature of mutation, the result of mutagenesis is completely random. Making it harder for researchers and breeders to determine treatment parameter.
2. The mutated seed quality is only visible after it has been planted, which takes quite a long time. This makes the researcher do a "guess game" to obtain the treatment parameter.
3. There's no easy-to-use application for researchers to know whether their treatment parameter is suitable or not for the seed.

1.3 Objectives

1. Develop a machine learning model that uses historical data from past treatment created by other researchers or breeders to predict whether the treatment is suitable or not for the seed
2. Develop a system that can give the researcher or breeder a list of treatment parameters that give the best result for a certain seed.
3. Develop a demo web app to showcase the model's capabilities.

1.4 Scope

1. Project Scope
   - Dataset collection

     The dataset utilized to train this model was meticulously curated from a diverse array of sources, primarily sourced from a multitude of scientific journals and articles available on the internet. A substantial portion of these sources were derived from research and studies conducted within the Indonesian scientific community. The dataset itself exhibits a well-structured format, featuring a range of pertinent variables that are instrumental in understanding the process of mutation in plants. These variables encompass the plant species under investigation, the duration of soaking, the concentration of Ethyl Methane Sulfonate (EMS), and, most significantly, the outcome of the experiment. This outcome is the product of careful observation and analysis, aimed at determining whether the resulting mutant plant exhibits improvements over its non-mutated counterpart, often referred to as the "control plant."
   - Treatment parameters

     The primary treatment parameters that are used to predict the result are limited to only four parameters, which are the plant species, the duration of soaking, the concentration of EMS. While these four parameters serve as the backbone of the predictive framework, it is worth noting that there exist other variable that can be inputted for additional context and refinement. For instance, there is the option to include the post-soak duration, which indicates the second soak process to further facilitate the mutagenic effects. However, these supplementary parameters are considered optional. The reasoning behind this selectivity is rooted in the predominant characteristics of the data. The majority of the dataset relies solely on the three primary parameters mentioned earlier, with the goal of maintaining a focused and consistent basis for analysis. By adhering to this core set of parameters, the dataset achieves a high degree of uniformity, ensuring that the model's predictions are grounded in a robust and dependable foundation of information.

- Chemical agent

  The choice of employing only Ethyl Methane Sulfonate (EMS) as the sole chemical agent in the mutagenesis process is a strategic decision, grounded in the principles of data collection consistency and the pursuit of accurate and dependable results. By limiting the mutagenic process to EMS exclusively, several important objectives are achieved. First and foremost, this singular focus on EMS streamlines the data collection process, significantly simplifying the acquisition and management of pertinent information. This simplicity enhances the efficiency of data gathering, ensuring that the dataset remains cohesive and uniform, with all experimental conditions and parameters standardized around a single mutagenic agent. Another key advantage of utilizing only EMS as the mutagenic agent is the assurance of consistent and reliable results. The model's predictive capabilities are honed and fine-tuned to respond to the unique nuances and outcomes associated with EMS-induced mutations. This focus on consistency is paramount, as it helps avoid potential discrepancies and confounding variables that might arise when dealing with multiple mutagenic agents. Furthermore, it is important to underscore that the predictions generated by the model, which are based on the use of EMS, are not intended to be extrapolated or applied to scenarios involving different mutagenic agents. Each mutagenic agent may introduce distinct molecular changes and responses in the plant species, and applying predictions made for EMS to other agents could lead to erroneous or misleading results in the treatment.

- Agrotechnology

  Agricultural production is under threat due to climate change in food insecure region (Habib-ur-Rahman et al., 2022). The challenges posed by climate change have significantly impacted agriculture, making it more crucial than ever for farmers to adapt and find innovative solutions to ensure food security and sustainability. As changing weather patterns, increased temperatures, and unpredictable climatic conditions continue to threaten crop yields, there is an urgent need to develop better crop varieties that can withstand these challenges and produce more reliable results. One promising avenue to address this pressing issue is the integration of cutting-edge biotechnology with traditional breeding methods, specifically through the utilization of mutation breeding. Mutation breeding involves inducing genetic changes or mutations in plants, which can lead to the development of novel traits, such as resistance to diseases, improved tolerance to environmental stressors, and increased yield potential. This approach has been employed for decades to enhance crop varieties. By leveraging the predictive capabilities of models, such as AI and machine learning-based systems, in conjunction with biotechnology and mutation breeding, farmers gain access to a powerful toolkit for crop improvement. These models can help predict the potential outcomes of mutations, and streamline the breeding process by reducing the time and resources needed to develop new crop varieties.

- Species

  There're a few numbers of species that is supported by the application which are:

  - Rice (Oryza sativa) (Oryza sativa - Malaysia) (Oryza sativa – Lokal Ende)
  - Chili (Capsicum frutescens) (Capsicum frutescens - Bara)
  - Black soybean (Glycine max (L) Merrit)
  - Carrot (Daucus carota)
  - Tobacco (Nicotiana tabacum)
  - Soybean (Glycine max) (Glycine max – Dering 1) (Glycine max – Grobogan) (Glycine max – Drought tolerant)
  - Purple sweet potato (Ipomoea batatas L. Poir)
  - Red pepper (Capsicum annuum L.)
  - Potato (Solanum tuberosum L - Granola)
  - Common wheat (Triticum aestivum L. – Dewata) (Triticum aestivum L. – Selayar) (Triticum aestivum L. – Nias)
  - Black pepper (Piper nigrum L.)
  - Oyster mushroom (Plaerotus ostreatus)
  - Shallot (Allium ascalonicum L.)
  - Porang (Amorphophallus muelleri)
  - Banana (Musa acuminata – Kepok)

- Fameflower (Talinum paniculatum (Jacq.) Gaertn.)

The reason behind the limitation of the supported species is because the data available in scientific journal is limited. Scientific research, particularly in the field of biology and taxonomy, often focuses on a select number of species that are of particular interest or relevance to ongoing studies.

2. System Scope
   - Users can specify the treatment parameters that will be applied to the seeds using the system's extensive input mechanism. This includes EMS (Ethyl Methane Sulfonate) percentage concentration, which specifies the strength or concentration of the mutagenic agent to be used, and soaking duration, which establishes how long the seeds are exposed to the treatment. With the ability to precisely and precisely customize the treatment process, this detailed input capability improves the system's prediction of the probability that the applied treatment will result in the development of a better crop variety.
   - The system's evaluation of the treatment's impact on the seeds is conveyed through a user-friendly interface, where the outcome is presented in a clear and informative manner. Based on its analysis, the model will provide one of two distinct responses: (A) "This treatment is not suitable for this seed": In cases where the specified treatment parameters are incompatible with the genetic characteristics of the seed or are unlikely to induce significant changes, the system will promptly communicate that this treatment option is unsuitable. This feedback helps users make informed decisions about their crop improvement strategies. (B) "This treatment will display phenotypic variation": When the system determines that the treatment is likely to induce noticeable changes in the seed's characteristics, it will present this positive outcome. This response indicates that the applied treatment has the potential to stimulate phenotypic variation, which is a critical step toward achieving improved crop varieties. This valuable information empowers users to proceed with confidence, knowing that the treatment holds promise for enhancing crop traits and genetic diversity.

1.5 Significance

By implementing ML in the mutagenesis process, the probability of creating a better crop variety is significantly increased. Thus, it helps address the continuous global problem of improving food security and sustainability. Not only that, but it also helps to maximize the time and financial efficiency needed to develop new seed varieties.

1.6 Summary

Global warming is causing crop yield losses, particularly in agriculture, which relies on plant breeding techniques for improved traits like disease resistance and nutritional content. Induced mutagenesis, a method of introducing mutations into an organism's DNA, is becoming increasingly popular in plant molecular biology. This intentional procedure can be achieved using chemical mutagens, physical mutagens, and radiation. EMS (Ethyl Methane Sulfonate) is the most popular method for induced mutagenesis, as it acts as an alkylating agent and can cause point mutations in DNA. This allows researchers and breeders to diversify the genetic makeup of populations and research novel features in organisms, particularly in plant breeding.

Advances in computing technology, such as artificial intelligence (ANI) and artificial general intelligence (AGI), have significantly impacted human life. Machine learning, a tool of AI, is a study of computer algorithms that enable machines to learn from and improve upon data analysis. This research project aims to leverage the wealth of data generated through EMS-induced mutagenesis by integrating machine learning techniques, particularly supervised learning, to predict and understand the outcomes of induced mutagenesis in plants.