www.journalworlds.com
AGRI JOURNAL WORLD
VOL. 5 ISSUE 3 JULY-SEPTEMBER, 2025

APPLICATION OF DATA MINING IN CROP PROTECTION

Riya Kundu, Neethu Narayanan, V. S. Rana, N. A Shakil, Parshant Kaushik*

Division of Agricultural Chemicals, ICAR-Indian Agricultural Research Institute, New Delhi-110012

**Corresponding author email: parshantagrico@gmail.com*

ABSTRACT

Artificial Intelligence (AI) has revolutionized agriculture by generating vast amounts of data through technologies like drones and GPS. However, managing and analyzing such complex data requires advanced approaches. Data mining has emerged as a vital tool for uncovering hidden patterns, supporting precise decision-making, and addressing critical challenges in crop protection. Applications include classifying herbicide resistance detecting invasive weeds using deep neural networks monitoring pesticide residues through intelligent systems like PRIAS diagnosing plant diseases using SVM classifiers and improving paddy leaf disease monitoring with big data models. Overall, data mining strengthens agricultural sustainability and food security in a data-driven era.

KEYWORDS: Crop protection, Data mining, Machine learning, Precision agriculture, Pest detection

INTRODUCTION

Artificial Intelligence (AI) significantly influences daily life by improving various processes in multiple sectors. It leads to greater efficiency, productivity, and improved decision-making through automation, tailored experiences, and comprehensive data analysis. Ranging from software tools that simplify tasks to voice-activated assistants that provide convenience, AI has revolutionized the interactions and operations of individuals and businesses alike (FITA, 2020). Technological advancements in agriculture have evolved from subsistence-based practices in Agriculture 1.0, characterized by the plow and animal drafts, to Agriculture 5.0, or digital agriculture, which employs technologies like AI, IoT, and precision farming for optimized crop yields and resource use. Agriculture 3.0 focuses on smart farming, utilizing GPS for data-driven management, while Agriculture 4.0 incorporates connected farming practices with drones and autonomous machinery for enhanced decision-making. These advancements ensure improved efficiency, productivity, and sustainability across agricultural processes (Fig. 1) (Mohd, J. *et al.* 2022 and Botta, A., *et al*.,2022).

With advancements in IT, efficiency improvements are now possible in nearly every industry and service sector, including agriculture. Modern farmers not only harvest crops but also gather increasing amounts of data. This data is often detailed and granular. However, managing large volumes of data can be both a benefit and a challenge. There is an abundance of data available, such as soil and yield characteristics, that should be leveraged to the farmer's advantage. This is a common challenge, often addressed by data mining. Data mining techniques are designed to uncover patterns or insights in the data that are both valuable and of interest to the farmer (Raorane *et al.*,2013).

EVOLUTION OF DATA MINING

The evolution and development of finding the right data for decision-making are shown in Figure 1 (Rohanizadeh, *et al.,* 2009). Data handling has evolved over the decades. In the late 1960s and 1970s, basic pre-formatted reports were generated from stored datasets. By the 1980s, users demanded more frequent and personalized information, prompting formal queries to access data. In the 1990s, the need for real-time, "just-in-time" information led users to create their own queries for precise data extraction. In recent years, the growth of large datasets has highlighted the need for advanced tools and techniques, giving rise to data mining, which helps uncover patterns and relationships to make data more relevant and actionable.

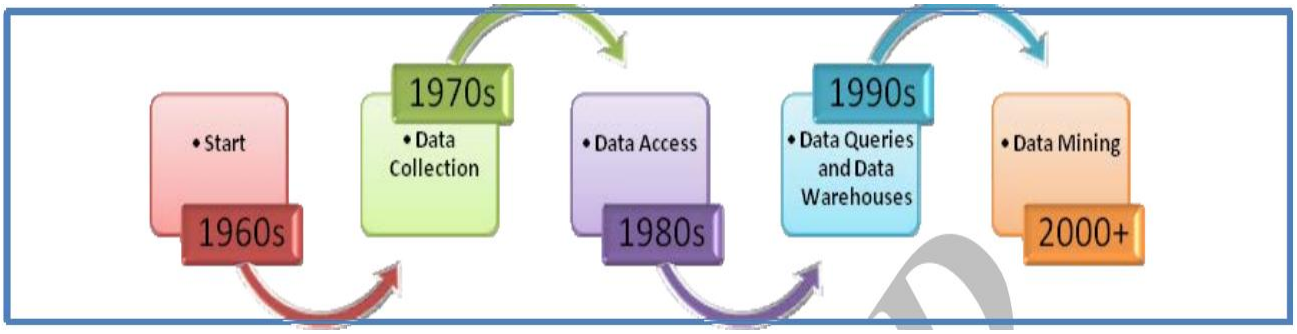


Fig. 1 Evolution of Data Mining

SOURCES OF DATA FOR MINING

- **Data Warehouses** – Data warehouses act as centralized repositories integrating historical data from multiple sources; ensure quality and consistency for trend analysis.
- **Transactional Databases** – Transactional databases are optimized for handling real-time operations and transactions; data can be mined directly or after aggregation into warehouses.

- **Spatial and Temporal Data** – Capture location and time-based information; useful for detecting geographic patterns and time-series trends.
- **Time-Series Data** – These are sequential data collected over intervals; supports prediction of future values and strategic planning.
- **Multimedia and Text Data** – It includes images, videos, and documents; mined using natural language processing and computer vision to extract insights on behavior and preferences.

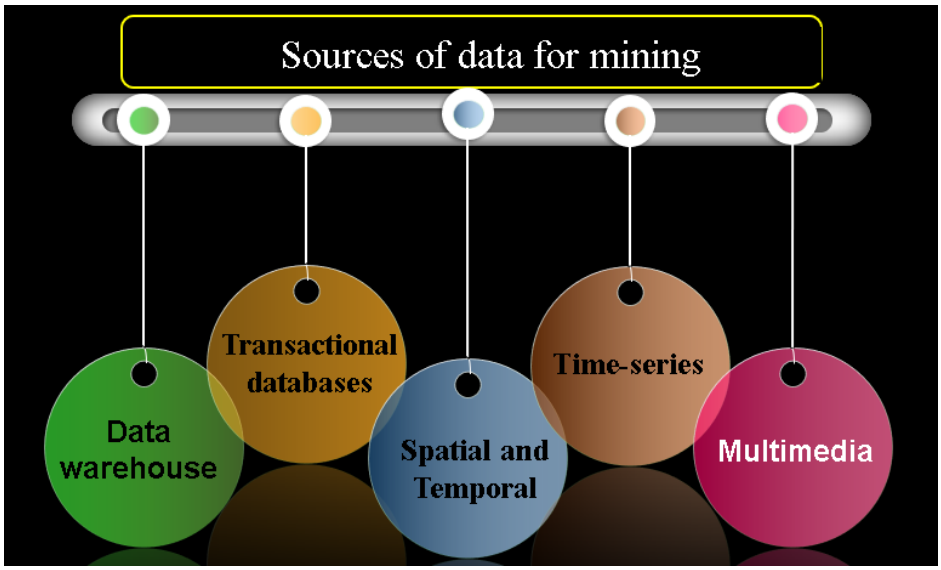


Fig. 2 Sources of data for mining

.

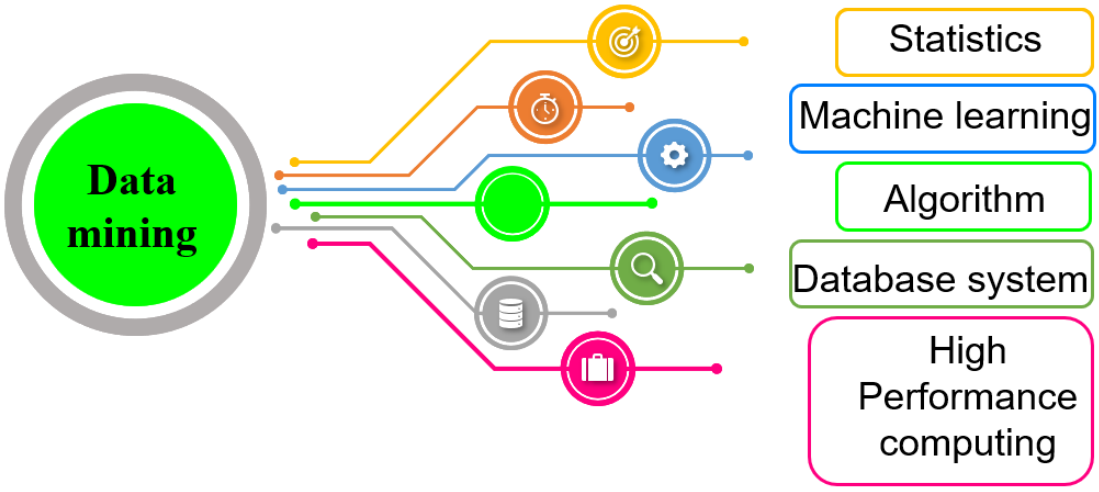


Fig. 3 Techniques of Data Mining

WHAT ARE THE TECHNIQUES USED IN DATA MINING?

Data mining employs a variety of techniques to uncover hidden patterns, relationships, and insights from large datasets. These techniques are designed to support prediction, classification, clustering, association analysis, and anomaly detection. The choice of technique depends on the type of data, the objective of analysis, and the desired outcomes. Some of the commonly used techniques in data mining include statistics, machine learning, algorithm, database system, and high performance computing

STATISTICS IN DATA MINING

Statistics is essential in analyzing large datasets to uncover patterns and correlations. It involves organizing, analyzing, and interpreting numerical data to aid decision-making. It is of two types

I. **Descriptive statistics:** Descriptive statistics focus on visualizing datasets using numerical and graphical methods to reveal trends, summarize essential information, and present it in a clear and understandable manner. This process helps users make better-informed decisions.

II. **Inferential statistics:** inferential statistics involve drawing conclusions, making estimates, predictions, and decisions based on data samples drawn from a larger dataset.

MACHINE LEARNING (ML)

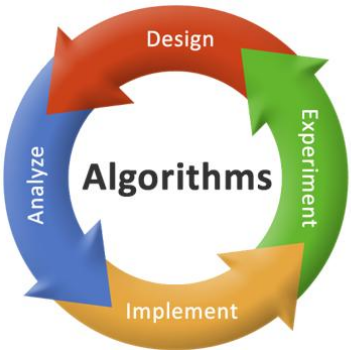
Machine learning is a field that enables computers to learn from data without explicit programming, allowing them to process information efficiently and extract meaningful insights. With the rise of large datasets, its demand has grown across industries to uncover hidden patterns and improve decision-making. The primary goal is for systems to learn and improve from experience using various algorithms, since no single method suits all problems. The choice of algorithm depends on factors like problem type, variables, and the best-fit model. Commonly used machine learning algorithms include classification, regression, clustering, and association methods.

- **Example:** Gmail spam filter learns from examples of spam and non-spam emails to improve accuracy.

ALGORITHMS IN DATA MINING

An algorithm is a set of step-by-step instructions that describe how to perform a task. In data mining and machine learning, algorithms play a crucial role in processing and analyzing data to extract valuable insights. These algorithms can be broadly categorized into several types, each serving a specific purpose.

Algorithms are mainly two types- supervised learning algorithm and unsupervised learning algorithm.



I. **Supervised Learning:** Uses labeled data to predict outputs. Examples: Linear regression, logistic regression, decision trees, SVM, neural networks.

II. **Unsupervised Learning:** Finds patterns without labels. Examples: K-means clustering, hierarchical clustering, PCA, association rule learning.

DATABASE SYSTEMS & DATA WAREHOUSES

- **Database:** A database system is like an organized collection of data. It stores information in a structured way, making it easy to retrieve, update, and manage data. Databases are used for day-to-day operations, like keeping track of customer orders or inventory.
- **Data Warehouse:** A data warehouse is a specialized type of database. It acts as a central repository integrating data from multiple sources for analysis and reporting. Enables complex queries and pattern discovery.

HIGH PERFORMANCE COMPUTING (HPC)

HPC plays a crucial role in data mining, particularly when dealing with large datasets, complex algorithms, and the need for rapid analysis. From a data mining perspective, HPC refers to the use of powerful computing resources to perform data analysis tasks more efficiently and effectively.

ROLES IN DATA MINING

1. Manage large datasets
2. Accelerate algorithm execution
3. Implement complex models (e.g., deep learning)
4. Parallel processing for faster computation

5. Scalability for growing data
6. Improve model accuracy
7. Optimize computational resources

HOW DOES DATA MINING WORK?

The data mining process begins with data source identification, where relevant information is collected from multiple sources. Next, data selection and sampling are performed to focus on subsets that reduce complexity. The data is then pre-processed by cleaning, handling missing values, and normalizing, followed by data transformation, which includes feature selection, dimensionality reduction, or deriving new variables. During data exploration, statistical analysis and visualization techniques are used to understand the data, which then guides the modeling stage where algorithms such as classification, clustering, or regression are applied to uncover meaningful patterns. In the pattern recognition stage, relationships, clusters, or rules are identified. The results are then subjected to evaluation and interpretation to assess accuracy and relevance. Insights are communicated through knowledge presentation via reports, dashboards, or visualizations, and finally, in the deployment stage, the models are applied to support real-world decision-making.

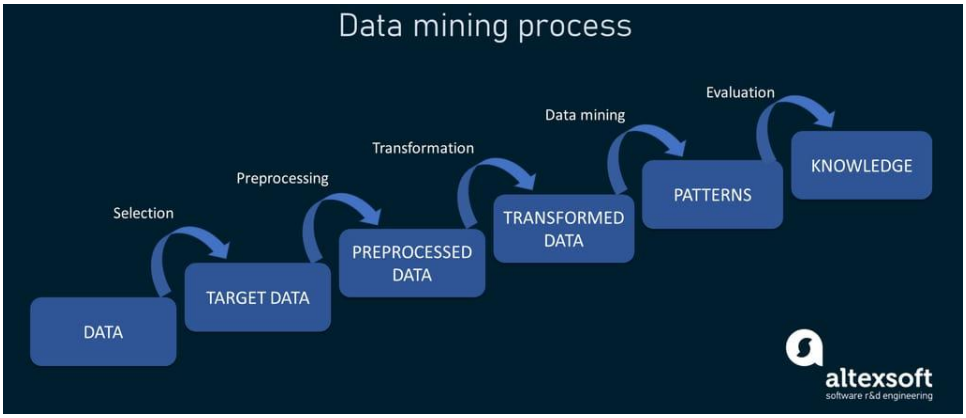


Fig. 4 Steps in data mining process

DATA MINING METHODOLOGIES

The main methodologies in data mining are **Classification, Clustering, Regression, and Association Rules**. These techniques are widely used in solving agricultural and other real-world problems.

1. CLASSIFICATION

Classification is the process of predicting the class of a new item. Therefore, to classify the new item and identify to which class it belongs. Classification is a supervised learning process that allows the prediction of a class label from a set of training data

- **Learning approaches:** Supervised, unsupervised, semi-supervised.
- **Common Algorithms:**
  - ✓ **Decision Tree:** Tree-like model using yes/no decisions.
  - ✓ **Artificial Neural Networks (ANN):** Pattern-based predictions inspired by the human brain.
  - ✓ **Bayesian Classification:** Probability-based predictions using Bayes' theorem.
  - ✓ **K-Nearest Neighbor (K-NN):** Classifies based on nearest data points.
  - ✓ **Support Vector Machine (SVM):** Creates boundaries between classes.
  - ✓ **Genetic Algorithm:** Evolution-inspired optimization for classification.
  - ✓ **Fuzzy Logic:** Handles uncertain or imprecise data.

2. CLUSTERING

As opposed to classification, clustering is an unsupervised learning process where classes are not known in advance. It consists in dividing objects into groups (clusters) based on information found within the data which describes these objects and their relationships. Clustering techniques are applied when there are no predefined classes and instances must be divided into groups

- **Goal:** To identify inherent patterns and relationships in data.
- **Common Methods:**
  - ➢ **Hierarchical Clustering:** It builds a tree of clusters.
  - ➢ **K-Means Clustering:** It divides data into K distinct clusters.
  - ➢ **Density-Based Clustering:** It forms clusters based on dense regions, handling noisy data.

3. REGRESSION

Regression is a data mining (machine learning technique) used to fit an equation to a dataset. A straight line is given by the equation y= mx +c and determines the approximate values for m and c to calculate the value of y based on a particular value of x. Multiple regression, uses more than one input variable and allows for the fitting of more complex models

- **Types:**
  - **Simple Linear Regression:** Models relationship between one predictor (X) and one outcome (Y).
  - **Multiple Linear Regression:** Models relationship between multiple predictors and an outcome.
- **Applications:** It is used to predict outcomes like crop yield, house prices, or sales.

4. ASSOCIATION RULES

An association algorithm creates rules that describe how often events have occurred together. **Example:** "If a customer buys a computer, there is a 90% chance they will buy software."

- **Applications:** Market basket analysis, customer segmentation, store layout optimization, telecommunication alarms.
- **Common Algorithms:** Apriori, FP-Growth, Eclat, Declat, DHP, DIC, SEAR, Spear, MaxEclat.

COMMERCIAL TOOLS

- Oracle Data Miner
  http://www.oracle.com
- Data To Knowledge
  http://alg.ncsa.uluc.edu
- SAS
  http://www.sas.com/
- Clementine
  http://spss.com/clemetine/
- Intelligent Miner
  http://www-306.ibm.com/software

APPLICATION OF DATA MINING IN CROP PROTECTION

❖ **MODELING & PREDICTION OF HERBICIDE RESISTANCE**

Weed species resistant to multiple herbicide modes of action (MoAs) are increasing, posing a threat to current herbicide effectiveness. Hierarchical clustering grouped MoAs into three clusters—HRAC 2, 4, 5, 9; HRAC 12, 14, 15; and HRAC 1, 3, 22—based on shared resistant weeds, reflecting similarities in physiological or biochemical targets. Network analysis showed that resistance to two MoAs is linked to resistance to individual MoAs. Farmers can manage

resistance more effectively by rotating herbicides between clusters rather than within clusters, while considering crop, weed, and environmental conditions (Hulme, 2022).

- **DETECTION OF PLANT DISEASES**

Padol, *et al.*, 2018, conducted a study focused on detecting grape leaf diseases by first applying K-means clustering for image segmentation to identify diseased regions, then extracting both color and texture features from the segmented images, and finally classifying the type of disease using a Support Vector Machine (SVM) classifier.

- **DATA MINING IN PESTICIDE RESIDUE DATA ANALYSIS**

An intelligent system was developed to collect, process, and analyze pesticide residue data by integrating detection results with maximum residue limit standards, applying statistical methods for multi-dimensional risk identification, and using the Apriori algorithm to mine hidden associations for pre-warning rules. The system also generated automatic reports and, when applied to residue data from 42 Chinese cities (2012–2015), significantly improved the depth, accuracy, and efficiency of analysis, providing stronger decision support for food safety supervision (Chen, *et al.*, 2022).

- **DETECTION & MANAGEMENT OF INVASIVE SPECIES and WEEDS**

Ullah, *et al.*, 2019, conducted a study aimed to develop a machine vision system to identify and classify wheat using pattern recognition techniques, and design a real-time robotics system to locate and categorize wheat for selective herbicide application with the help of morphological algorithms. The proposed approach was further evaluated and optimized to achieve high classification accuracy, with a target of 94% success under diverse field conditions.

- **DISEASE MONITORING SYSTEMS (DMS) FOR PLANT DISEASES**

An efficient Disease Monitoring System (DMS) for paddy leaf diseases was designed using big data mining techniques (Suresh et al., 2020). The proposed model aims to improve the accuracy of both segmentation and classification in monitoring paddy leaf diseases. Data mining methodologies are deployed in the working platform of MATLAB.

- **PEST INFESTATION DETECTION**

Early Detection of Red Palm Weevil, *Rhynchophorus ferrugineus* (Olivier), Infestation Using Data Mining (Kurdi, *et al*, 2021, Plants)**.** This study evaluates ten classification algorithms for predicting early Red Palm Weevil (RPW) infestation using plant-size and temperature data. Their performance will be assessed through accuracy, precision, recall, and F-measure on a real RPW dataset. Key features such as temperature and plant circumference will be analyzed for their

predictive importance, while highlighting the need for more RPW datasets to improve validation and reliability.

Table 1: Open access databases (Wang, *et al.*, 2022)

Database	Last update	Substance Records
Compendium of Pesticide Common Names (CPCN)	June 2022	>1200 official common names by ISO for >1800 different active ingredients and 350 ester and salt derivatives.
Pesticides Properties DataBase (PPDB)	June 28,2022	1906 pesticides physicochemical properties
Bio-Pesticides DataBase (BPDB)	June 28,2022	738 bio-pesticides physicochemical properties
ChEMBL	2022	Over 115 million chemical structures, properties, and associated information integrated from hundreds of sources
Arthropod Pesticide Resistance Database (APRD)	2022	17,863 Case reports of drug resistance (Citations of Resistance and Locations) in the United States, Europe, and other countries from 1914 to the present.
International Survey of Herbicide Resistant Weeds	July 3, 2022	513 unique cases of herbicide resistant weeds globally resistant 267 species
PAN Pesticide Database	September 7, 2000	About 15,300 active ingredients details

ADVANTAGES OF DATA MINING IN AGRICULTURE

1. **IMPROVED DECISION MAKING**: Data mining helps farmers make data-driven decisions on crop selection, planting, irrigation, fertilization, and pest management by identifying patterns and correlations, leading to more effective and informed practices.

2. **INCREASED EFFICIENCY**: Data mining identifies inefficiencies in farming and optimizes resource use like water, fertilizers, and labor, reducing waste and costs while improving yields and overall productivity.
3. **EARLY DETECTION OF PESTS AND DISEASES**: Data mining enables early detection of pest and disease outbreaks through predictive models, allowing timely interventions that reduce crop losses and limit chemical use.
4. **PRECISION AGRICULTURE**: By analyzing data from soil sensors, satellites, and weather forecasts, data mining supports precision farming, helping farmers optimize inputs for specific plots or plants, boosting productivity and conserving resources.
5. **ENVIRONMENTAL BENEFITS**: Data mining promotes sustainability by guiding precise use of fertilizers and pesticides, reducing runoff and soil contamination, while supporting practices that enhance soil health, biodiversity, and carbon sequestration.
6. **HISTORICAL ANALYSIS**: It helps analyze past agricultural data to identify trends, evaluate policy impacts, and adapt to climate or market changes, providing benchmarks for future planning and risk management.

DISADVANTAGES

1. **INITIAL COST:** Implementing data mining involves significant initial costs, including investments in software, hardware, and data storage infrastructure, as well as training personnel. This can be a barrier for smaller farms or organizations with limited budgets.
2. **TECHNICAL EXPERTISE REQUIRED:** Data mining requires specialized knowledge and skills in data analysis, machine learning, and statistics. The need for experts to interpret complex data patterns can make it challenging for organizations without dedicated data science teams.
3. **DATA QUALITY CONCERNS:** The effectiveness of data mining depends heavily on the quality of the data being analyzed. Incomplete, outdated, or inaccurate data can lead to unreliable results, reducing the utility and trustworthiness of insights derived.
4. **OVERRELIANCE ON TECHNOLOGY:** There is a risk of becoming overly dependent on data mining tools and technology, potentially leading to neglect of other important factors, such as local knowledge, experience, and intuition, which can be crucial in decision-making.

5. **INTEGRATION CHALLENGES:** Integrating data mining systems with existing workflows, databases, and technologies can be complex and time-consuming. Compatibility issues and resistance to change from staff can also hinder the successful implementation of data mining solutions.

FUTURE PROSPECTS

1. **ADVANCED PREDICTIVE MODELING**: Future data mining will use advanced machine learning to more accurately predict weather, pest outbreaks, and yields, improving farmer decision-making
2. **INTEGRATION WITH EMERGING TECHNOLOGIES**: Data mining will integrate with IoT, blockchain, and edge computing—IoT for real-time data collection, blockchain for data integrity, and edge computing for timely interventions
3. **IMPROVED DATA INTEGRATION**: Unified platforms will combine data from satellites, weather stations, and sensors, offering a holistic view of farm operations for better analysis and coordinated responses
4. **AI-DRIVEN DRONES**: AI-powered drones will monitor crop health, detect pests, and apply treatments precisely where needed, minimizing waste and boosting efficiency
5. **AUTONOMOUS ROBOTS**: Robots guided by data mining will enable targeted pest removal, automated weeding, and plant-level health monitoring, reducing labor costs and chemical use.

CONCLUSION

Data mining is transforming crop protection by converting vast agricultural data into actionable intelligence. Through machine learning, predictive modeling, and pattern recognition, it enhances early pest and disease detection, optimizes pesticide use, and improves decision-making. Integration with technologies like AI, IoT, and remote sensing enables precision agriculture and sustainable pest management. Despite challenges such as high implementation costs and data quality issues, the potential benefits are immense. Future advancements in AI-driven analytics, drones, and robotics promise more efficient, eco-friendly, and resilient farming systems, positioning data mining as a cornerstone of modern, data-driven agricultural innovation and sustainability

REFERENCES

Botta, A., Cavallone, P., Baglieri, L., Colucci G., Tagliavini, L., and Quaglia G. (2022). A Review of Robots, Perception, and Tasks in Precision Agriculture. *Applied mechanics*, *3*(3), 830–854. https://doi.org/10.3390/applmech3030049 .

Chen, Y., Dou, H., Chang, Q., and Fan, C. (2022). PRIAS: an intelligent analysis system for pesticide residue detection data and its application in food safety supervision. *Foods*, *11*(6), 780- 795.

FITA (2020). The importance of artificial intelligence in everyday life. (https://www.fita.in/the-importance-of-artificial-intelligence-in-everyday-life/).

Hulme, P. E. (2022). Hierarchical cluster analysis of herbicide modes of action reveals distinct classes of multiple resistance in weeds. *Pest Management Science*, *78*(3), 1265-1271.

Kaufman, K. A., and Michalski, R. S. (2005). From data mining to knowledge mining. *Handbook of Statistics*, *24*, 47-75.

Kurdi, H., Al-Aldawsari, A., Al-Turaiki, I., and Aldawood, A. S. (2021). Early detection of red palm weevil, *Rhynchophorus ferrugineus* (Olivier), infestation using data mining. *Plants*, *10*(1), 95-103.

Madni, H. A., Anwar, Z., and Shah, M. A. (2017). Data mining techniques and applications—A decade review. In *2017 23rd International Conference On Automation And Computing (ICAC)* (pp. 1-7). IEEE.

Padol, P. B., and Yadav, A. A. (2018). SVM classifier based grape leaf disease detection. In *2016 Conference On Advances In Signal Processing* ,175-179. IEEE.

Raorane, A., and Kulkarni, R. (2013). Role of data mining in Agriculture. *International Journal of Computer Science Information Technology*, *4*(2), 270-272.

Rohanizadeh, S. S., and Bameni, M. M. (2009). A proposed data mining methodology and its application to industrial procedures. *Journal of Industrial Engineering*, *4*, 37-50.

Suresh, K., Karthik, S., & Hanumanthappa, M. (2020). Design an efficient disease monitoring system for paddy leaves based on big data mining. *Inteligencia Artificial*, *23*(65), 86-99.

Ullah, A., Nawi, N. M., Arifianto, A., Ahmed, I., Aamir, M., and Khan, S. N. (2019). Real-time wheat classification system for selective herbicides using broad wheat estimation in deep neural network. *International Journal of Advance Science Engineering Information Technology*, *9*, 153-158.

Wang, D., Deng, H., Zhang, T., Tian, F., and Wei, D. (2022). Open access databases available for the pesticide lead discovery. *Pesticide Biochemistry and Physiology*, *188*, 105267.

How to cite:

Kundu, R., Narayanan, N., Rana, V.S., Shakil, N.A., and Kaushik, P. (2025). Application of data mining in crop protection. Leaves and Dew Publication, New Delhi 110059. *Agri Journal World* 5 (3): 57-69.

**********XXXXX**********