SMART INSECT-PEST MANAGEMENT FOR COTTON CROPS

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Engineering

Raúl TOSCANO

Universidad EAFIT

Escuela de Ingeniería

Medellín, Colombia

15/03/2024

Director of dissertation research:

Prof. Jose Aguilar — Systems Engineering Department

Examining committee:

Prof. OSWALDO TERÁN — Universidad Católica del Norte, Coquimbo, Chile

Prof. HÉCTOR CANCELA — Universidad de la Republica Uruguay, Montevideo, Uruguay

Prof. DIEGO VILLANUEVA — Universidad EAFIT, Medellín, Colombia

Dissertation committee chairperson:

Prof. Jorge Prieto Salazar — Director of the Doctorate in Engineering

President of the Board of the Doctoral Program in Engineering:

Prof. Ricardo Taborda Rios — Dean School of Engineering

Dissertation date : 2024

Abstract

In this research, we address the problem of smart insect-pest management for cotton crops. For the study of this problem, we have positioned it in the framework of the paradigm of Smart agriculture. In this context, Smart agriculture, also known as precision agriculture or digital agriculture, involves the use of advanced technologies to improve agricultural productivity, efficiency, and sustainability. Its focus is to use data-driven and innovative approaches to optimize farming practices and reduce resource waste while ensuring food security. The development of approaches to aid in decision-making for smart insect-pest management for agriculture is necessary to avoid the massive spread of insect pests and the increase in environmental impact. Despite the existence of advances in smart agriculture, integrated management of insect pests remains a challenge. To address this problem, our objective was to develop methodologies, models, and approaches to support decision-making related to smart insect-pest management for cotton crops. To achieve this objective, several sub-objectives were raised, the first one was to design a metacognitive architecture for the smart management of cotton pests, the second was to implement knowledge models for the smart management of cotton pests, and the third was to implement novel AI concepts for the development of knowledge models. Particularly, several research articles were developed to meet the objectives proposed in this thesis. Initially, a review article on the latest trends in Smart agriculture using artificial intelligence and sensing techniques for the management of insect pests and diseases in cotton was carried out. On the other hand, for the first sub-objective, an article was conducted where a metacognitive architecture with metacognitive tasks (meta-memory, meta-learning, meta-reasoning, meta-comprehension, and meta-knowledge) was proposed for smart-pest management of cotton. To meet the second sub-objective, two articles were proposed. The first article is a classification model of the cotton boll-weevil population and the second article presented a fuzzy classification system to analyze the yield of cotton production. Regarding the third sub-objective, two articles were proposed. The first article is about a system with autonomous cycles of data analysis tasks for the integrated management of cotton. And the second article shows how to enhance the insect pest classification in cotton using Transfer Learning techniques. In each article, the strategies/models were evaluated using various datasets. The results showed the capacity of the developed methodologies and models for decision-making in smart insect-pest management for cotton crops. Specifically, our proposals allow the prediction of the boll-weevil behaviors, the diagnosis/prediction of cotton yield, and the prescription of strategies for cotton management into a framework of a meta-cognitive architecture, with good results in performance metrics.

Keywords: Artificial intelligence, Machine learning, Predictive modeling, Prescriptive modeling, Smart agriculture, Precision agriculture, Cotton

Resumen

En esta investigación, abordamos el problema del manejo inteligente de plagas de insectos para cultivos de algodón. Para el estudio de esta problemática nos hemos posicionado en el marco del paradigma de la Agricultura Inteligente. En este contexto, la agricultura inteligente, también conocida como agricultura de precisión o agricultura digital, implica el uso de tecnologías avanzadas para mejorar la productividad, la eficiencia y la sostenibilidad agrícolas. Su objetivo es utilizar enfoques innovadores y basados en datos, para optimizar las prácticas agrícolas y reducir el desperdicio de recursos al tiempo que se garantiza la seguridad alimentaria. El desarrollo de enfoques que ayuden en la toma de decisiones para la gestión inteligente de plagas de insectos para la agricultura es necesario para evitar la propagación masiva de plagas de insectos y el aumento del impacto ambiental. A pesar de la existencia de avances en agricultura inteligente, el manejo integrado de plagas de insectos sigue siendo un desafío. Para abordar este problema, nuestro objetivo fue desarrollar metodologías, modelos y enfoques para apoyar la toma de decisiones relacionadas con el manejo inteligente de plagas de insectos para cultivos de algodón. Para lograr este objetivo se plantearon varios subobjetivos, el primero fue diseñar una arquitectura metacognitiva para el manejo inteligente de plagas del algodón, el segundo fue implementar modelos de conocimiento para el manejo inteligente de plagas del algodón, y el tercero fue implementar novedosos conceptos de Inteligencia Artificial para el desarrollo de modelos de conocimiento. Particularmente, se desarrollaron varios artículos de investigación para cumplir con los objetivos propuestos en esta tesis. Inicialmente se realizó un artículo de revisión sobre las últimas tendencias en Agricultura Inteligente utilizando inteligencia artificial y técnicas de sensado para el manejo de plagas de insectos y enfermedades en algodón. Por otro lado, para el primer subobjetivo se realizó un artículo donde se proponía una arquitectura metacognitiva con tareas metacognitivas (meta-memoria, meta-aprendizaje, meta-razonamiento, meta-comprensión y meta-conocimiento) para la gestión inteligente de plagas. Para cumplir con el segundo subobjetivo se propusieron dos artículos. El primer artículo es un modelo de clasificación de la población del picudo algodonero y el segundo artículo presentó un sistema de clasificación difusa para analizar el rendimiento de la producción de algodón. En cuanto al tercer subobjetivo, se propusieron dos artículos. El primer artículo trata sobre un sistema con ciclos autónomos de análisis de datos para la gestión integrada del algodón. Y el segundo artículo muestra cómo mejorar la clasificación de plagas de insectos en algodón utilizando técnicas de Aprendizaje por Transferencia. En cada artículo, las estrategias/modelos se evaluaron utilizando varios conjuntos de datos. Los resultados mostraron la capacidad de las metodologías y modelos desarrollados para la toma de decisiones en el manejo inteligente de plagas de insectos en cultivos de algodón. Específicamente, nuestras propuestas permiten la predicción de los comportamientos del picudo del algodonero, el diagnóstico/predicción del rendimiento del algodón y la prescripción de estrategias para el manejo del algodón en el marco de una arquitectura metacognitiva, con buenos resultados en las métricas de rendimiento.

Palabras Clave: Inteligencia artificial, Aprendizaje automático, Modelización predictiva, Modelización prescriptiva, Agricultura inteligente, Agricultura de precisión, Algodón

Scientific contributions

Several scientific articles were generated and published during the development process of this research project.

Published articles:

- R. Toscano-Miranda, M. Toro, J. Aguilar, M. Caro, A. Marulanda, and A. Trebilcok, "Artificial-intelligence and sensing techniques for the management of insect pests and diseases in cotton: a systematic literature review", *The Journal of Agricultural Science*, vol. 160, pp. 16-31, 2022. doi:10.1017/S002185962200017X
  Q2 Scientific Journal Rankings

- R. Toscano-Miranda, W. Hoyos, M. Caro, J. Aguilar, A. Trebilcok, and M. Toro, "A Classification Model of Cotton Boll-Weevil Population", *2022 XVLIII Latin American Computer Conference (CLEI)*, pp. 1-5, 2022. doi:10.1109/CLEI56649.2022.9959893
  In IEEE Xplore

- R. Toscano-Miranda, W. Hoyos, M. Caro, J. Aguilar, A. Trebilcok, and M. Toro, "Different transfer learning approaches for insect pest classification in cotton", *Applied Soft Computing*, vol. 153, pp. 111283, 2024.
  Q1 Scientific Journal Rankings

Articles submitted to journals:

- R. Toscano-Miranda, M. Caro, J. Aguilar, A. Trebilcok, and M. Toro, "MASMC: A Metacognitive Architecture for Smart-Pest Management of Cotton", preprint submitted to *International Journal of Computational Science and Engineering*, 2023.
  Q3 Scientific Journal Rankings

- R. Toscano-Miranda, M. Caro, J. Aguilar, A. Trebilcok, and M. Toro, "A fuzzy classification system to analyze the yield of cotton production", preprint submitted to *Information Processing and Management*, 2023.
  Q1 Scientific Journal Rankings

- R. Toscano-Miranda, M. Caro, J. Aguilar, A. Trebilcok, and M. Toro, "Precision farming using autonomous data analysis cycles for integrated cotton management", preprint submitted to *Information Processing in Agriculture*, 2023.
  Q1 Scientific Journal Rankings

Project context

This study was partially funded by the Colombian Administrative Department of Science, Technology and Innovation - COLCIENCIAS and Colombian Ministry of Science and Technology *Bicentennial PhD Grant*.

Acknowledgments

This thesis is the result of a process that began several years ago. This Ph.D. research would not have been possible without the support, collaboration, and contribution of many wonderful people whom I am grateful to have in my life.

First and foremost, I thank God for His divine guidance, wisdom, and grace, which have sustained me throughout my academic journey. Without His blessings, I would not have had the strength, determination, and inspiration to complete this thesis. I am immensely grateful to my parents, Elida Miranda and Julián Toscano, for their unconditional love, support, and sacrifice. Their unwavering encouragement and belief in me have been a constant source of motivation and inspiration. To my wife Diana Quintero, I am forever grateful for her love, patience, and understanding. Her unwavering support, encouragement, and sacrifices have been instrumental in helping me pursue my academic goals while balancing my family life. I am also deeply grateful to my children Raúl and Andrés, whose boundless love, enthusiasm, and curiosity have inspired me to push through the challenges and pursue my academic dreams. I am very grateful to my siblings because they have always been there when I need them. I also want to thank my mother-in-law Guadalupe Riáscos and my nephew Guillermo González for their words of encouragement and unconditional help.

I also want to extend my sincere thanks to my supervisors José Aguilar, Mauricio Toro, and Manuel Caro, for their invaluable guidance, experience, and comments throughout this research project. Their constructive criticism, insightful comments, and patient mentoring have been instrumental in shaping my research ideas and helping me overcome obstacles. I feel blessed to have had my supervisors by my side, who have excellent human qualities. I would also like to express my sincere gratitude to Professors David Fernández and Pablo Muñoz from the University of Alcalá de Henares, Spain, for their valuable guidance and support during my research internship. I am truly grateful for their mentorship and for the opportunity to learn from them. I feel privileged to have worked with them. I also want to thank Abraham Arenas for motivating me and for his academic and research advice, which helped me to answer many questions in this research process. I also like to extend my sincere thanks to my friend William Hoyos, whose support, encouragement, and intellectual curiosity have been instrumental in shaping my research ideas and helping me stay motivated throughout this academic journey. His insightful comments and thoughtful feedback have been invaluable in refining my thesis and making it a stronger work. I am truly grateful for his friendship and support.

I also want to thank the engineers Anibal Trebilcok from the University of Córdoba, Rodolfo Alvarez from CONALGODON, and Luis Sotomayor from ICA, who supported me with their experience and gave me their guidance and wise advice.

Finally, I would like to thank Colciencias (*Bicentennial Ph.D. Grant*), the University of EAFIT, and the University of Córdoba for their financial support during my doctoral studies.

Contents

Chapter 1

Introduction and research context

1.1 Problem statement and motivation

Cotton (*Gossypium hirsutum L.*) is an economically important crop. Cotton is the main source of natural textile fiber and one of the most important oil crops [1]. Cotton contains 49 species distributed throughout most tropical and subtropical regions of the world. The world's cotton industry represents a multibillion-dollar enterprise, from the production of raw fiber to finished textile products [2]. About 25 million tons of cotton are annually produced in more than 100 countries [3]. Cotton is cultivated on about 33 million hectares around the world [3, 4].

Insect pests and diseases in cotton crops generate large economic losses. If they are not controlled in time, that is, at an early stage, they can cause an infestation, and decrease the production yield and quality of the harvested product [5]. As an example, in Brazil, annual losses in agricultural production due to insect pests can reach an average of 7.7%, equivalent to approximately US$ 17.7 billion [6]. Entomological and pathogenic problems are one of the causes of low yields and economic losses in cotton crops [7, 8].

On the other hand, Cognitive Informatics is a multidisciplinary research area that investigates the internal information-processing mechanisms of the brain and natural intelligence [9]. Cognitive Computing is an emerging paradigm of Artificial Intelligence (AI) based on Cognitive Informatics, which implements computational intelligence by autonomous inferences and perceptions, mimicking the mechanisms of the brain and natural intelligence [9]. For independence, an intelligent system should have cognitive (planning, understanding, and learning) and metacognitive processes (control and monitoring of cognitive processes) [10]. Metacognition is cognition about cognition [11] and the term metacognition in AI refers to the ability of an intelligent system to monitor and control its own learning and reasoning processes [12, 13]. Thus, metacognitive processes like monitoring, controlling, and goal setting, are related to the cognitive processes, therefore they are one of the main parts of a cognitive architecture [14]. Cognitive architecture refers to a theory about the structure of the human mind and its computational instantiation in the fields of AI and computational cognitive science [15]. Cognitive architectures offer the following advantages to intelligent agents (1) greater autonomy in decision-making [16, 17]; (2) fault tolerance since the system can identify faults and fix them without human intervention [18, 19]; (3) better response to unexpected events or to situations for which they were not designed [17].

In addition, AI has contributed to several areas; in this case, this work focuses on its application in agriculture, especially, in cotton crops. Smart agriculture (SA) plays an important role in cotton crops, including the detection and control of insect pests and diseases. SA uses the interrelationship of sensor-network technology, cloud-computing technology, and context-aware computing technology, in order to manage the agriculture process [20]. SA includes strategies of integrated pest management (IPM), which seeks to minimize the environmental impact of pesticide application and reduce risks to human and animal health [21, 22]. IPM is based on very important aspects, such as the prevention and monitoring of pests and diseases, which today are being assisted with detection equipment and AI techniques. However, there are no solutions that integrate different knowledge models of the AI (predictive, descriptive, and prescriptive models, among others) for smart management of cotton pests [23].

Since one of the problems in cotton cultivation is the control of the boll weevil [24–26], we have taken it as a case study to apply our research. Boll weevil (*Anthonomus grandis grandis*) is an insect that feeds on the squares and bolls of the cotton plant, and causes huge losses in cotton crops [27]. Recent work has demonstrated the application of IPM using AI for the control of boll weevils in cotton cultivation [28]. Cognitive architecture has been applied to the domain of agriculture. [29] presented a cognitive architecture for automatic gardening, which is composed of a decision-making framework with robotics techniques for sensing and acting to autonomously treat plants. However, nowadays, to the best of our knowledge, there is no metacognitive architecture with knowledge models applied to the management of the boll weevil in cotton crops. In the context of the identified problem, the following research question is formulated:

What would be the contributions that AI can give to a more intelligent management of cotton crops?

In accordance with the research question, this work identified and focused on the following challenges: i) develop predictive models to know the behavior of insect pests attack in cotton, ii) implement prescriptive models in smart management of cotton related to insect pests, iii) develop diagnostic models for cotton yield related to insect pests, iv) define a cotton-crop management system using a cognitive-computing architecture, v) select the most useful variables related to insect pests attack in cotton, carrying out a feature engineering process. These challenges were addressed under the approach of smart insect-pest management for cotton crops.

In summary, there is a need to develop different knowledge models for the smart management of cotton pests. Our expectation is to develop knowledge models that can be integrated, in the future, into our metacognitive architecture with autonomous tasks.

1.2 Research objectives

1.2.1 General objective

Build an intelligent system based on AI for smart insect-pest management in cotton crops.

1.2.2 Specific objectives

- Design a metacognitive architecture for the smart management of cotton pests.
- Implement knowledge models for the smart management of cotton pests.
- Implement novel AI concepts for the development of smart management systems for cotton pests.

1.3 Contributions and research scope

This research focuses on the development of models, methodologies, and computational approaches to support decision-making in the management of pests in cotton (our case study is the boll weevil). The study makes several contributions to the field, which are outlined in this section. Firstly, a systematic literature review (SLR) was conducted to identify the challenges and opportunities associated with managing insect pests and diseases in cotton [23]. The review focused on two areas of interest, namely AI and sensing techniques, which are relevant to the management of insect pests and diseases in cotton. With this in mind, we identify challenges and opportunities for future work. Based on the information reported in the SLR, we set out specific objectives related to metacognitive architectures, diagnostic, predictive and prescriptive modeling, and transfer learning.

In this sense, we design a metacognitive architecture that integrates smart agriculture technologies, knowledge models, and metacognitive functions. This integration points to a more efficient use of agricultural, and technological resources, with greater autonomy and assistance to the farmer in decision-making. We applied AI techniques on datasets related to the boll weevil in cotton crops to generate predictive models with explanatory capacity, which means, models that allow us to evaluate the behavior of included variables. Furthermore, we created models that can both diagnose and predict the yield of cotton. We also developed prescriptive models that utilize optimization techniques to suggest the most effective strategies for managing cotton. By combining these predictive, diagnostic, and prescriptive models, we established an autonomous cycle. In addition, we implemented transfer learning techniques (TL) to enhance our previous knowledge models. Transfer learning involves using knowledge from previously learned tasks to improve the performance of new tasks. In general, the data sources used to develop these knowledge models were obtained from *Colombian Agricultural Institute* (ICA in Spanish), *Colombian Cotton Confederation* (CONALGODON in Spanish), and *Institute of Hydrology, Meteorology and Environmental Studies* (IDEAM in Spanish). Data sources include capture records of the boll weevil in pheromone traps, weather data, cotton production data, and the knowledge of cotton growing and marketing experts. The case study used cotton-growing regions of Córdoba, Colombia. The models/systems were evaluated using particular scenarios advised by experts in cotton agriculture.

All the contributions made in this research are represented in several research articles. A total of six (6) scientific articles were generated, of which two (2) are published and the other four (4) are under review.

1.4 Thesis organization

This thesis is presented as a collection of articles developed to meet each of the proposed objectives. Chapter 2 describes the results of our SLR. Chapter 3, Chapter 4, and Chapter 5 correspond to the fulfillment of the first, second, and third objectives, respectively. One article was generated for the first sub-objective, two articles for the second sub-objective, and two articles for the third sub-objective. These articles will be presented in each chapter.

A brief description of each chapter is presented below. Chapter 2 describes the results of our SLR on AI and sensing techniques for the management of insect pests and diseases in cotton. This SLR allowed us to identify trends, challenges, and research opportunities in this field. Chapter 3 shows the architecture for Smart agriculture with metacognitive functions that achieve sub-objective one. Chapter 4 presents two articles to meet the second sub-objective proposed in this thesis. The first article corresponds to the development of classification models of the cotton boll-weevil population. The second article corresponds to the development of a diagnostic model for cotton yield management. Chapter 5 presents two articles, in the first one, we propose the development of an autonomous cycle of data analysis tasks for integrated cotton management, and in the other one, we propose TL techniques to improve the accuracy of the predictive/classification models developed in sub-objective two. Finally, Chapter 6 presents a summary of the conclusions of all the articles presented in the previous sections. We also show the limitations of our research and possible future work.

Chapter 2

State of the art on smart insect-pest management for cotton crops

2.1 Motivation

In this chapter, we present an SLR that aims to provide a comprehensive overview of the recent advances and applications of AI and sensing techniques for the management of insect pests and diseases in cotton. At the time of the start of the research project, there was no review in the literature that encompassed this approach in agriculture. The literature review highlights the potential of AI and sensing techniques to help the farmer in decision-making related to insect pests and diseases in cotton. Results point out that AI techniques with remote and field sensing techniques have a wide potential in the detection and diagnosis of diseases and insect pests in cotton. However, existing challenges to face, and therefore, we propose challenges such as i) develop predictive models to know the behavior of insect pests attack in cotton, ii) implement prescriptive models in smart management of cotton related to insect pests, iii) develop diagnostic models for cotton yield related to insect pests, iv) define a cotton-crop management system using a cognitive-computing architecture, v) select the most useful variables related to insect pests attack in cotton using feature engineering. The article about the SLR is in Appendix A.

2.2 Identification of the article

R. Toscano-Miranda, M. Toro, J. Aguilar, M. Caro, A. Marulanda, and A. Trebilcok, "Artificial-intelligence and sensing techniques for the management of insect pests and diseases in cotton: a systematic literature review", *The Journal of Agricultural Science*, vol. 160, pp. 16-31, 2022, doi:10.1017/S002185962200017X.

2.3 Abstract

Integrated pest management (IPM) seeks to minimize the environmental impact of pesticide application and reduce risks to human and animal health. IPM is based on two important aspects -prevention and monitoring of diseases and insect pests- which today are being assisted by sensing and artificial-intelligence (AI) techniques. In this paper, we surveyed the detection and diagnosis, with AI, of diseases and insect pests, in cotton, which have been published between 2014 and 2021. This research is a systematic literature review. The results show that AI techniques were employed -mainly- in the context of (i) classification, (ii) image segmentation, and (iii) feature extraction. The most used algorithms, in classification, were support vector machines, fuzzy inference, back-propagation neural-networks and recently, convolutional neural networks; in image segmentation, k-means was the most used; and, in feature extraction, histogram of oriented gradients, partial least-square regression, discrete wavelet transform, and enhanced particle-swarm optimization were equally used. The most used sensing techniques were cameras, and field sensors such as temperature and humidity sensors. The most investigated insect pest was the whitefly, and the disease was root rot. Finally, this paper presents future works related to the use of AI and sensing techniques, to manage diseases and insect pests, in cotton; for instance, implement diagnostic, predictive and prescriptive models to know when and where the diseases and insect pests will attack and make strategies to control them.

2.4 Link to the full article

https://doi.org/10.1017/S002185962200017X

Chapter 3

Metacognitive Architecture for Smart-Pest Management

3.1 Motivation

Cognitive architectures are important because they provide a framework for building smart systems that can mimic humans to adapt to new situations and learn from experience. These architectures are designed to handle different types of tasks, such as learning, reasoning, problem-solving, and decision-making. They can be based on diagnostic, predictive, and prescriptive models that can be used to improve performance in a wide range of fields, from healthcare to agriculture. We designed a cognitive architecture for smart-pest management of cotton with metacognitive tasks, and a case study of the architecture for predictive and prescriptive problems in the context of integrated pest management in cotton. Thus, in this chapter, we present the paper for the fulfillment of the first objective (see the article in Appendix B).

3.2 Identification of the article

R. Toscano-Miranda, M. Caro, J. Aguilar, A. Trebilcok, and M. Toro, "A Smart-Pest Management of Cotton based on a Metacognitive Architecture", preprint submitted to *International Journal of Computational Science and Engineering.*

3.3 Abstract

The use of information technology in agriculture plays an important role in Smart-Pest Management. Particularly, Artificial intelligence helps to identify, monitor, control and make decisions about pests in

crops. In this paper, we present a new metacognitive architecture, called Metacognitive Architecture for a Smart-Pest Management of Cotton. Especially, this paper presents several contributions: (1) a new architecture that implements several metacognitive tasks (meta-memory, meta-learning, meta-reasoning, meta-comprehension, meta-knowledge); (2) a case study of the architecture for predictive and prescriptive problems, in the context of integrated pest management in cotton; (3) an integrated approach of data analytics and metacognition in smart systems.

3.4 Link to the full article

Appendix B

Chapter 4

Knowledge models for the smart management of cotton pests

4.1 Motivation

Knowledge models (KMs) are important for the smart management of cotton pests for several reasons: i) Pest identification: KMs can provide accurate and up-to-date information about cotton pests. They can help farmers and agronomists identify behaviors. This knowledge is crucial for implementing effective pest management strategies. ii) Early detection: KMs can assist in the early detection of cotton pests. By analyzing data such as weather conditions, and pest occurrence patterns, they can identify signs of pest infestations at an early stage. Early detection enables timely intervention, reducing the risk of widespread damage and improving pest control. iii) Decision support: cotton pest management involves making complex decisions, such as selecting appropriate control measures, determining the optimal time for treatments, and evaluating economic expenses. KMs can provide decision support by analyzing various factors and recommending suitable management strategies based on current conditions and best practices. This helps farmers optimize their pest control efforts and minimize costs. In summary, this chapter shows how the KMs for cotton pest management enhance pest identification, enable early detection, and provide information for decision support, among other things. Thus, in this chapter, we present two articles to fulfill the second objective. The first article presents one KM for the classification of the boll-weevil population in cotton crops. The second article presents a second KM to analyze the yield of cotton production.

4.2 A Classification Model of Cotton Boll-Weevil Population

4.2.1 Motivation

The motivation for a classification model of the cotton boll-weevil population is to help farmers and agronomists identify the behavior of the pests and assist in the early detection of cotton pests. By analyzing data such as weather conditions, and pest occurrence patterns, the classification model can identify signs of pest infestations at an early stage. Also, it can aid in monitoring cotton pests by analyzing data from various sources such as remote sensing and pest trap records. With this analysis, the KM can predict potential outbreaks. Thus, in this section, we present the first article to fulfill the second objective. The complete article can be found in Appendix C.

4.2.2 Identification of the article

R. Toscano-Miranda, W. Hoyos, M. Caro, J. Aguilar, A. Trebilcok, and M. Toro, "A Classification Model of Cotton Boll-Weevil Population", *2022 XVLIII Latin American Computer Conference (CLEI)*, pp. 1-5, 2022. doi:10.1109/CLEI56649.2022.9959893
In IEEE Xplore

4.2.3 Abstract

Integrated pest management (IPM) seeks to minimize the environmental impact of pesticide application. IPM is based on two important aspects -prevention and monitoring of diseases and insect pests- which today are being assisted by sensing and artificial-intelligence (AI). Particularly, AI helps to identify, monitor, control and make decisions about pests in crops. In this paper, we present a comparison among five machine-learning models to classify the population of the boll weevil in cotton into three classes: low, medium, and high. Weather data (average daily rainfall, humidity, and temperature) were used to classify the population of the boll weevil in the department of Córdoba, Colombia. The results showed that XGBoost obtained the highest accuracy (88%). Results showed that it is possible to classify boll-weevil populations using weather data.

4.2.4 Link to the full article

https://doi:10.1109/CLEI56649.2022.9959893

4.3 A fuzzy classification system to analyze the yield of cotton production

4.3.1 Motivation

The motivation for a fuzzy classification system to analyze the yield of cotton production is to provide decision support by studying various factors. This allows the integration of diverse data sources and provides a holistic view of pest management. The fuzzy classification system considers factors such as pheromone control, cultural practices, and pesticide usage to develop sustainable and environmentally friendly pest management plans. Particularly, this KM uses the classification model resulting from the previous study (classification of the weevil population), and then, by incorporating new input variables (crop stage, rainfall, fertilizer, pheromone traps, boll-weevil killing tube), the KM diagnoses the crop yield. Thus, in this section, we present the second article to fulfill the second objective. The complete article can be found in Appendix D.

4.3.2 Identification of the article

R. Toscano-Miranda, M. Caro, J. Aguilar, A. Trebilcok, and M. Toro, "A fuzzy classification system to analyze the yield of cotton production", preprint submitted to *Information Processing and Management*, 2023.

4.3.3 Abstract

Properly managing the cultivation of cotton is essential because it directly impacts the amount of cotton that is produced, making it a highly significant task. The aim of this work is the proposal of a fuzzy classification system for diagnosis-prediction tasks of the cotton crop yield. We used a soft computing method to handle/describe experts' knowledge. Seven input variables (attack level of the red boll weevil, attack level of the black boll weevil, crop stage, rainfall, fertilizer, pheromone traps, and boll-weevil killing tube) were considered in the system to analyze the cotton production. System tests were carried out on different agricultural scenarios, to determine their robustness and adaptability. According to the results, the fuzzy system has the capability to generate outputs that correspond with the experts' evaluations, which can be used to help farmers select the best practices in cotton crop management, in order to obtain the best yield in a specific context. The developed models enhance our capacity to predict crop yields based on climate data, the soil and pest behaviors, a valuable indicator for decision-making and overall sustainability.

4.3.4 Link to the full article

Appendix D

Chapter 5

Novel AI concepts for the development of smart management systems for cotton pests

5.1 Motivation

This chapter integrates previous KMs using a new paradigm known as autonomous cycles of data analysis tasks. It allows an integrated pest management (IPM): IPM is an approach that emphasizes the effective management of pests while minimizing environmental impact through a combination of strategies. The autonomous cycles play a vital role in IPM by integrating diverse data sources and KMs, providing a holistic view of pest management. Furthermore, this chapter explores the application of TL techniques to enhance KMs. In summary, this chapter shows the integration of KMs for cotton pest management and the improvement of the KMs using TL. Thus, in this chapter, we present two articles to fulfill the third sub-objective. The first article presents the integration of several KMs using autonomous cycles of data analysis tasks for integrated cotton management. The second article presents TL techniques for improving previous KMs about the classification of the boll-weevil population.

5.2 Precision farming using autonomous cycles of data analysis tasks for integrated cotton management

5.2.1 Motivation

The motivation for "Precision Farming using autonomous data analysis cycles for integrated cotton management" is to implement integrated management with features such as pest identification, early detection, monitoring, and decision support, managed by an autonomous cycle. This approach integrates all previously implemented KM as data analysis tasks, such that can assist in the early detection of cotton pests, can identify signs of pest infestations at an early stage, and can carry out complex decisions, such as selecting appropriate control measures, determining the optimal time for treatments, and evaluating economic expenses. In this sense, the autonomous cycle uses diverse data sources, so that their KMs can autonomously monitor, analyze and make decisions about the cotton crop to provide holistic view of pest management. The autonomous cycle monitors the crop conditions, analyzes the data, predicts the crop yield, and recommends the best strategy to manage the crop. This information helps stakeholders to efficiently allocate resources, guide interventions and prevent extensive pest damage. Thus, in this section, we present the first article to fulfill the third sub-objective. The complete article can be found in Appendix E.

5.2.2 Identification of the article

R. Toscano-Miranda, M. Caro, J. Aguilar, A. Trebilcok, and M. Toro, "Precision farming using autonomous data analysis cycles for integrated cotton management", preprint submitted to *Information Processing in Agriculture*, 2023.

5.2.3 Abstract

Precision farming (PF) allows the efficient use of resources such as water, and fertilizers, among others; as well, it helps to analyze the behavior of insect pests, in order to increase production and decrease the cost of crop management. This paper introduces an innovative approach to integrated cotton management, involving the implementation of an Autonomous Cycle of Data Analysis Tasks (ACODAT). The proposed autonomous cycle is composed of a classification task of the population of pests (boll weevil) (based on XGBoost), a diagnosis-prediction task of cotton yield (based on a fuzzy system), and a prescription task of strategies for the adequate management of the crop (based on genetic algorithms). The proposed system can evaluate several variables according to the conditions of the crop, and recommend the best strategy for getting increase the cotton yield. In

particular, the classification task has an accuracy of 88%, the diagnosis/prediction task obtained a 98% of accuracy, and the genetic algorithm recommends the best strategy for the context analyzed. Focused on integrated cotton management, our system offers flexibility and adaptability, which facilitates the incorporation of new tasks.

5.2.4 Link to the full article

Appendix E

5.3 Different transfer learning approaches for insect pest classification in cotton

5.3.1 Motivation

The motivation for ”Improving Insect Pest Classification in Cotton Using Transfer Learning Techniques” is to improve the quality of the KM for classification. Given a KM that is not attaining the desired precision, TL techniques allow us to improve it by reusing previous KMs which obtained better results. Particularly, the quality of the KM by regions in Córdoba is very different due to the data available of each one. These data have information on captures of the boll weevil in pheromone traps. Three types of TL techniques, namely instance-, feature-, and parameter-based, were used. These techniques help when there are few instances or characteristics in the training data. In this way, we increased the number of instances and/or characteristics to train the models, when necessary. TL techniques were used to improve the lowest results in some regions. In this sense, TL is useful for transferring knowledge learned in one context, then applying it in different but related contexts to improve the results in these contexts. Thus, in this section, we present the second article to fulfill the third sub-objective. The complete article can be found in Appendix F.

5.3.2 Identification of the article

R. Toscano-Miranda, W. Hoyos, M. Caro, J. Aguilar, A. Trebilcok, and M. Toro, “Different transfer learning approaches for insect pest classification in cotton”, *Applied Soft Computing*, vol. 153, pp. 111283, 2024.

5.3.3 Abstract

Boll weevil is an important pest that affects cotton crops worldwide, causing significant economic losses. The classification of the boll-weevil population is crucial for developing effective pest management strategies. However, the limited availability of data and features makes classification a challenging task. This study aimed to investigate the use of Transfer Learning (TL) techniques to improve the classification of boll weevil populations. Three types of TL techniques, instance-based, feature-based, and parameter-based, were studied to improve the classification performance of the machine learning algorithms. This work used data from two domains, one with a limited number of instances and the other with a limited number of features, to test the proposed approach. Climate variables were incorporated as features to predict the level of the boll-weevil attack. The proposed approach achieved significant improvements in classification accuracy for both the limited instances and limited feature domains. The case with few instances initially, reached an accuracy of 90.79%, while the case with few features reached an accuracy of 96.28%. The results demonstrate the effectiveness of TL techniques in improving the classification of boll-weevil populations in cotton crops when there is a limited amount of data and/or features.

5.3.4 Link to the full article

Appendix F

Chapter 6

Conclusions

This thesis made contributions on smart insect-pest management for cotton crops. In this chapter, we present a summary of the results of all the work presented above. In addition, we show limitations and research opportunities for the future.

6.1 Summary

The integration of information technology, specifically artificial intelligence, in agriculture is crucial for the implementation of Smart-Pest Management systems. This study introduces a novel metacognitive architecture, named Metacognitive Architecture for a Smart-Pest Management of Cotton, which addresses various metacognitive tasks, including meta-memory, meta-learning, meta-reasoning, meta-comprehension, and meta-knowledge. The contributions of this study are threefold. Firstly, it presents a new architecture that incorporates multiple metacognitive tasks, providing a comprehensive framework for Smart-Pest Management. Secondly, a case study is conducted to demonstrate the effectiveness of the proposed architecture in addressing predictive and prescriptive problems within the context of integrated pest management in cotton. Lastly, the study emphasizes the importance of integrating data analytics with metacognitive capabilities in the development of intelligent systems for agriculture. By leveraging the capabilities of artificial intelligence and metacognition, this study significantly advances the field of Smart-Pest Management by enabling the identification, monitoring, control, and decision-making processes related to pests in agricultural crops. The integrated approach presented in this study contributes to sustainable and efficient agricultural systems.

In addition, we developed a KM to classify the population of the boll weevil in cotton. The classification model utilized information on pheromone traps and sensors. The sensors got weather data such as average daily rainfall, humidity, and temperature from regions of the department of Córdoba, Colombia. Also, this study proposed a fuzzy classification system for the diagnosis and prediction of cotton crop yield, employing a soft computing method to leverage expert knowledge. The system incorporates seven input variables, including the attack levels of the red and black boll weevils, crop stage, rainfall, fertilizer usage, pheromone trap data, and boll-weevil killing tube information. By analyzing these variables, the system provides insights into cotton production. Rigorous testing was conducted across diverse agricultural scenarios to assess the system's robustness and adaptability. The results indicate that the fuzzy system is capable of generating outputs that align with expert evaluations. This capacity enables farmers to make informed decisions and select the most suitable practices for cotton crop management in specific contexts, thereby optimizing yield. By employing a fuzzy classification system and leveraging expert knowledge, this research significantly contributes to improving cotton crop management. The developed models enhance our ability to predict crop yields by incorporating climate data, soil conditions, and pest behaviors. This information serves as a valuable indicator for decision-making and contributes to overall sustainability in cotton cultivation.

On the other hand, this study introduces an autonomous cycle of data analysis tasks for integrated cotton management. The autonomous cycle for integrated cotton management consists of three main tasks: a classification task for population estimation of boll weevils using XGBoost, a diagnosis-prediction task for cotton yield utilizing a fuzzy system, and a prescription task employing genetic algorithms to recommend optimal crop management strategies. The system evaluates various variables based on crop conditions and provides recommendations for maximizing cotton yield. The classification task achieves an accuracy of 88%, demonstrating its effectiveness in estimating the population of boll weevils. The diagnosis task exhibits a high accuracy of 98%, enabling an accurate analysis of cotton yield. Furthermore, the genetic algorithm effectively recommends the best strategy for the specific context analyzed. Focused on integrated cotton management, this system offers flexibility and adaptability, facilitating the inclusion of new tasks as needed. By leveraging advanced technologies and data analysis, the proposed approach optimizes resource allocation, enhances decision-making, and ultimately, increases cotton yield. The autonomous cycle for integrated cotton management presented in this paper represents a significant advancement in precision farming and cotton management. Its accuracy in estimating pest populations and predicting yield, and recommendation capabilities contribute to the efficient and sustainable management of cotton crops.

Finally, as was mentioned before, the classification of boll weevil populations is a critical task for effective pest management in cotton crops, considering the significant economic losses caused by this pest worldwide. However, limited data availability and features pose challenges to accurate classification. This study investigated the application of TL techniques to enhance the classification performance of boll weevil populations. Three types of TL techniques, namely instance-based, feature-based, and parameter-based, were studied to improve the performance of machine learning algorithms in classification. The proposed approach analyzed two data problems: one with a limited number of instances and another with a limited number of features. The results showcased significant improvements in classification accuracy for both domains with limited instances and limited features. In the case of limited instances, the proposed approach achieved an accuracy of 90.79%, while in the case of limited features, the accuracy reached 96.28%. These outcomes demonstrate the effectiveness of TL techniques in enhancing the classification of boll weevil populations in cotton crops, especially when data and/or features are scarce. Ultimately, these advancements contribute to the development of more efficient and sustainable approaches to combat boll weevil infestations in cotton crops.

6.2 Limitations and future work

With the results of the present thesis, we were able to meet the proposed objectives. However, this thesis had some limitations in terms of region data, the models, and of the variables that were missing to incorporate, which are summarized below. Regarding the data of the regions, this study included several cotton regions of Córdoba, other regions were left out of the study due to the lack of sensors. The data was collected from sources with few sensors. It would be very useful to increase the number of sensors to have more accurate data from the region. On the other hand, it would be important to incorporate sensors that can record more climatic variables. The use of more climatic variables would enrich the KMs. This study only included regions of Córdoba. If all the cotton-growing regions of Colombia are incorporated, then a much more complete behavior of the weevil could be determined, and better control of the insect in the country obtained. With a greater number of sensors and pheromone traps in all cotton-growing regions, then the models would have more complete information to predict the spread of insects. Finally, pheromone traps are not electronic with real-time data collection. If electronic traps with the ability to read the data in real-time are used, then the KMs predictions would have greater precision.

On the other hand, regarding the models, for crop yield predictions, only the boll weevil was considered in the case study. If information about other insects and diseases is included, then it is possible to have a more complete vision of the attacks that crops have. This would allow, for example, to develop multi-detection models of diseases or pest attacks.

Other limitation of the current study is regarding the variables that were missing to incorporate. The current study refers to the general use of the amount of fertilizer. We did not include each specific fertilizer class. This could be overcome by specifying the amount of each fertilizer and adding this information to the model. Also, this study did not include other environmental variables such as soil organic matter, weed coverage percentage, and tillage system management, which can be considered in future works. The analysis could also include other variables such as: the rainfall distribution, the mean temperatures (day/night gradient), the amount of solar radiation, the density of plants in cotton cultivation, the date of planting, the timely weed control, and so on.

Finally, according to our metacognitive architecture, sensor information (Ground Level) was used and KMs (Object Level) were implemented. Future works would be oriented to the implementation of the metacognitive functions on the Meta Level. In this sense, this allows AI systems to reason and adapt to the situation with self-awareness. Also, we plan to integrate our metacognitive architecture with the multi-agent systems paradigm, to take advantage of the existing modeling capabilities and implementations in agent theory.

References

[1] *Method for monitoring the cotton plant vigor based on the WSN technology*, Computers and Electronics in Agriculture **133**, 68–79 (2017).

[2] C. W. Smith and J. T. Cothren, *Cotton: Origin, History, Technology, and Production*, John Wiley & Sons (1999).

[3] M. Azam, A. Wahid, M. Ahmad, M. Tayab, M. Ahmed, S. Ahmad, and M. Hasanuzzaman. *World cotton production: An overview.* In *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies*, chapter 1, pages 1–7. (2020).

[4] J. Carvalho, L. Ruiz, F. FAO, A. Organization, and Gobierno de Brasil, *Cooperación sur-sur trilateral estudio nichos de mercados del algodón* (2018).

[5] N. El-Wakeil and A. M. Abdallah, *Cotton pests and the actual strategies for their management control* (2014).

[6] C. M. Oliveira, A. M. Auad, S. M. Mendes, and M. R. Frizzas, *Crop losses and the economic impact of insect pests on Brazilian agriculture*, Crop Protection **56**, 50–54 (2014).

[7] M. Anees and S. A. Shad. *Insect pests of cotton and their management.* In *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies*, chapter 11, pages 177–212. Springer, (2020).

[8] S. Chohan, R. Perveen, M. Abid, M. Nouman, and M. Sajid. *Chapter 13 Cotton Diseases and Their Magement.* In *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies*, chapter 13, pages 239–270. Springer, (2020).

[9] *Cognitive informatics: Towards cognitive machine learning and autonomous knowledge manipulation*, International Journal of Cognitive Informatics and Natural Intelligence **12**, 1–13 (2018).

[10] M. T. Cox, *Perpetual self-aware cognitive agents*, AI Magazine **28**, 32–45 (2007).

[11] M. T. Cox, *Metacognition in computation: A selected research review*, Artificial Intelligence **169**, 104–141 (2005).

[12] M. F. Caro, D. P. Josyula, M. T. Cox, and J. A. Jiménez, *Design and validation of a metamodel for metacognition support in artificial intelligent systems*, Biologically Inspired Cognitive Architectures **9**, 82–104 (2014).

[13] M. T. Cox and A. Raja, *Metareasoning: An Introduction*, Metareasoning , 3–14 (2013).

[14] R. Sun, X. Zhang, and R. Mathews, *Modeling meta-cognition in a cognitive architecture*, Cognitive Systems Research **7**, 327–338 (2006).

[15] A. Lieto, *Cognitive design for artificial minds*, Routledge (2021).

[16] M. T. Cox and Z. A. Dannenhauer, *Perceptual Goal Monitors for Cognitive Agents in Changing Environments*, Advances in Cognitive Systems , 1–16 (2017).

[17] M. Paisner, M. T. Cox, M. Maynord, and D. Perlis. *Goal-Driven Autonomy for Cognitive Systems.* In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, (2014).

[18] M. T. Cox and A. Raja, *Metareasoning: Thinking about Thinking* (2011).

[19] C. M. Kennedy and A. Sloman. *Reflective Architectures for Damage Tolerant Autonomous Systems.* Technical report, University of Birmingham, School of Computer Science, (2002).

[20] Aqeel-Ur-Rehman and Z. A. Shaikh, *Smart agriculture*, Applications of Modern High Performance Networks , 120–129 (2009).

[21] G. Vétek, A. Timus, M. Chubinishvili, G. Avagyan, V. Torchan, Z. Hajdú, A. Veres, and A. Nersisyan, *Integrated pest management of major pests and diseases in eastern Europe and the Caucasus*, Food and Agriculture Organization of the United Nations (2017).

[22] D. Simberloff and M. Rejmanek, *Encyclopedia of biological invasions*, Univ of California Press, USA (2011).

[23] R. Toscano-Miranda, M. Toro, J. Aguilar, M. Caro, A. Marulanda, and A. Trebilcok, *Artificial-intelligence and sensing techniques for the management of insect pests and diseases in cotton: a systematic literature review*, The Journal of Agricultural Science **160**, 16–31 (2022).

[24] E. Ben Guerrero, R. Salvador, and P. Talia, *Evaluation of hydrolytic enzyme activities from digestive fluids of Anthonomus grandis (Coleoptera: Curculionidae)*, Archives of Insect Biochemistry and Physiology **105**, 1–8 (2020).

[25] R. R. Coelho, J. D. A. de Souza Júnior, A. A. Firmino, L. L. de Macedo, F. C. Fonseca, W. R. Terra, G. Engler, J. de Almeida Engler, M. C. M. da Silva, and M. F. Grossi-de Sa, *Vitellogenin knockdown strongly affects cotton boll weevil egg viability but not the number of eggs laid by females*, Meta Gene **9**, 173–180 (2016).

[26] J. F. Grigolli, L. A. Souza, M. G. Fernandes, and A. C. Busoli, *Spatial Distribution of Adult Anthonomus grandis Boheman (Coleoptera: Curculionidae) and Damage to Cotton Flower Buds Due to Feeding and Oviposition*, Neotropical Entomology **46**, 442–451 (2017).

[27] H. C. Ellis and D. L. Horton. *Boll Weevil*, (1997).

[28] A. N. Alves, W. S. Souza, and D. L. Borges, *Cotton pests classification in field-based images using deep residual networks*, Computers and Electronics in Agriculture **174**, 105488 (2020).

[29] A. Agostini, G. Alenyà, A. Fischbach, H. Scharr, F. Wörgötter, and C. Torras, *A cognitive architecture for automatic gardening*, Computers and Electronics in Agriculture **138**, 69–79 (2017).

Appendix A

Artificial-intelligence and sensing techniques for the management of insect pests and diseases in cotton: a systematic literature review

The Journal of Agricultural Science

cambridge.org/ags

Artificial-intelligence and sensing techniques for the management of insect pests and diseases in cotton: a systematic literature review

Crops and Soils Review

Cite this article: Toscano-Miranda R, Toro M, Aguilar J, Caro M, Marulanda A, Trebilcok A (2022). Artificial-intelligence and sensing techniques for the management of insect pests and diseases in cotton: a systematic literature review. *The Journal of Agricultural Science* **160**, 16–31. https://doi.org/10.1017/S002185962200017X

Received: 1 July 2021
Revised: 1 March 2022
Accepted: 28 March 2022

Key words:
Image processing; Internet of Things; machine learning; pest detection; precision agriculture; remote sensing; smart agriculture

Author for correspondence:
R. Toscano-Miranda,
E-mail: rtoscano@correo.unicordoba.edu.co

R. Toscano-Miranda[1], M. Toro[2], J. Aguilar[2,3,4], M. Caro[1], A. Marulanda[5] and A. Trebilcok[6]

[1]Department of Educational Informatics, Universidad de Córdoba, Montería, Colombia; [2]GIDITIC, Department of Informatics and Systems, Universidad Eafit, Medellín, Colombia; [3]CEMISID, Universidad de Los Andes, Mérida, Venezuela; [4]Dpto. de Automática, Universidad de Alcalá, Alcalá de Henares, Spain; [5]Department of Physics Engineering, Universidad Eafit, Medellín, Colombia and [6]Department of Agronomic Engineering, Universidad de Córdoba, Montería, Colombia

Abstract

Integrated pest management (IPM) seeks to minimize the environmental impact of pesticide application, and reduce risks to human and animal health. IPM is based on two important aspects – prevention and monitoring of diseases and insect pests – which today are being assisted by sensing and artificial-intelligence (AI) techniques. In this paper, we surveyed the detection and diagnosis, with AI, of diseases and insect pests, in cotton, which have been published between 2014 and 2021. This research is a systematic literature review. The results show that AI techniques were employed – mainly – in the context of (i) classification, (ii) image segmentation and (iii) feature extraction. The most used algorithms, in classification, were support vector machines, fuzzy inference, back-propagation neural-networks and recently, convolutional neural networks; in image segmentation, *k*-means was the most used; and, in feature extraction, histogram of oriented gradients, partial least-square regression, discrete wavelet transform and enhanced particle-swarm optimization were equally used. The most used sensing techniques were cameras, and field sensors such as temperature and humidity sensors. The most investigated insect pest was the whitefly, and the disease was root rot. Finally, this paper presents future works related to the use of AI and sensing techniques, to manage diseases and insect pests, in cotton; for instance, implement diagnostic, predictive and prescriptive models to know when and where the diseases and insect pests will attack and make strategies to control them.

Introduction

Cotton (*Gossypium hirsutum* L.) is an economically important crop. Cotton is the main source of natural textile fibre, and one of the most important oil crops (Zhang *et al.*, 2017). Cotton contains 49 species distributed throughout the most tropical and subtropical regions of the world. The world's cotton industry represents a multibillion-dollar enterprise, from the production of raw fibre to finished textile products (Smith and Cothren, 1999). Between 2016 and 2017, 32.4 million hectares were planted in more than 80 countries (Carvalho *et al.*, 2018).

Diseases and insect pests, in cotton, generate large economic losses. If they are not controlled in time, that is, at an early stage, they can cause an infestation, and decrease production yield and quality of the harvested product (El-Wakeil and Abdallah, 2014). As an example, in Brazil, annual losses, in agricultural production, due to insect pests, can reach an average of 7.7%, equivalent to approximately US$ 17.7 billion (Oliveira *et al.*, 2014). Entomological and pathogenic problems are one of the causes of low yields and economic losses in cotton (Anees and Shad, 2020; Chohan *et al.*, 2020).

One of the ways to control diseases and insect pests is through agrochemicals, but this (i) increases production costs and (ii) generates a negative impact on the environment. This is why integrated pest management (IPM) seeks to minimize the environmental impact of pesticide application, and reduce risks to human and animal health (Simberloff and Rejmanek, 2011; FAO, 2017). IPM is based on two important aspects – prevention and monitoring of diseases and insect pests – which today are being assisted by sensing and artificial-intelligence (AI) techniques. Research on object-recognition and computer vision has led to advances in factory automation, assembly-line industrial inspection systems and medical imaging (Andreopoulos and Tsotsos, 2013). In the detection of diseases and insect pests, computer vision has led to advances in the development of precision agriculture (Solis-Sánchez *et al.*,

Table 1. Summary of reviews related to this SLR

Article	Objective
Patrício and Rieder (2018)	Grain crops: disease detection, grain quality and phenotyping using computer vision and AI.
Zhang *et al.* (2019)	Monitoring plant diseases and pests through remote-sensing technology.
Boissard *et al.* (2013)	Image processing for identification of agricultural pests on various crops.
Bannerjee *et al.* (2018)	AI in agriculture.
Iqbal *et al.* (2018)	Citrus plant: detection and classification of diseases using image processing techniques.

2009; Patrício and Rieder, 2018; Habib *et al.*, 2020). In addition to computer vision, sensors (e.g. temperature sensors, soil and moisture sensors) allow data acquisition for analysis and predictions (Pratheepa *et al.*, 2016).

For the analysis of diseases and insect pests, expert systems have also helped to make better decisions and to assist farmers to prevent or control diseases and insect pests (Boissard *et al.*, 2008; Abu-Nasser and Abu-Naser, 2018; Alzamily, 2018). Expert systems contributed to improve the productivity and quality of the cultivated products, which is related to the objectives of smart agriculture (SA): (i) to increase productivity, (ii) to improve food security, (iii) to improve adaptation and resilience to climate change and variability and (iv) to reduce greenhouse gas emissions (FAO, 2013).

SA uses the interrelationship of (i) sensor-network, (ii) grid-computing and (iii) context-aware computing to manage the agriculture process. SA allows decisions to be made based on the acquisition of data from the agricultural context (Aqeel-Ur-Rehman and Shaikh, 2009; García *et al.*, 2020). Some of the technologies used in SA are robotic automation, data analytics and remote sensing (Grady *et al.*, 2019). AI techniques are employed in SA for the following tasks (Bannerjee *et al.*, 2018): (i) general crop management, (ii) pest management, (iii) disease management, (iv) agricultural product monitoring and storage control, (v) soil and irrigation management, (vi) weed management and (vii) yield prediction. In this systematic literature review (SLR), we focus on pest management.

Some SLRs have been made in relation to pest management, but none – specifically – to cotton. Patrício and Rieder (2018) conducted an SLR of computer vision and AI, in precision agriculture, for grain crops (maize, rice, wheat, soybean and barley), and aspects related to disease detection, grain quality and phenotyping. Zhang *et al.* (2019) conducted an SLR on monitoring plant diseases and pests through remote-sensing technology, not including *in-situ* sensors: the main topics were sensing technologies and feature extraction. The previous study included four papers related to cotton between 2004 and 2011. Boissard *et al.* (2013) conducted a brief review of the application of image processing to identify agricultural pests on various crops not including cotton. Bannerjee *et al.* (2018) conducted a literature survey on AI, in agriculture, in general. For cotton cultivation, they included three papers, in the category of crop management, before 1989; and one paper, on yield prediction, in 2008. Finally, Iqbal *et al.* (2018) made an SLR of automated detection and classification of citrus-plant diseases using image-processing techniques. Table 1 shows the summary of these reviews. According to the above, an SLR for cotton is needed (i) to analyse articles that used AI techniques to manage diseases and insect pests, in cotton; (ii) to know the most recent state-of-the-art, given the continuous advances in the area and (iii) to identify future work directions.

The objective of this SLR is to establish the state-of-the-art research on the management (detection, prediction, diagnosis and prescription), of diseases and insect pests, in cotton. The main contributions of this article are the following. Firstly, a description of the cotton diseases and insect pests was investigated through the use of AI and sensing techniques, from 2014 to 2021. Secondly, an analysis of the selected papers. Finally, a definition and discussion of the current challenges on AI techniques, for pest and disease management, in cotton.

Background

In what follows, a background of cotton diseases and insect pests is presented.

Cotton insect pests

Insects are classified as pests when the damage they cause decreases the yield of the farmer's products (Dent and Binks, 2020). The insect pests described as follows were selected due to the severity of their damage and greater presence (Presley, 1954; Carpenter, 1983; Ñañez, 2012), or because they have been much studied.

*Boll weevil (*Anthonomus grandis*)*

Boll weevil is the main pest in cotton around the world, directly affecting cotton production (Coelho *et al.*, 2016; Grigolli *et al.*, 2017; Ben Guerrero *et al.*, 2020). Adults feed on fruiting forms, leaf petioles and terminal growth (Ellis and Horton, 1997).

*Whitefly (*Bemisia tabaci*)*

Whitefly infests cotton and many other plants, for example, tomato, soybean, paprika and rose (Martin *et al.*, 2008; Xia, 2012; Barbedo, 2014; El-Wakeil and Abdallah, 2014; Xia *et al.*, 2014). The whitefly reduces the performance, or even kills the plant, by feeding on the sap. In addition, the whitefly can also transport viruses (El-Wakeil and Abdallah, 2014). The whitefly is one of the most prominent insect pests and it is present in two stages of cotton: growing and fruiting.

*Thrips (*Thrips tobacco*)*

Thrips can occur in plants such as cotton, tomato, avocado, broccoli and lettuce (Solis-Sánchez *et al.*, 2011; Xia *et al.*, 2014; Shahzadi *et al.*, 2016). Thrips' damage can stunt growth which impacts crop performance (El-Wakeil and Abdallah, 2014). The thrips infest in the seedling stage.

Bollworms

There are two types: (i) American bollworm (*Helicoverpa armigera*) and (ii) pink bollworm (*Pectinophora gossypiella*).

Table 2. Research question and search queries of this SLR

Research question	Search query
RQ1 – How is AI used to manage diseases and insect pests of cotton?	('artificial intelligence' OR 'machine learning' OR 'computer vision') AND ('cotton crop' OR 'cotton farming' OR 'cotton yield') AND (disease OR pest OR insect)
RQ2 – How are sensing techniques used to detect diseases and insect pests of cotton?	('remote sensing' OR 'wireless sensor networks') AND ('cotton crop' OR 'cotton farming' OR 'cotton yield') AND (disease OR pest OR insect)

Bollworms can infest cotton, tomato and okra. In a severe infestation, bollworms may cause high damage to the plant. The larvae feed on cotton boll in the fruiting stage (El-Wakeil and Abdallah, 2014).

Cotton diseases

Diseases may be caused by fungus (mainly), bacteria or nematodes (Presley, 1954; Carpenter, 1983; Ñañez, 2012). The most important diseases based on the damage they do to cotton are described below.

Cotton diseases with the most damage

*Bacterial blight (*Xanthomonas campestris *pv.* Malvacearum*)*
The damage of bacterial blight disease is that the leaf veins blacken causing a 'blighting' appearance, causing defoliation and rotting. It can infect all growth stages of cotton, and can quickly spread to other areas of the field through wind-driven rain or irrigation (Cox *et al*., 2019).

*Fusarium wilt (*Fusarium oxysporum *f. sp.* Vasinfectum*)*
Damage caused by fusarium wilt disease includes brown discolouration of the vascular system, plant stunting, plant wilt, necrosis and death. The pathogen that causes this disease is difficult to control. It spreads through the soil, in which it can survive for a long time, and through plant debris and seeds (Cox *et al*., 2019).

Cotton diseases most studied

The most studied diseases in the literature are described below.

*Root rot (*Phymatotrichum omnivorum*)*
The fungus attacks the plant root, blocks the vascular elements, inhibiting the movement of water. The leaves turn yellow or brown and then wilt rapidly, causing death in a few days (Pammel, 1888; Uppalapati *et al*., 2010). The symptoms usually begin during extensive vegetative growth, are more visible during flowering and fruit development and continue through the growing season (Smith *et al*., 1962).

*Grey mildew (*Ramularia areola*)*
The disease is produced by a fungus. Initial symptoms appear, firstly, on lower leaves after the first boll set. They are light green to yellow-green translucent spots bounded by veinlets (called areolate) on the upper surface of the leaves. The severe infection leads to defoliation and premature boll opening (Chohan *et al*., 2020). *R. areola* is the most important foliar cotton disease; its infection can cause boll abortion, malformation of bolls and lower fibre quality (Xavier *et al*., 2019).

Pest management and AI

AI supports decision-making activities of pest management, such as monitoring and control. Some examples of the application of AI in pest management are (i) pest identification (Deng *et al*., 2018; Roldan-Serrato *et al*., 2018), (ii) pest counting (Xia *et al*., 2014; Yao *et al*., 2014) and (iii) pest-spread prediction (Hudgins *et al*., 2017; Chen *et al*., 2020; Ji *et al*., 2020). In disease management, AI has been used for (i) disease recognition (Habib *et al*., 2020; Velasquez *et al*., 2020) and for (ii) early plant-disease forecast (Khattab *et al*., 2019). AI has also been used for soil and irrigation management (Navinkumar *et al*., 2020; Talaviya *et al*., 2020) and weed management (Partel *et al*., 2019; Sudars *et al*., 2020; Monteiro *et al*., 2021). In this paper, we only focus on disease and insect pest management.

Materials and methods

The methodology for reviewing the papers was based on Kitchenham *et al.* (2010). The bibliographic analysis, in the domain under study, involved two steps: (a) collection of related work and (b) detailed review and analysis of these collected works. In the first step, a keyword-based search for scientific papers, between 2014 and 2021, was performed to know the most recent state-of-the-art. Although there are earlier works (Willers *et al*., 1999, 2005, 2009; Boissard *et al*., 2008; Martin *et al*., 2008), they were not included because they did not comply with the range of dates. Sources were the scientific databases: Scopus, ScienceDirect, Taylor and Francis, Springer and Google Scholar. The results of each database were merged, and later, the duplicates were deleted. Table 2 lists the research questions and their search queries.

The following inclusion criteria were used. IC-1: include publications in journals and conferences whose titles are related to the management and diagnosis of insect pests or diseases. IC-2: include publications in journals and conferences that contain keywords that match those defined in the search string. IC-3: include publications whose summary and/or introduction and/or conclusions are related to the selected topic. Finally, IC-4: include studies in English. The following exclusion criteria were applied. EC-1: exclude publications that do not match the previous inclusion criteria. EC-2: exclude all duplicates. EC-3: exclude books. EC4: exclude documents in the form of editorial, abstract, keynote, poster. EC5: exclude opinion pieces or position papers.

The list of the final papers, by research questions, is presented in Table 3. Papers that answer RQ2 also answer RQ1. Google Scholar included results of Scopus, as well Scopus included results of Science Direct, Springer and Taylor and Francis.

In general, the search strings were applied to search in the article title, abstract and keywords. In the scientific databases, 2057 papers were found. With these papers, the selection filters were applied and, finally, 30 were selected. Of the 30 papers, 30

Table 3. Data sources and results of the search queries of this SLR

RQ1	RQ2	Final selection
10	7	10
15	10	15
5	3	5
30	20	30

respond to RQ1 and 20 respond to RQ2. Figure 1 shows the flowchart of the selection process.

In the second step, the 30 selected papers from the first step were analysed one by one, considering the research questions.

Results

This section explains the SLR results, particularly, it analyses the selected papers.

General characteristics of the selected studies

Thirty papers met the eligibility criteria and were included in this review – according to the inclusion/exclusion process. In total, 66.7% were about diseases and 33.3% about insect pests. The distribution of the studies, by country, is shown in Fig. 2. These studies were conducted in five countries: India (66.7%), the United States (13.3%), Brazil (10%), China (6.7%) and Pakistan (3.3%). Not surprisingly, India represents the highest percentage, as it is the major producer of cotton, ranked number one in the world. The production of cotton in these five countries is among the top ten in the world. India produces about ~6000 metric tonnes, China ~5000, the United States ~4000 and Pakistan and Brazil ~2000 (Azam *et al.*, 2020).

AI techniques used for the management of diseases and insect pests

This section describes the techniques used for the management (detection, diagnosis, etc.), of diseases and insect pests, in cotton agriculture.

Insect pests

To identify whitefly, Sangari and Saraswady (2016) presented a pest-image segmentation using Marker-Controlled Watershed Transformation (MWT), which was compared with a fuzzy c-means (FCM) clustering. The results showed that MWT performs better than FCM, with a better convergence rate. Sangari and Saraswady (2016) used nonlinear assessments for the measurement of image distortion: the parameters evaluated were structural content (SC), peak signal-to-noise ratio (PSNR), normalized correlation coefficient (NK), normalized absolute error (NAE) and average difference (AD).

Shahzadi *et al.* (2016) proposed a rule-based system to diagnose whitefly and other insect pests. The rule-based system used moisture sensors, temperature sensors, humidity sensors and leaf-wetness sensors. For knowledge acquisition of the expert system, they used three inputs: (i) domain experts, (ii) research and (iii) field observations.

Kandalkar *et al.* (2014) used the following techniques to identify *H. armigera*: (i) for image segmentation, a saliency map; (ii) for feature extraction, the energy of an image as a feature vector with the discrete wavelet transform (DWT) – instead of colour, shape and texture features and (iii) for pest classification, a back-propagation neural-network (BPNN).

Pratheepa *et al.* (2016) used Shannon's information theory (SIT) to find significant factors that affect *H. armigera* incidence. The results showed that correlation analysis revealed that crop stage is negatively correlated with pest population, which is true because the *H. armigera* population started to increase when the crop was in an earlier stage of fruiting and boll formation, and started to decline when the crop was in boll-bursting stage. The crop stage, followed by the number of rainy days in a week and relative humidity, were crucial in the pest population fluctuation, which also had seasonal effects. In addition, they found that SIT is more suitable to find significant factors, in pest surveillance data, rather than regression analysis.

To evaluate the severity of mealybug, Singh *et al.* (2016) developed a model to map mealybug damage using remote-sensing indices. They used multiple linear regression for data analysis and evaluated the relationship between spectral vegetation indices (SVIs) and severity index. These two indices had a huge correlation between healthy and mealybug-infested cotton.

Ranjitha *et al.* (2014) used Pearson correlation to predict thrips damage. They determined the correlation between canopy reflectance and SVIs. Recently, Alves *et al.* (2020) used convolutional neural networks (CNNs) to classify 13 insect pests (e.g. *H. armigera*, *Aphis gossypii*, *A. grandis*, etc.). They used a modified deep residual learning (RestNet34*). RestNet34* improved the accuracy of other algorithms: local binary patterns with support vector machine (LBP-SVM), AlexNet, ResNet34 and ResNet50.

The distribution of the reviewed papers according to whether they used the classification algorithms, image segmentation or a combination of both is shown in Table 4. Seven papers focused on insect-pest classification (the majority). For image classification, the AI techniques used were based on artificial neural networks (ANNs), regression and rules. Finally, two papers studied image segmentation.

Diseases

To detect *Phyllosticta gossypina*, Zhang *et al.* (2018) proposed an active-contour model (ACM) – based on a global gradient and local information – to detect the disease from images. ACM was more accurate in segmentation – and with lower running time – than geodesic active contour, chan-vese and local binary fitting. In a complex background, ACM can segment the leaves of cotton with uneven illumination, shadow and fuzzy edges. The results showed that ACM is the most suitable for the segmentation of diseased leaves under natural conditions.

Rothe and Rothe (2019) used another technique in image segmentation to detect bacterial leaf blight, *Myrothecium* and *Alternaria*. They used Otsu's segmentation to capture the image of a diseased leaf in such a way that its background is kept intact. This allowed the separation of the spot from the underlying organic background of the leaf. In the stage of classification, they used a BPNN. The accuracy of the classification was 97.14% for *Alternaria*, 93.3% for bacterial blight and 96% for *Myrothecium*.

Patil and Zambre (2014) also used Otsu's segmentation, but their research focused on cotton-leaf spot classification. They also used other techniques in the process: (i) for image segmentation, global threshold, variable threshold and Otsu' segmentation for an automatic threshold; (ii) for feature extraction,

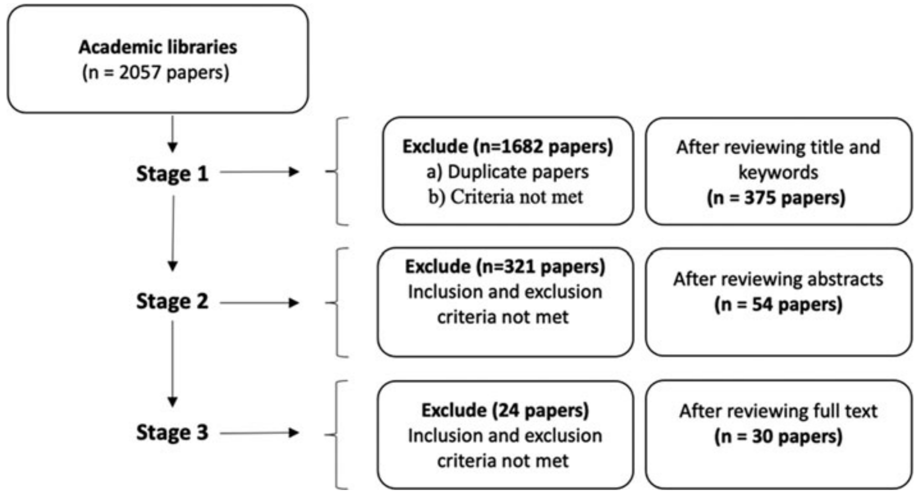


Fig. 1. Flowchart of the selection process for this SLR.

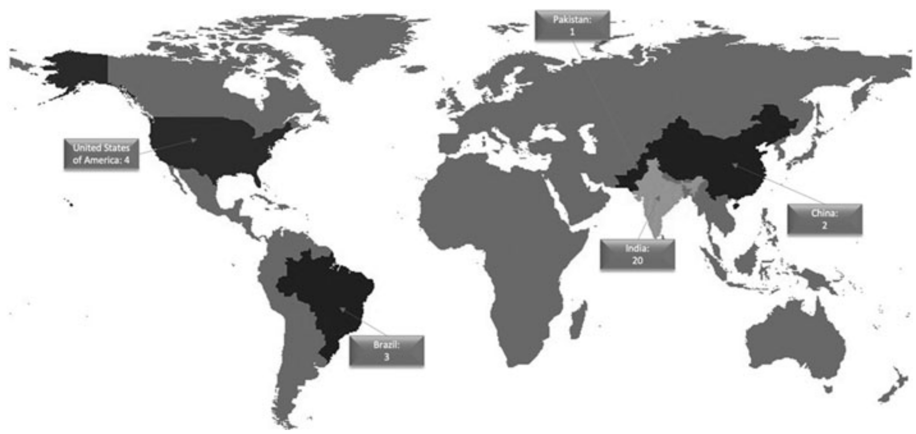


Fig. 2. World map of reviewed research articles in this SLR.

Table 4. Use of AI for insect-pest management in cotton according to the problem of classification or segmentation

	Problem		AI technique			
References	Classification	Image segmentation	ANN	Rule-based	Regression	Clustering
Nigam *et al.* (2016), Ranjitha *et al.* (2014), Singh *et al.* (2016)	X				X	
Dalmia *et al.* (2020), Alves *et al.* (2020), Kandalkar *et al.* (2014)	X		X			
Shahzadi *et al.* (2016)	X			X		
Pagariya and Bartere (2014), Sangari and Saraswady (2016)		X				X

ANN, artificial neural networks.

features of diseased leaf spot; (iii) for shape feature extraction, general descriptors such as the number of the object, area of the shape object, width and length of the object and area of the image. They used the SVM as the classification algorithm. For the classification, they determined that the morphology and the colour of leaf spots were very important because it provided critical information on the visual representation of the disease.

To detect bacterial blight, *Alternaria*, and root rot, Prashar *et al.* (2017) created an automatic cotton-crop disease-recognition method using the different invariant feature descriptors and SVM. In the pre-processing of the images, all the images were standardized by resizing them to the same size. After, the images were converted into two-dimensional images, using a grey-scale conversion, and a Gaussian filter was used for noise removal of the grey-scale images. As a feature descriptor, the histogram of oriented gradients (HOG) was used. Finally, for classification, they used SVM with 85% accuracy.

https://doi.org/10.1017/S002185962200017X Published online by Cambridge University Press

Table 5. Type of problem to solve *v.* AI techniques (ANNs, rule-based, regression, clustering, SVM, DT or gradient-based) to detect cotton diseases

	Problem			AI techniques						
References	Classification	Image segmentation	Feature extraction	ANN	Rule-based	Regression	Clustering	SVM	DT	Gradient-based
Caldeira *et al.* (2021), Liang, (2021), Patil and Burkpalli (2021), Patil and Patil (2021)	X			X						
Rastogi and Solanki (2015)	X				X			X		
Toseef and Khan (2018)	X				X					
Yang *et al.* (2014)	X			X				X		
Xavier *et al.* (2019)	X					X		X	X	
Chopda *et al.* (2018)	X								X	
Dumare and Mungona (2017), Usha Kumari *et al.* (2019)	X	X					X	X		
Sarangdhar and Pawar (2017)	X	X				X	X	X		
Prashar *et al.* (2017)	X	X					X	X		X
Patil and Zambre (2014)	X	X					X	X		
Rothe and Rothe (2019)	X	X		X			X			
Revathi and Hemalatha (2014)	X		X		X					
Zhang *et al.* (2018)		X								X

https://doi.org/10.1017/S002185962200017X Published online by Cambridge University Press

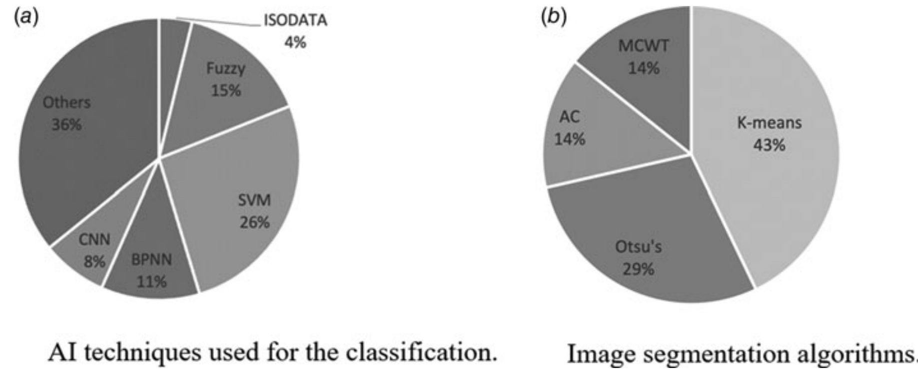


Fig. 3. AI techniques used for the classification and image segmentation of cotton diseases and insect pests. AC, active contour model based on global gradient and local information; MCWT, marker-controlled watershed transformation; ISODATA, iterative self-organizing data analysis; SVM, support vector machine; BPNN, back-propagation neural-network; CNN, convolutional neural networks.

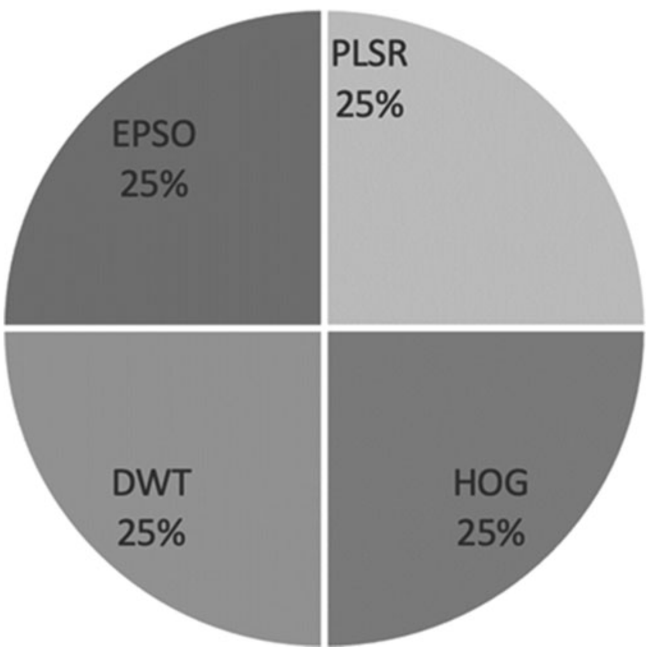


Fig. 4. Image feature-extraction algorithms of cotton diseases and insect pests. HOG, histogram of oriented gradients; PLSR, partial least square regression; DWT, discrete wavelet transform; EPSO, enhanced particle swarm optimization.

Another study, by Sarangdhar and Pawar (2017), included SVM to classify five cotton leaf diseases (bacterial blight, *Alternaria*, grey mildew, *Cercospora* and fusarium wilt). The main steps for detection were: (i) image acquisition; (ii) pre-processing (the images were resized, and the noise was removed); (iii) segmentation (colour transformation and threshold were used to extract from the region of interest of the lesion region); (iv) feature extraction (colour and texture features were extracted using partial least-square regression (PLSR)) and (v) classification (SVM regression with Gaussian kernel). The overall classification accuracy was 83.26%.

For the detection of ramularia leaf blight, Xavier *et al.* (2019) used multispectral classifications, with four classifiers, in Waikato environment for knowledge analysis software: (i) multinomial logistic regression (MLR), (ii) multinomial logistic regression with boosting (MLRb), (iii) SVM and (iv) random forest (RF). Xavier *et al.* (2019) focused on the application of different algorithms to minimize the possibility that the obtained performance of infection level may be caused by the specifications of a single classifier. The MLR used a linear-predictor function and required small training data to estimate the parameters for classification. SVM was adjusted to nonlinear class predictors and performed well in the multi-spectral remote-sensing classification. In addition, the results showed that the other two approaches – MLRb and RFT – were affected more by overfitting of training data or higher amounts of training data demanded: MLRb because of its underlying boosting and RF because it was an ensemble approach based on bootstrap aggregating (bagging).

To map cotton root rot, Yang *et al.* (2014) evaluated (i) iterative self-organizing data analysis (ISODATA) unsupervised classification applied to multi-spectral images, (ii) unsupervised classification applied to normalized difference vegetation index (NDVI) and (iii) two supervised-classification techniques, BPNN and SVM. Images were taken from airborne multi-spectral imagery. All methods appeared to be equally effective and accurate, for the detection of cotton root rot, for site-specific management of this disease. Especially, the NDVI-based classification can be easily implemented without the need for complex image processing capabilities. Results demonstrated that ISODATA applied to multi-spectral imagery (94%), NDVI combined with unsupervised classification (94.5%) and the supervised classifiers (BPNN (95.5%) and SVM (95%)) are all effective to detect root rot.

Yang *et al.* (2016) used multispectral imaging to detect consistency and changes, in cotton root rot disease, for 10 years. They used ISODATA for root rot classification. The result showed that NDVI-based ISODATA classification appears to be a simple and effective method to generate root rot infection maps.

Similarly, Song *et al.* (2017) classified the root rot with ISODATA, with the minimum spectral distance to group each pixel into a class, based on the four spectral bands (e.g. red-green-blue (RGB) and near-infrared (NIR)) and the NDVI combination.

For automatic detection of alternaria leaf spot, grey mildew and rust foliar, Usha Kumari *et al.* (2019) created an automatic disease detection for the three diseases. Usha Kumari *et al.* used the *k*-means clustering algorithm for disease image segmentation of the cotton leaf. The diseased cluster was segmented into three clusters. From each cluster, the features mean, contrast, energy, correlation, standard deviation, variance, entropy and kurtosis were extracted. The extracted features were given to a BPNN and an SVM for classification. The performance of these classifiers was compared, and the following results were obtained: The alternaria leaf spot disease was classified 77.4% for BPNN and 84.3% for SVM; grey mildew disease was 87.8% for BPNN and 98.7% for SVM; rust foliar fungal disease was 90.1% for BPNN and 93.2% for SVM. The overall average accuracy of the BPNN classifier was 85.1% and for SVM was 92.06%. SVM classifiers gave more accurate disease detection than BPNN.

https://doi.org/10.1017/S002185962200017X Published online by Cambridge University Press

To identify bacterial blight, two articles were based on *k*-means clustering: Dumare and Mungona (2017) used *k*-means for image segmentation and SVM for classification; and Pagariya and Bartere (2014) used *k*-means to identify the disease. To identify different diseases, Rastogi and Solanki (2015) developed an expert system with a fuzzy inference to identify the diseases at an early stage. The classification serves at two levels: (i) classification and grouping of disease, having the same causing agents – such as viruses, bacteria and fungi – based on a feature vector extraction and (ii) reclassification based on SVM.

To diagnose 21 cotton diseases, Toseef and Khan (2018) proposed a fuzzy inference system for the diagnosis of crop diseases. The system diagnosed the main diseases of cotton and wheat. Twenty-one diseases (e.g. bacterial blight, leaf curl, root rot, verticillium wilt and anthracnose) were diagnosed with 99% of accuracy. Toseef and Khan (2018) had three main reasons to apply fuzzy logic for decision-making: (i) rules are derived from expert knowledge and described in natural language; (ii) fuzzy logic is a powerful knowledge representation mechanism for linguistic knowledge and (iii) fuzzy logic handles the vagueness and uncertainty inherent in the problem domain, which is not handled by classical set theory. Seventy-three inference rules were built for decision-making.

To detect bacterial blight, fusarium wilt, leaf blight, root rot, micro-nutrient and verticillium wilt diseases, Revathi and Hemalatha (2014) proposed a new feature extraction method using enhanced particle swarm optimization (EPSO) with skew-divergence. The obtained features were classified using SVM, BPNN and fuzzy classifiers. The accuracy was of 91, 93 and 94% for SVM, BPNN and fuzzy, respectively. The results showed higher accuracy when EPSO is combined with fuzzy classifiers.

To predict anthracnose and grey mildew diseases, Chopda *et al.* (2018) used temperature sensors and soil-moisture sensors. They used a decision-tree (DT) classifier because it is a simple classification technique that implies a set of questions about the attributes of the test data set. The results showed that the system can predict the disease with parameters such as temperature and soil moisture, based on the previous year data.

Recently, Caldeira *et al.* (2021) used CNN (GoogleNet and Resnet50 with 86.6 and 89.2% of accuracy, respectively) for cotton disease classification. The results were better for the processing of images compared with traditional approaches such as SVM, KNN, ANN and neuro-fuzzy. Liang (2021) also used CNNs (Vgg, DesenNet, ResNet and S-DesneNet). These CNNs were optimized with the spatial structure optimizer (SSO). The result showed more accuracy in classification in small samples.

To detect the diseases, the authors of the reviewed papers focused on classification, image segmentation or feature extraction (see Table 5). Ten papers focused on disease classification. Six papers combined image segmentation and classification. One paper combined classification and feature extraction. Among the AI techniques used, those based on SVMs and ANNs were the most used. Regarding image segmentation, all the works used *k*-means.

Summary

Many AI techniques were used to detect cotton diseases and insect pests. Such AI techniques allowed an automatic detection by crop symptoms, environmental conditions or physical characteristics of the pest or disease. Particularly, it was found that the articles focused on classification algorithms, image segmentation and feature extraction. The classification algorithms that stood out most for their results were CNNs, ISODATA, BPNN, fuzzy inference and SVMs, the latter more frequently (see Fig. 3). For segmentation, algorithms such as ACM, MWT and Otsu's segmentation were used (Fig. 3). And, finally, for feature extraction, DWT, EPSO, HOG, and PLSR were used in equal proportion (Fig. 4).

Some reviewed papers made comparisons of different algorithms. Revathi and Hemalatha (2014) found that EPSO for feature extraction, and SVM or BPNN for classification, work best. Yang *et al.* (2014) compared four classification algorithms (ISODATA applied to multispectral imagery, NDVI combined with unsupervised classification and two supervised classifiers (BPNN, SVM)) with very close results. Usha Kumari *et al.* (2019) found that *k*-means for image segmentation and SVMs for classification gave better results than *k*-means for image segmentation and BPNN for classification. Xavier *et al.* (2019) evaluated the classification algorithms MLR, MLRb, SVMs and RFT, obtaining similar results. The most used technique was SVM (26%), outperforming BPNN (11%) (see Fig. 3). Nonetheless, in 2021, CNNs got the best performance (Caldeira *et al.*, 2021; Liang, 2021).

Sensing techniques used for the detection of pests and diseases

This section describes sensing techniques, used to detect diseases and insect pests, on cotton.

Cameras

Cameras were used alone or through platforms such as satellites, aircraft or unmanned airborne vehicles (UAVs). The details are presented below.

Without platform. To detect *P. gossypina*, Zhang *et al.* (2018) proposed an automatic segmentation of a diseased leaf, to improve the image-segmentation performance of cotton leaves in a natural environment. They used a digital single-lens reflex (DSLR) camera with a Canon electro-optical system.

Satellite. To evaluate mealybug severity, Singh *et al.* (2016) evaluated the relationship between mealybug severity and remote-sensing indices. The authors used Landsat TM5 satellite images with spectral bands RGB, NIR, shortwave infrared and thermal. The mealybug-infested cotton crop had a significantly lower reflectance (33%) in the NIR region, and higher (14%) in the visible range of the spectrum, when compared with the non-infested cotton crop, having NIR of 48% and visible-region reflectance of 9%. These results indicate that remote sensing has the potential to distinguish damage by mealybug and quantify its abundance in cotton.

Unmanned airborne vehicle. To detect ramularia leaf blight, Xavier *et al.* (2019) used multispectral imagery from an UAV. The camera captured wavelengths of 520–600 nm (green band), 630–690 nm (red band) and 760–900 nm (NIR band). The camera was on an UAV with flight heights of 100, 300, 500 and 700 m. This type of imaging helped to detect the disease; however, the images were not sufficient to differentiate finer-scaled disease severity levels. The results show that a camera with a higher resolution is needed to improve the disease classification.

https://doi.org/10.1017/S002185962200017X Published online by Cambridge University Press

Aircraft. Yang *et al.* (2016) used multispectral imaging to detect consistency and changes, in the root rot disease, over 10 years (2002–2012). Firstly, the authors used cameras in three bands: green (555–565 nm), red (625–635 nm) and NIR (845–857 nm). Later, they used cameras in four spectral bands: blue (430–470 nm), green (530–570 nm), red (630–670 nm) and NIR (810–850 nm). Finally, the authors used an RGB camera and NIR (720 nm). All images were acquired from an aircraft at an altitude of 3050 m. Results demonstrated that root rot tends to occur in the same general areas within fields in recurring years, even though variations in infection patterns exist over the years.

To identify root rot, Song *et al.* (2017) used two methods for image acquisition – airborne multispectral imagery and satellite imagery – to identify infested areas. In the first case, they used two Nikon D810 digital complementary metal–oxide–semiconductor cameras. One camera was used to capture RGB images, and the other camera captured NIR images. Airborne images were taken at an altitude of 3050 m. Both cameras simultaneously and independently captured images. For the satellite imagery, they used the Sentinel-2A in the bands RGB and NIR. The authors assessed the potential of 10-m Sentinel-2A satellite imagery for root rot detection and compared it with airborne multispectral imagery. Accuracy assessment showed that the classification maps from the Sentinel-2A imagery were better than the airborne-image classification. However, they found some small root-rot areas were undetectable, and some non-infested areas within large root-rot areas were incorrectly classified as infested due to the images' coarse spatial resolution. These results demonstrate that freely-available Sentinel-2 imagery can be used as an alternative data source for identifying root rot and creating prescription maps for site-specific management.

Sensors

Three types of sensors were used: field sensors, spectroradiometer and microscope. In what follows, we explain each type.

Field sensors. To detect bacterial blight, *Alternaria*, grey mildew, *Cercospora* and fusarium wilt, Sarangdhar and Pawar (2017) used image acquisition, environment temperature sensors, humidity sensors, soil-moisture sensors and water sensors to detect and control diseases in cotton. A Nikon camera (non-specified model) captured RGB images. Environmental temperature sensors, humidity sensors and moisture sensors were used to monitor the soil. A water sensor was used to monitor the water level of a pesticide tank. The results showed that, with timely detection and permanent monitoring, cotton production can be improved.

To predict anthracnose and grey mildew, Chopda *et al.* (2018) used environment temperature sensors and soil-moisture sensors. The results showed that the system can predict cotton-crop diseases with temperature, soil moisture, based on the previous year data. In the same way, Shahzadi *et al.* (2016) used sensors to determine the conditions that favour the appearance of whitefly, thrips, jassid and pink bollworm. Soil sensors collected data on soil conditions, soil moisture, soil content and leaf-wetness sensors. In addition, weather sensors collected data about humidity and temperature.

Finally, Pratheepa *et al.* (2016) used data mining to find the significant factors that affect the incidence of the pest *H. armigera*. The authors considered for the analysis, as incidence factors, the crop stage of the cotton, season and abiotic factors such as maximum temperature, minimum temperature, morning relative humidity, evening relative humidity, rainfall and number of rainy days in a week. The results showed that among all the factors, crop stage played a major role, followed by the number of rainy days in a week, and relative humidity, for the insect pest incidence.

Spectroradiometer. To detect and estimate the damage caused by *T. tobacco* (Lind), Ranjitha *et al.* (2014) used a spectroradiometer, from 70 to 90 days, after sowing. Canopy reflectance was recorded and SVIs were estimated. The hyper-spectral radiometer recorded the spectral reflectance in blue (450–520 nm), green (520–590 nm), red (620–680 nm) and NIR (770–860 nm), at 30 cm above the cotton canopy. The results showed that the reflectance decreased in NIR, while RGB reflectance increased compared to undamaged plants. Red band (at wavelengths 691 and 710 nm) and green-red vegetative index were found to be more sensitive to thrips damage. The sensitivity curve shows a single peak in the blue region (at about 496 nm), which is characteristic of the thrips damage.

To determine a whitefly infestation, Nigam *et al.* (2016) determined the relation of the infestation with and biotic stress with remote sensors. They used a spectroradiometer with different sampling intervals across the spectral region of 350–2500 nm at 1 nm. Chlorophyll concentration was measured to determine the relationship between whitefly infestation damage severity and chlorophyll concentration. A whitefly-infested cotton crop showed a low-reflectance value between 350–1335 nm and 1526–1769 nm. Whitefly-infested leaf-tissue was damaged and reflectance, in NIR, also went down drastically compared to healthy plants.

Microscope. To detect bacterial leaf blight, myrothecium and alternaria present, Rothe and Rothe (2019) used a DSLR camcorder and Leica Wild M3C microscope in natural situations.

Summary

Sensing techniques were used to detect cotton diseases and insect pests. These techniques allowed capturing information of crop symptoms, environmental conditions or physical characteristics of the insect pest or disease. Sensing techniques captured (i) images and (ii) environmental conditions. In the case of the images, multi-spectral cameras, DSLR cameras, spectroradiometers and microscopes were used (Ranjitha *et al.*, 2014; Yang *et al.*, 2016; Rothe and Rothe, 2019).

In the case of the cameras, most were installed in UAVs with a flight altitude between 100 and 700 m, aircrafts with a flight altitude of 3050 m or satellites of low Earth orbit at an altitude of 705 km for Landsat TM5 and 786 km for Sentinel-2A. Cameras, in 38% of the articles, were the most used in the detection of diseases and insect pests in cotton.

In the case of *in-situ* sensors, environment-temperature sensors, humidity sensors, soil-moisture sensors, leaf-wetness sensors and water sensors were used (Nigam *et al.*, 2016; Sarangdhar and Pawar, 2017). In the selected studies, it was demonstrated that with low-cost sensors, it was possible to take enough information to make predictions of insect pests or diseases. The detection techniques that were used in the selected papers include cameras, soil moisture sensors, temperature sensors, water sensors, humidity sensors, leaf wetness sensors, spectroradiometers and microscopes (see Fig. 5).

Discussion

In this study, we have systematically searched the scientific literature, from 2014 to 2021, to establish the state-of-the-art on

https://doi.org/10.1017/S002185962200017X Published online by Cambridge University Press

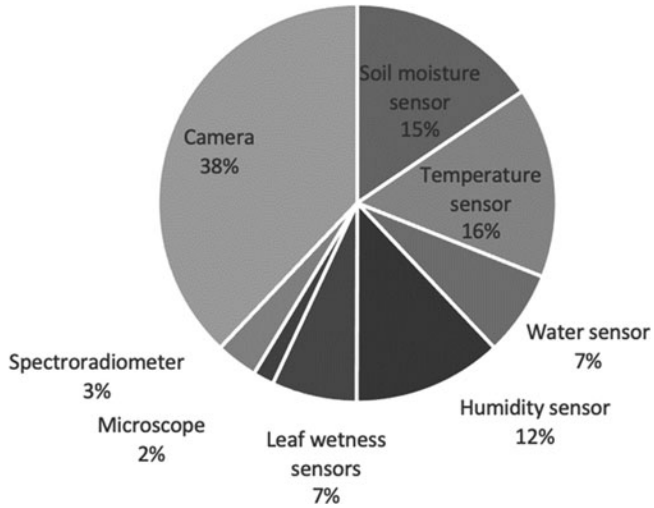


Fig. 5. Sensing techniques used to detect insect pests or diseases in cotton.

management (detection, diagnosis, etc.) of diseases and insect pests in cotton. Many AI algorithms were used to detect cotton diseases and insect pests. AI algorithms allowed an automatic detection of crop symptoms, environmental conditions or physical characteristics of the pest or disease.

AI algorithms

Regarding AI algorithms, it was found that articles focused on classification algorithms, image segmentation and feature extraction. The classification algorithms that stood out most for their results were CNNs, ISODATA, BPNN, fuzzy inference and SVM – the latter most frequently. For segmentation, algorithms, ACM, MCWT and Otsu's segmentation were used. Finally, for feature extraction, DWT, EPSO, HOG and PLSR were equally used.

HOG focuses on the structure or the shape of an object. HOG + SVM gave better results than scale-invariant feature transform + SVM or than spectral asymmetry index + SVM. PLSR is often used when there are a lot of explanatory variables, possibly correlated. A key advantage of DWT is that it has temporal resolution: DWT captures both frequency and location in timed information. Regarding EPSO, it has several advantages such as simplicity, convergence speed and robustness.

More recent works have focused on the use of CNN, which has led to greater accuracy in the classification of pests and diseases (Caldeira *et al.*, 2021; Liang, 2021). The results were better compared with traditional approaches for the processing of images. The main disadvantage is that CNN can, sometimes, take much longer to train (Kamilaris and Prenafeta-Boldú, 2018). To solve this, Liang (2021) used SSO on the training process, in different architectures (including Vgg, DesenNet, and ResNet and S-DesneNet), in small samples.

Insect-pest detection techniques

The most investigated insect pest was the whitefly, followed by thrips and pink bollworm (see Fig. 6). The sensors that were used for the whitefly were spectroradiometers, soil-moisture sensors, temperature sensors, humidity sensors and leaf-wetness sensors. In the case of thrips, in addition to the sensors used with

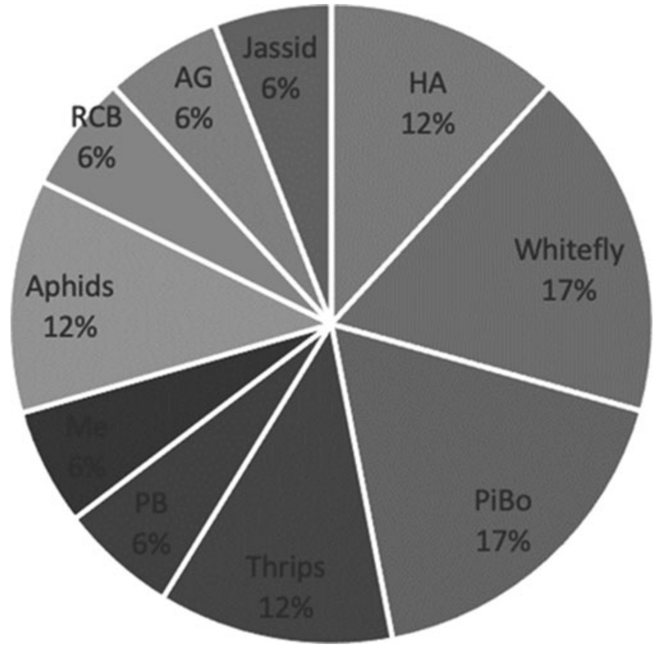


Fig. 6. Insect pests for cotton that were studied in the reviewed papers. HA, *Helicoverpa armigera*; WF, whitefly; PiBo, pink and American bollworm; PB, pod borer; RCB, red cotton bug; AG, *Anthonomus grandis*; mealybug, Me.

whitefly, satellites were used. These insect pests have been the most studied because they have the capacity to infest several crops; for instance, cotton, rose, soybean, corn, pepper, tomato and lettuce. The degree of infestation is rapid, for example, the whitefly can lay 130 eggs (Boissard *et al.*, 2008), which is why early detection of this pest is important. Image recognition was mainly used, but the analysis of environmental conditions was also important to discriminate biotic stress (Nigam *et al.*, 2016) that can encourage the appearance of insect pests; for instance, this is the case of the whitefly which develops rapidly in warm weather.

According to the AI problem, whether it was classification, image segmentation or feature extraction, the authors combined image sensing or field sensing techniques to detect insect pests in cotton (see Table 6). Two papers used field sensors to classify insect pests and four papers used images. Other papers focused on AI techniques for other tasks (segmentation and feature extraction) and did not describe the sensing techniques (N/A means that sensing techniques were not available).

Disease-detection techniques

The most investigated diseases were root rot using cameras (see Fig. 7), followed by bacterial blight using sensors. The sensors that were used to detect the bacterial blight were soil-moisture sensors, temperature sensors, humidity sensors, cameras, microscopes and water sensors (Pagariya and Bartere, 2014; Sarangdhar and Pawar, 2017; Toseef and Khan, 2018; Rothe and Rothe, 2019). The research studies were, mainly, conducted in disease recognition. For this purpose, the authors used close-up images of diseased leaves of cotton captured with cameras (Revathi and Hemalatha, 2014; Zhang *et al.*, 2018; Rothe and Rothe, 2019). Cameras were mounted on aircrafts, satellites or UAV platforms, to capture images of entire fields of cotton, to

Table 6. Sensing techniques and AI problems to detect insect pests in cotton

References	AI problem			Sensing technique	
	Classification	Image segmentation	Feature extraction	Images	Field sensor
Dalmia *et al.* (2020), Nigam *et al.* (2016), Ranjitha *et al.* (2014), Singh *et al.* (2016)	X			X	
Pratheepa *et al.* (2016), Shahzadi *et al.* (2016)	X				X
Sangari and Saraswady (2016)	X	X		N/A	N/A
Kandalkar *et al.* (2014)	X		X	N/A	N/A
Alves *et al.* (2020)	X			N/A	N/A
Pagariya and Bartere (2014)		X		N/A	N/A

Sensing technique for images: cameras or spectroradiometers; mainly cameras.
N/A, not available.

make prescription maps or monitor diseases, in cotton (Yang *et al.*, 2016; Song *et al.*, 2017; Xavier *et al.*, 2019).

The reviewed papers were classified according to the AI problem and the sensing techniques to detect diseases (see Table 7). Most studies used classification with images and some with field sensors. For the combination of classification and image segmentation, mainly cameras or microscopes (images) were used. Only two papers used field sensors, one of which supplemented it with cameras.

Trends in the reviewed articles

The main results from the selected articles are summarized in Fig. 8. In the upper part, we can note from left to right aspects such as: (i) AI tasks, (ii) the most used AI techniques, (iii) the most researched pest and disease in cotton and (iv) the sensing techniques (for detection). The width of the nodes and their links is proportional to the number of reviewed articles in each of the categories. Different colours were used in the links, to facilitate the visualization of the connections. Overall, there was more research on diseases compared to insect pests for cotton. For diseases and insect pests, the AI tasks were classification, image segmentation and feature extraction. SVM and fuzzy inference were widely used for disease classification, and *k*-means was used for image segmentation of diseases and insect pests. The images were taken with microscopes, spectroradiometers and cameras – the latter most frequently. Finally, regarding sensing techniques, soil-moisture sensors, humidity sensors and temperature sensors were frequently used in combination, for both, diseases and insect pests.

Comparison of the results with respect to previous SLRs

Same as this SLR, Patrício and Rieder (2018) found that the SVM classifier was the most used with good results. In addition to SVM, Iqbal *et al.* (2018) found neural networks as the most used for classification. Patrício and Rieder (2018) worked in precision agriculture for grain crops and Iqbal *et al.* worked in the detection of citrus-plant diseases. Iqbal *et al.* also found that *k*-means clustering performs well for image segmentation. Regarding remote sensing, Zhang *et al.* (2019) found studies that used hyperspectral and multispectral systems to monitor plant diseases and pests.

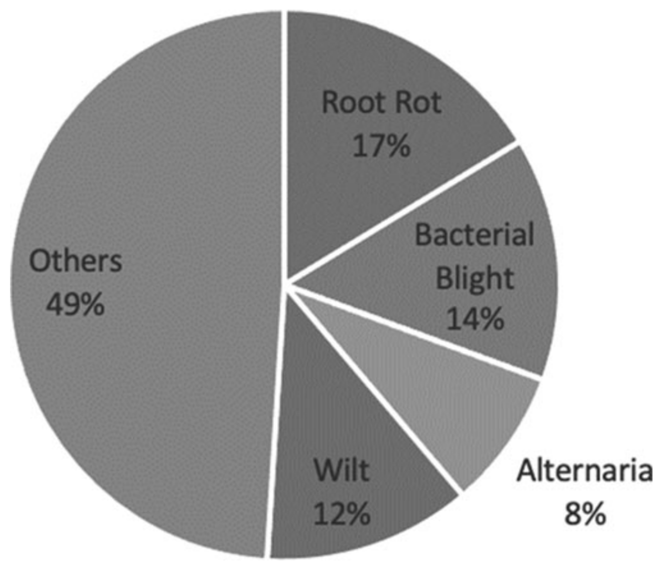


Fig. 7. Diseases for cotton that were studied in the reviewed papers.

In the papers reviewed by Patrício and Rieder (2018), they found that deep learning has been used for the detection of some stored-grain insects, while Iqbal *et al.* (2018) found that deep learning has been used for the detection of some citrus-fruit diseases. Regarding remote sensing, Zhang *et al.* (2019) found studies that used fluorescence and thermal systems, synthetic aperture radar and light detection and ranging equipment. It is worth noting that, Zhang *et al.* did not include studies with *in-situ* sensors like this SLR.

Limitations of the selected papers

The limitations of the papers selected according to the research questions in this SLR are described below.

Limitations of AI to manage diseases and insect pests in cotton

AI has contributed to the development of agriculture, in pest and disease management, and specifically, to detect and diagnose diseases and insect pests in cotton. AI allows detecting diseases and insect pests in, timely, quickly and with more precision (Xia *et al.*, 2014; Khattab *et al.*, 2019). To harvest a high cotton yield, it is

Table 7. Sensing techniques and AI problems to detect diseases in cotton

References	AI problem			Sensing technique	
	Classification	Image segmentation	Feature extraction	Images	Field sensors
Patil and Patil (2021), Rastogi and Solanki (2015), Toseef and Khan (2018)	X			N/A	N/A
Caldeira *et al.* (2021), Liang (2021), Patil and Burkpalli (2021), Song *et al.* (2017), Wang *et al.* (2020), Xavier *et al.* (2019), Yang *et al.* (2014, 2016)	X			X	
Dumare and Mungona (2017), Rothe and Rothe (2019)	X	X		X	
Patil and Zambre (2014), Prashar *et al.* (2017), Sarangdhar and Pawar (2017), Usha Kumari *et al.* (2019)	X	X		N/A	N/A
Chopda *et al.* (2018)	X				X
Revathi and Hemalatha (2014)	X		X	X	
Zhang *et al.* (2018)		X		X	
Sarangdhar and Pawar (2017)	X	X		X	X

The image sensing techniques used cameras or microscopes, mainly cameras.
N/A, not available.

very important the integrated management of diseases and insect pests (Anees and Shad, 2020; Chohan *et al.*, 2020). Despite the importance of insect pest and disease monitoring with AI, there are few works with neural networks, deep learning, deep residual learning and no work on meta-cognition, which have been used successfully in other studies (Cheng *et al.*, 2017; Li *et al.*, 2020; Shi *et al.*, 2020). Finally, no works include simultaneously diseases and insect pests.

Limitations of sensing techniques to detect diseases and insect pests

Remote sensing has allowed us to obtain monitoring data, in real-time, of diseases and insect pests (Singh *et al.*, 2016). This allows us to provide an overview for large areas using, for example, satellites, airplanes and UAV platforms (Ranjitha *et al.*, 2014; Song *et al.*, 2017; Xavier *et al.*, 2019). *In-situ* sensors allow obtaining data, in real-time, of environmental variables (e.g. temperature, humidity, moisture) (Pratheepa *et al.*, 2016; Chopda *et al.*, 2018). However, only one study tried data fusion and included computer vision with cameras and *in-situ* sensors to measure simultaneously soil moisture, temperature, humidity and water (Sarangdhar and Pawar, 2017). The combination of a larger set of sensor data can increase the accuracy and truthfulness of the data. There was no research that combined sensors, AI and pheromone traps for boll weevil. There are also no studies, on the presence of the cotton boll weevil in post-harvest, which take into account that boll weevil takes refuge for a long period during post-harvest.

Future research trends

In this review, we found that many AI and sensing techniques were used in the selected articles. For future research, however, there are different challenges due to the need to use data to make better decisions on the treatment of diseases and insect pests, with the possibility of anticipation of an outbreak. These challenges are summarized in the next sessions.

Develop predictive models to know when and where the diseases and insect pests attack

This challenge represents the opportunity to work with the prediction of diseases and insect pests simultaneously, using new techniques of prediction. Previous research did not deal, specifically, with both at the same time.

Implement prescriptive models to define how to control diseases and insect pests

The prescriptive models help to determine what needs to be done to attack diseases and insect pests. The prescriptive models define what activities and tasks are necessary to do when this type of problem appears in cotton. These models are important because they will allow the farmer to know what to do.

Make a smart pheromone traps system to predict the spread of pests

A smart pheromone traps system must define two aspects: firstly, determine how the pest spread; secondly, where to put geo-spatially the traps. This way, the system has the prediction task and also self-defines where put the traps. None of the previous systems did this simultaneously.

Develop diagnostic disease models

The diagnostic models allow defining the causes of the disease. This is important for those who make the decisions because might attack the causes to solve the problem. Iqbal *et al.* (2018) and Zhang *et al.* (2019) found some papers of detection and diagnosis for other crops, but not for cotton.

Make multi-detection models of diseases or pest attacks

This challenge includes multi-label techniques using labels related to diseases and insect pests (Araujo *et al.*, 2003). This technique allows detecting, simultaneously, if there are pests and diseases.

https://doi.org/10.1017/S002185962200017X Published online by Cambridge University Press

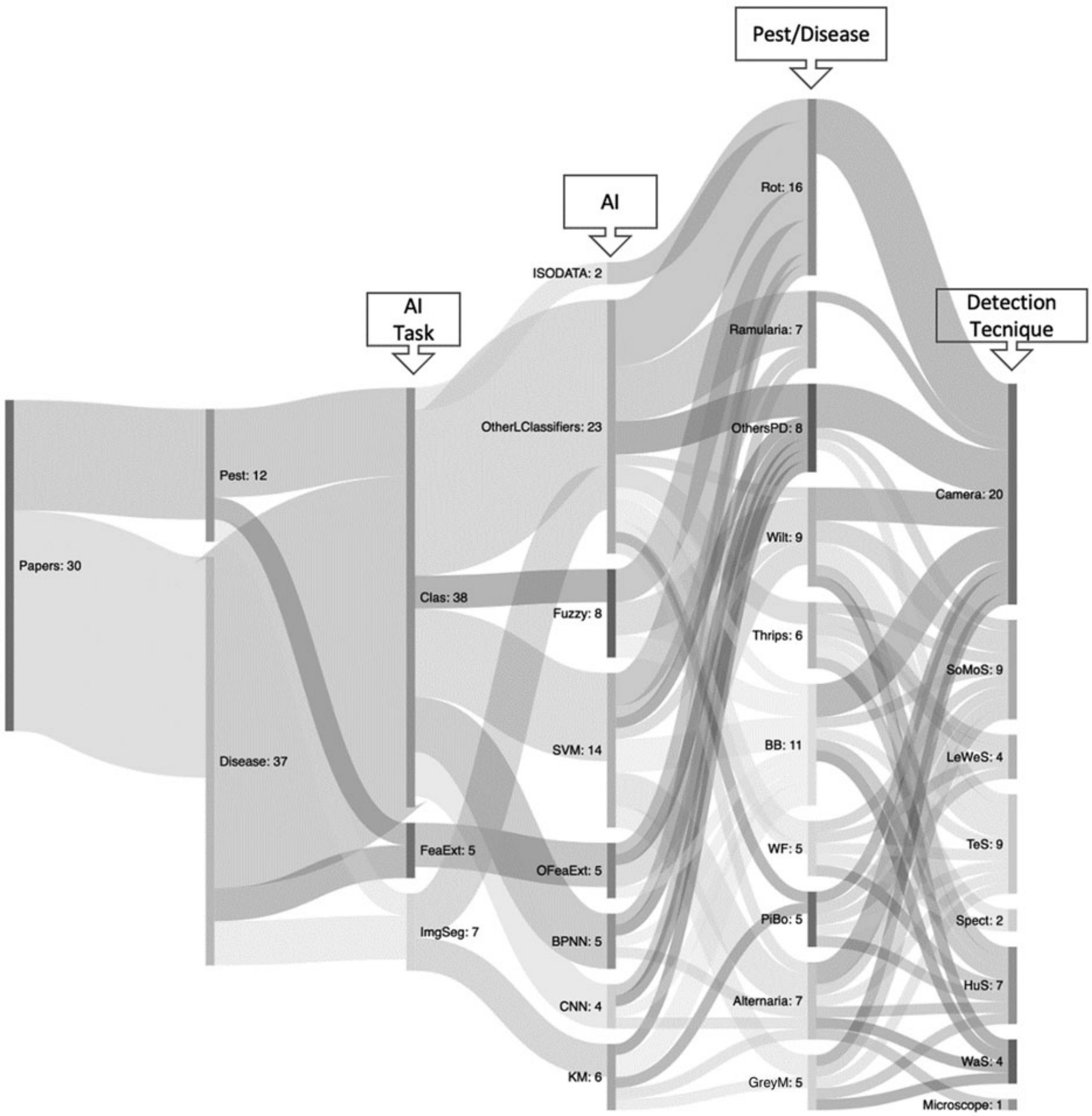


Fig. 8. Trends in the reviewed articles are divided into diseases and insect pests, AI techniques and sensing techniques. Clas, classification task; FeaExt, feature extraction task; ImgSeg, image segmentation task; AI techniques (ISODATA, iterative self-organizing data-analysis technique algorithm; SVM, support vector machine; fuzzy, fuzzy logic; KM, *k*-means; BPNN, back-propagation neural-network; CNN, convolutional neural networks; OFeaExt, others feature extraction algorithms); pest (WF, whitefly; PiBo, pink bollworm); disease (Rot, root rot; Ramularia, ramularia leaf blight; BB, bacterial blight; GreyM, grey mildew); sensing techniques (SoMoS, soil-moisture sensor; TeS, temperature sensor; WaS, water sensor; HuS, humidity sensor; LeWeS, leaf-wetness sensors; Spect, spectroradiometer).

Define a cotton-crop management system using a cognitive-computing architecture

No articles dealt with cognitive computing. In previous reviews, only (Boissard *et al.*, 2013) found one study where cognitive vision was used for pest detection. One challenge is the use of cognitive computing to manage cotton better. According to Crowder and Friess (2011), metacognition and metamemory allow AI systems to reason and adapt to the situation with self-awareness. Meta-learning facilitates the selection of appropriate AI algorithms, or adjusts them according to the task (Grąbczewski, 2014). Meta-reasoning gives systems the ability to reason, deliberate and self-optimize a decision-making process to produce effective action on time (Russell and Wefald, 1991; Svegliato and Zilberstein, 2018).

Select the most useful variables

All challenges that have been described need to establish the right variables for each model. These features can be used to improve the performance of machine-learning algorithms. The stress in crops may be generated by variables such as climatic conditions, pest damage and diseases, and the study and selection of the right variables can be complex (Ranjitha *et al.*, 2014; Yang *et al.*, 2014). It is needed to analyse which variables are most useful for each knowledge model (Pacheco *et al.*, 2014; Jiménez *et al.*, 2021).

Develop smart sticky-traps for whitefly, aphids and thrips

As an example, the whitefly eggs have a length of about 0.2 mm. Xia *et al.* (2014) used sticky traps to take whitefly-egg samples in

https://doi.org/10.1017/S002185962200017X Published online by Cambridge University Press

tomato crops. Nonetheless, in the articles analysed in this SLR, no sticky traps were used. These traps could be combined with AI in a similar manner as pheromone traps. These traps could be used for boll weevil (*A. grandis* Boheman), which affects cotton, in several countries (Neupert *et al*., 2018), but only one research was found related to the boll weevil and the use of AI for its detection (Alves *et al*., 2020). Boissard *et al*. (2013) found one study with sticky traps, for whitefly and aphids, using video to record the insects flying.

Conclusions

AI techniques were employed mainly in the context of (i) image classification, (ii) image segmentation and (iii) feature extraction of images. These techniques were successfully used for insect pests and diseases for cotton. The most used sensors were cameras, and field sensors such as temperature and humidity sensors. Future work should apply knowledge models, combined with the Internet of Things, to monitor and control diseases and insect pests.

Acknowledgements. The authors thank the anonymous reviewers for their insights, comments and suggestions.

Author contributions. J. Aguilar, M. Toro, M. Caro and R. Toscano-Miranda conceived and designed the study. J. Aguilar, M. Toro, M. Caro and R. Toscano-Miranda conducted data gathering. J. Aguilar, M. Toro, M. Caro, A. Marulanda, A. Trebilcok and R. Toscano-Miranda performed statistical analyses. J. Aguilar, M. Toro, M. Caro, A. Marulanda, A. Trebilcok and R. Toscano-Miranda wrote the article.

Financial support. This work was supported by the Colombian Science, Technology and Innovation Fund (FCTeI) of the General Royalty System (SGR); Universidad EAFIT and Universidad de Córdoba.

Conflict of interest. The authors declare there are no conflicts of interest.

References

Abu-Nasser B and Abu-Naser S (2018) Cognitive system for helping farmers in diagnosing watermelon diseases. *International Journal of Academic Information Systems Research (IJAISR)* **2**, 1–7.

Alves AN, Souza WSR and Borges DL (2020) Cotton pests classification in field-based images using deep residual networks. *Computers and Electronics in Agriculture* **174**, 105488.

Alzamily JY (2018) A rational system for diagnosing banana disorders. *International Journal of Academic Information Systems Research (IJAISR)* **2**, 1–8.

Andreopoulos A and Tsotsos JK (2013) 50 years of object recognition: directions forward. *Computer Vision and Image Understanding* **117**, 827–891.

Anees M and Shad SA (2020) Chapter 11: Insect pests of cotton and their management. In *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies*. Multan, Pakistan: Springer, pp. 177–212. https://doi.org/10.1007/978-981-15-1472-2_11.

Aqeel-Ur-Rehman and Shaikh ZA (2009) Smart agriculture. In *Applications of Modern High Performance Networks*, June 2014, Bentham Science Publishers, pp. 120–129. https://doi.org/10.2174/978160805077210901010120.

Araujo M, Aguilar J and Aponte H (2003) Fault detection system in gas lift well based on artificial immune system. *International Joint Conference on Neural Networks* **3**, 1673–1677.

Azam M, Wahid A, Ahmad M, Tayab M, Ahmed M, Ahmad S and Hasanuzzaman M (2020) Chapter 1: World cotton production: an overview. In *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies*. Multan, Pakistan: Springer, pp. 1–7.

Bannerjee G, Sarkar U, Das S and Ghosh I (2018) Artificial intelligence in agriculture: a literature survey. *International Journal of Scientific Research in Computer Science Applications and Management Studies* **7**, 1–6. https://doi.org/10.15838/alt.2018.1.4.5.

Barbedo JGA (2014) Using digital image processing for counting whiteflies on soybean leaves. *Journal of Asia-Pacific Entomology* **17**, 685–694.

Ben Guerrero E, Salvador R and Talia P (2020) Evaluation of hydrolytic enzyme activities from digestive fluids of *Anthonomus grandis* (Coleoptera: Curculionidae). *Archives of Insect Biochemistry and Physiology* **21**, 1–8. https://doi.org/10.1002/arch.21730.

Boissard P, Martin V and Moisan S (2008) A cognitive vision approach to early pest detection in greenhouse crops. *Computers and Electronics in Agriculture* **62**, 81–93.

Boissard P, Martin V and Moisan S (2013) A review of literature on application of image processing for identification of agricultural pests on various crops. *Certified International Journal of Engineering Science and Innovative Technology* **9001**, 2319–5967.

Caldeira RF, Santiago WE and Teruel B (2021) Identification of cotton leaf lesions using deep learning techniques. *Sensors* **21**, 3169. https://doi.org/10.3390/s21093169.

Carpenter ZL (1983) *Texas Plant Disease Handbook: Cotton*. Texas: Texas Agricultural Extension Service, Texas A&M University.

Carvalho J, Ruiz L, FAO F and Gobierno de Brasil AO (2018) *Cooperación sur-sur trilateral estudio nichos de mercados del algodón*. FAO and ABC.

Chen P, Xiao Q, Zhang J, Xie C and Wang B (2020) Occurrence prediction of cotton pests and diseases by bidirectional long short-term memory networks with climate and atmosphere circulation. *Computers and Electronics in Agriculture* **176**, 105612.

Cheng X, Zhang Y, Chen Y, Wu Y and Yue Y (2017) Pest identification via deep residual learning in complex background. *Computers and Electronics in Agriculture* **141**, 351–356.

Chohan S, Perveen R, Abid M, Nouman M and Sajid M (2020) Chapter 13: Cotton diseases and their management. In *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies*, Multan, Pakistan: Springer, pp. 239–270.

Chopda J, Raveshiya H, Nakum S and Nakrani V (2018) Cotton Crop Disease Detection using Decision Tree Classifier. *Vitellogenin Knockdown Strongly Affects Cotton Boll Weevil Egg Viability but Not the Number of Eggs Laid by Females International Conference on Smart City and Emerging Technology, ICSCET 2018*. https://doi.org/10.1109/ICSCET.2018.8537336.

Coelho RR, de Souza Júnior JDA, Firmino AAP, de Macedo LLP, Fonseca FCA, Terra WR, Engler G, de Almeida Engler J, da Silva MCM and Grossi-de-Sa MF (2016) Vitellogenin knockdown strongly affects cotton boll weevil egg viability but not the number of eggs laid by females. *Meta Gene* **9**, 173–180.

Cox KL, Babilonia K, Wheeler T, He P and Shan L (2019) Return of old foes – recurrence of bacterial blight and Fusarium wilt of cotton. *Current Opinion in Plant Biology* **50**, 95–103.

Crowder JA and Friess S (2011) Metacognition and metamemory concepts for AI systems. Proceedings of the 2011 International Conference on Artificial Intelligence, ICAI 2011. Vol. 2, pp. 419–424.

Dalmia A, White J, Chaurasia A, Agarwal V, Jain R, Vora D, Dhame B, Dharmaraju R and Panicker R (2020) Pest Management in Cotton Farms: An AI-System Case Study from the Global South. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3119–3127. https://doi.org/10.1145/3394486.3403363.

Deng L, Wang Y, Han Z and Yu R (2018) Research on insect pest image detection and recognition based on bio-inspired methods. *Biosystems Engineering* **169**, 139–148.

Dent D and Binks R (2020) *Insect Pest Management*, 3rd Edn edn, UK: Cabi.

Dumare NV and Mungona S (2017) Identification of cotton leaf diseases using raspberry pi. *International Journal on Recent and Innovation Trends in Computing and Communication* **5**(5), 3590–3592. https://doi.org/10.15680/IJIRSET.2017.0603111.

El-Wakeil N and Abdallah AMM (2014) Cotton pests and the actual strategies for their management control. In *Agricultural Research Updates*, Vol. **7**, Issue September 2012.

Ellis HC and Horton DL (1997) *Boll Weevil*. Available at https://web.archive.org/web/20150219082851/http://www.bugwood.org/factsheets/99-001.html.

FAO (2013) Climate-smart agriculture sourcebook. Food and agriculture organization of the United Nations. In *Sourcebook on Climate-Smart Agriculture, Forestry and Fisheries*.

FAO (2017) *Integrated Pest Management*, Vol. **4**. FAO. https://doi.org/10.1007/978-94-007-7802-3.

García R, Aguilar J, Toro M, Pinto A and Rodríguez P (2020) A systematic literature review on the use of machine learning in precision livestock farming. *Computers and Electronics in Agriculture* **179**, 105826.

Grąbczewski K (2014) *Meta-Learning in Decision Tree Induction*. Switzerland: Springer.

Grady MJO, Langton D, Hare GMPO, O'Grady MJ, Langton D and O'Hare GMP (2019) Edge computing: a tractable model for smart agriculture? *Artificial Intelligence in Agriculture* **3**, 42–51.

Grigolli JFJ, Souza LA, Fernandes MG and Busoli AC (2017) Spatial distribution of adult *Anthonomus grandis* Boheman (Coleoptera: Curculionidae) and damage to cotton flower buds due to feeding and oviposition. *Neotropical Entomology* **46**, 442–451.

Habib MT, Majumder A, Jakaria AZM, Akter M, Uddin MS and Ahmed F (2020) Machine vision based papaya disease recognition. *Journal of King Saud University – Computer and Information Sciences* **32**, 300–309.

Hudgins EJ, Liebhold AM and Leung B (2017) Predicting the spread of all invasive forest pests in the United States. *Ecology Letters* **20**, 426.

Iqbal Z, Khan MA, Sharif M, Shah JH, ur Rehman MH and Javed K (2018) An automated detection and classification of citrus plant diseases using image processing techniques: a review. *Computers and Electronics in Agriculture* **153**, 12–32.

Ji W, Han K, Lu Y and Wei J (2020) Predicting the potential distribution of the vine mealybug, *Planococcus ficus* under climate change by MaxEnt. *Crop Protection* **137**, 105268.

Jiménez M, Aguilar J, Monsalve-Pulido J and Montoya E (2021) An automatic approach of audio feature engineering for the extraction, analysis and selection of descriptors. *International Journal of Multimedia Information Retrieval* **10**, 33–42.

Kamilaris A and Prenafeta-Boldú FX (2018) A review of the use of convolutional neural networks in agriculture. *Journal of Agricultural Science* **156**, 312–322.

Kandalkar G, Deorankar AV and Chatur PN (2014) Classification of agricultural pests using DWT and back propagation neural networks. *International Journal of Computer Science and Information Technologies (IJCSIT)* **5**, 4034–4037.

Khattab A, Habib SED, Ismail H, Zayan S, Fahmy Y and Khairy MM (2019) An IoT-based cognitive monitoring system for early plant disease forecast. *Computers and Electronics in Agriculture* **166**, 105028.

Kitchenham B, Pretorius R, Budgen D, Brereton O, Turner M, Niazi M and Linkman S (2010) Systematic literature reviews in software engineering-A tertiary study. *Information and Software Technology* **52**, 792–805.

Li Y, Wang H, Dang LM, Sadeghi-Niaraki A and Moon H (2020) Crop pest recognition in natural scenes using convolutional neural networks. *Computers and Electronics in Agriculture* **169**, 105174.

Liang X (2021) Few-shot cotton leaf spots disease classification based on metric learning. *Plant Methods* **17**, 1–11.

Martin V, Moisan S, Paris B and Nicholas O (2008) Towards a video camera network for early pest detection in greenhouses. *ENDURE International Conference 2008 Diversifying Crop Protection, 2008 La Grande-Motte, France, October*, 12–15.

Monteiro AL, de Freitas Souza M, Lins HA, Teófilo TMdaS, Barros Júnior AP, Silva DV and Mendonça V (2021) A new alternative to determine weed control in agricultural systems based on artificial neural networks (ANNs). *Field Crops Research* **263**, 108075. https://doi.org/10.1016/j.fcr.2021.108075.

Ñañez L (2012) *Manejo fitosanitario del cultivo de algodón*, Produmedios.

Navinkumar TM, Ranjith Kumar R and Gokila PV (2020) Application of artificial intelligence techniques in irrigation and crop health management for crop yield enhancement. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2020.10.227.

Neupert S, Marciniak P, Köhler R, Nachman RJ, Suh CPC and Predel R (2018) Different processing of CAPA and pyrokinin precursors in the giant mealworm beetle *Zophobas atratus* (Tenebrionidae) and the boll weevil *Anthonomus grandis grandis* (Curculionidae). *General and Comparative Endocrinology* **258**, 53–59.

Nigam R, Kot R, Sandhu SS, Bhattacharya B K, Chandi RS, Singh M, Singh J and Manjunath KR (2016) Ground-based hyperspectral remote sensing to discriminate biotic stress in cotton crop. In *Multispectral, Hyperspectral, and Ultraspectral Remote Sensing Technology, Techniques and Applications VI*, vol. **9880**, SPIE, pp. 89–98.

Oliveira CM, Auad AM, Mendes SM and Frizzas MR (2014) Crop losses and the economic impact of insect pests on Brazilian agriculture. *Crop Protection* **56**, 50–54.

Pacheco F, Rangel C, Aguilar J, Cerrada M and Altamiranda J (2014) Methodological Framework for Data Processing based on the Data Science Paradigm. *XL Latin American Computing Conference (CLEI)*. https://doi.org/10.1109/CLEI.2014.6965184.

Pagariya R and Bartere M (2014) Identification of cotton crop diseases and provide their remedies using *k* means clustering. *International Journal of Advanced Research in Computer Science* **5**, 58–61.

Pammel L (1888) *Root Rot of Cotton or 'Cotton Blight'*. Bulletin/Texas Agricultural Experiment Station; no. 4. Texas, USA: Texas Agricultural Experiment Station.

Partel V, Charan Kakarla S and Ampatzidis Y (2019) Development and evaluation of a low-cost and smart technology for precision weed management utilizing artificial intelligence. *Computers and Electronics in Agriculture* **157**, 339–350.

Patil BM and Burkpalli V (2021) A perspective view of cotton leaf image classification using machine learning algorithms using WEKA. *Advances in Human–Computer Interaction* **2021**, 1–21.

Patil BV and Patil PS (2021) Computational method for cotton plant disease detection of crop management using deep learning and Internet of Things platforms. *Lecture Notes on Data Engineering and Communications Technologies* **53**, 875–885.

Patil SP and Zambre RS (2014) Classification of cotton leaf spot disease using support vector machine. *Journal of Engineering Research and Applications* **4**, 92–97.

Patrício DI and Rieder R (2018) Computer vision and artificial intelligence in precision agriculture for grain crops: a systematic review. *Computers and Electronics in Agriculture* **153**, 69–81.

Prashar K, Talwar R and Kant C (2017) Robust Automatic Cotton Crop Disease Recognition (ACDR) Method using the Hybrid Feature Descriptor with SVM. *International Conference on Computing on Sustainable Global Development, March*.

Pratheepa M, Verghese A and Bheemanna H (2016) Shannon information theory a useful tool for detecting significant abiotic factors influencing the population dynamics of *Helicoverpa armigera* (Hübner) on cotton crops. *Ecological Modelling* **337**, 25–28.

Presley JT (1954) *Cotton Diseases and Methods of control.*U.S. Department of Agriculture.

Ranjitha G, Srinivasan MR and Rajesh A (2014) Detection and estimation of damage caused by thrips *Thrips tabaci* (Lind) of cotton using hyperspectral radiometer. *Agrotechnology* **3**, 123. https://doi.org/10.4172/2168-9881.1000123.

Rastogi V and Solanki A (2015) A classification based target specific expert system for cotton crop. *Journal of Agroecology and Natural Resource Management* **2**, 134–140.

Revathi P and Hemalatha M (2014) Cotton leaf spot diseases detection utilizing feature selection with skew divergence method. *International Journal of Science Engineering and Technology* **3**, 22–30.

Roldan-Serrato KL, Escalante-Estrada JAS and Rodriguez-Gonzalez MT (2018) Automatic pest detection on bean and potato crops by applying neural classifiers. *Engineering in Agriculture, Environment and Food* **11**, 245–255.

Rothe P and Rothe J (2019) Intelligent pattern recognition system with application to cotton leaf disease identification. *Innovations in Computer Science and Engineering* **32**, 19–27.

Russell S and Wefald E (1991) Principles of metareasoning. *Artificial Intelligence* **49**, 361–395.

Sangari S and Saraswady D (2016) Analyzing the optimal performance of pest image segmentation using non linear objective assessments. *International Journal of Electrical and Computer Engineering* **6**, 2789–2796.

Sarangdhar AA and Pawar V (2017) Machine learning regression technique for cotton leaf disease detection and controlling using IoT. *Proceedings of the International Conference on Electronics, Communication and Aerospace Technology, ICECA 2017*, 2017 January, pp. 449–454.

Shahzadi R, Ferzund J, Tausif M and Asif M (2016) Internet of things based expert system for smart agriculture. *International Journal of Advanced Computer Science and Applications* **7**, 341–350. https://doi.org/10.14569/ijacsa.2016.070947.

Shi R, Zhai D, Liu X, Jiang J and Gao W (2020) Rectified meta-learning from noisy labels for robust image-based plant disease diagnosis. *ArXiv* **14**, 1–10.

Simberloff D and Rejmanek M (2011) Encyclopedia of biological invasions. In *Choice Reviews Online*, Vol. **48**. USA: Univ of California Press. https://doi.org/10.5860/choice.48-6030.

Singh SK, Dutta S and Dharaiya N (2016) A study on geospatial technology for detecting and mapping of Solenopsis mealybug (Hemiptera: Pseudococcidae) in cotton crop. *Journal of Applied and Natural Science* **8**, 2175–2181.

Smith CW and Cothren JT (1999) *Cotton: Origin, History, Technology, and Production*. John Wiley & Sons.

Smith HE, Elliott FC and Bird LS (1962) Root rot losses of cotton can be reduced. *Texas Agricultural Experiment Station* **361**, 1–8.

Solis-Sánchez LO, García-Escalante JJ, Castañeda-Miranda R, Torres-Pacheco I and Guevara-González R (2009) Machine vision algorithm for whiteflies (*Bemisia tabaci* Genn.) scouting under a greenhouse environment. *Journal of Applied Entomology* **133**, 546–552.

Solis-Sánchez LO, Castañeda-Miranda R, Garcia-Escalante JJ, Torres-Pacheco I, Guevara-Gonzalez RG, Castañeda-Miranda CL and Alaniz-Lumbreras PD (2011) Scale invariant feature approach for insect monitoring. *Computers and Electronics in Agriculture* **75**, 92–99.

Song X, Yang C, Wu M, Zhao C, Yang G, Hoffmann WC and Huang W (2017) Evaluation of Sentinel-2A satellite imagery for mapping cotton root rot. *Remote Sensing* **9**, 1–17.

Sudars K, Jasko J, Namatevs I, Ozola L and Badaukis N (2020) Dataset of annotated food crops and weed images for robotic computer vision control. *Data in Brief* **31**, 105833.

Svegliato J and Zilberstein S (2018) Adaptive Metareasoning for Bounded Rational Agents. *CAI-ECAI Workshop on Architectures and Evaluation for Generality, Autonomy and Progress in AI (AEGAP).*

Talaviya T, Shah D, Patel N, Yagnik H and Shah M (2020) Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. *Artificial Intelligence in Agriculture* **4**, 58–73.

Toseef M and Khan MJ (2018) An intelligent mobile application for diagnosis of crop diseases in Pakistan using a fuzzy inference system. *Computers and Electronics in Agriculture* **153**, 1–11.

Uppalapati SR, Young CA, Marek SM and Mysore KS (2010) Phymatotrichum (cotton) root rot caused by *Phymatotrichopsis omnivora*: retrospects and prospects. *Molecular Plant Pathology* **11**, 325–334.

Usha Kumari C, Arun Vignesh N, Panigrahy AK, Ramya L and Padma T (2019) Fungal disease in cotton leaf detection and classification using neural networks and support vector machines. *International Journal of Innovative Technology and Exploring Engineering* **8**, 3664–3667.

Velasquez D, Sanchez A, Sarmiento S, Toro M, Maiza M and Sierra B (2020) A method for detecting coffee leaf rust through wireless sensor networks, remote sensing, and deep learning: case study of the caturra variety in Colombia. *Applied Sciences* **10**, 697.

Wang T, Thomasson JA, Yang C, Isakeit T, Nichols RL, Collett RM, Han X and Bagnall C (2020) Unmanned aerial vehicle remote sensing to delineate cotton root rot. *Journal of Applied Remote Sensing* **14**, 34522. https://doi.org/10.1117/1.jrs.14.034522.

Willers J, Seal MR and Luttrell RG (1999) Remote sensing, line-intercept sampling for tarnished plant bugs (Heteroptera: Miridae) in mid-south cotton. *Journal of Cotton Science* **3**, 160–170.

Willers J, Jenkins JN, Ladner WL, Gerard PD, Boykin DL, Hood KB, Mckibben PL, Samson SA and Bethel MM (2005) Site-specific approaches to cotton insect control. Sampling and remote sensing analysis techniques. *Precision Agriculture* **6**, 431–452.

Willers J, Jenkins JN, McKinion JM, Gerard P, Hood KB, Bassie JR and Cauthen MD (2009) Methods of analysis for georeferenced sample counts of tarnished plant bugs in cotton. *Precision Agriculture* **10**, 189–212.

Xavier TWF, Souto RNV, Statella T, Galbieri R, Santos ES, Suli GS and Zeilhofer P (2019) Identification of ramularia leaf blight cotton disease infection levels by multispectral, multiscale UAV imagery. *Drones* **3**, 33.

Xia C (2012) *In situ* detection of small-size insect pests sampled on traps using multifractal analysis. *Optical Engineering* **51**, 027001.

Xia C, Chon TS, Ren Z and Lee JM (2014) Automatic identification and counting of small size pests in greenhouse conditions with low computational cost. *Ecological Informatics* **29**, 139–146.

Yang C, Odvody GN, Fernandez CJ, Landivar JA, Minzenmayer RR and Nichols RL (2014) Evaluating unsupervised and supervised image classification methods for mapping cotton root rot. *Precision Agriculture* **16**, 201–215.

Yang C, Odvody GN, Thomasson JA, Isakeit T and Nichols RL (2016) Change detection of cotton root rot infection over 10-year intervals using airborne multispectral imagery. *Computers and Electronics in Agriculture* **123**, 154–162.

Yao Q, Xian D-x, Liu Q-j, Yang B-j, Diao G-q and Tang J (2014) Automated counting of rice planthoppers in paddy fields based on image processing. *Journal of Integrative Agriculture* **13**, 1736–1745.

Zhang R, Ren Z, Sun J, Tang W, Ning D and Qian Y (2017) Method for monitoring the cotton plant vigor based on the WSN technology. *Computers and Electronics in Agriculture* **133**, 68–79.

Zhang J, Kong F, Wu J, Han S and Zhai Z (2018) Automatic image segmentation method for cotton leaves with disease under natural environment. *Journal of Integrative Agriculture* **17**, 1800–1814.

Zhang J, Huang Y, Pu R, Gonzalez-Moreno P, Yuan L, Wu K and Huang W (2019) Monitoring plant diseases and pests through remote sensing technology: a review. *Computers and Electronics in Agriculture* **165**, 104943.

Appendix B

A Smart-Pest Management of Cotton based on a Metacognitive Architecture

A Smart-Pest Management of Cotton based on a Metacognitive Architecture

Raul Toscano-Miranda[a,c], Jose Aguilar[b,c,e], Manuel Caro[a], Anibal Trebilcok [d], Mauricio Toro[c]

[a]Department of Educational Informatics, Universidad de Córdoba, Colombia
[b]CEMISID, Universidad de Los Andes, Mérida, Venezuela
[c]GIDITIC, Universidad Eafit, Medellín, Colombia
[d]Department of Agronomic Engineering, Universidad de Córdoba, Colombia
[e] IMDEA Networks Institute, Leganés, Madrid, Spain

Abstract

The use of information technology —in agriculture— plays an important role in Smart-Pest Management. Particularly, Artificial intelligence helps to identify, monitor, control and make decisions about pests in crops. In this paper, we present a new metacognitive architecture, called Metacognitive Architecture for a Smart-Pest Management of Cotton. Especially, this paper presents several contributions: (1) a new architecture that implements several metacognitive tasks (meta-memory, meta-learning, meta-reasoning, meta-comprehension, meta-knowledge); (2) a case study of the architecture for predictive and prescriptive problems, in the context of integrated pest management in cotton; (3) an integrated approach of data analytics and metacognition in smart systems.

Keywords: Architecture, Metacognition, Artificial Intelligence, Pest Control, Pest Insects, Insect Pest Management, Cotton Crop

1. Introduction

Cognitive Informatics is a multidisciplinary research area that investigates the internal information processing mechanisms of the brain and natural intelligence (Wang et al., 2018). *Cognitive Computing* is an emerging paradigm of *Artificial Intelligence* (AI), based on Cognitive Informatics, which implements computational intelligence by autonomous inferences and perceptions, mimicking the mechanisms of the brain and natural intelligence

(Wang et al., 2018). According to Cox & Raja (2007), to be an intelligent system should have three cognitive processes (planning, understanding and learning), and a metacognition mechanism to control and monitor of such cognitive processes.

Metacognition is cognition about cognition (Cox, 2005) and, this term in AI, refers to the ability of an intelligent system to monitor and control its own learning and reasoning processes (Caro et al., 2014; Cox & Raja, 2011). Metacognitive processes —such as monitoring, controlling and goal setting— are related to cognitive processes; therefore, they should be an integral part of a *cognitive architecture* (Sun et al., 2006).

A cognitive architecture refers to the structure of the human mind and to its computational instantiation in the fields of AI and computational cognitive science (Lieto, 2021). *Metacognitive architectures* differ from cognitive architectures in that the agent itself is the referent of the cognitive processing; however, both represent knowledge and memory to store domain content and their processes (e.g., of knowledge acquisition) (Langley et al., 2009; Sun, 2009).

Metacognitive architectures offer the following advantages to intelligent agents: (1) greater autonomy in decision-making (Cox & Dannenhauer, 2017; Paisner et al., 2014); (2) fault tolerance since the system can identify faults and fix them without human intervention (Cox & Raja, 2011; Kennedy & Sloman, 2002); and (3) a better response to unexpected events or to situations for which they were not designed (Paisner et al., 2014).

In this paper, we focus on the application of metacognition in agriculture, in particular, to cotton. In particular, *smart agriculture* (SA) plays an important role in cotton crops, including the detection and control of insect pests and diseases (Toscano-Miranda et al., 2022). SA uses the interrelationship of (1) sensor-network technology, (2) grid-computing technology, (3) context-aware computing technology, and (4) AI, in order to manage the agriculture process (Aqeel-Ur-Rehman & Shaikh, 2009).

SA includes strategies of *integrated pest management* (IPM), which seeks to minimize the environmental impact of pesticide application, and reduce risks to human and animal health (FAO, 2017; Simberloff & Rejmanek, 2011). IPM is based on very important aspects, such as the prevention and monitoring of pests and diseases, which today are being assisted by detection equipment and AI techniques.

One of the main problems faced by IPM, in cotton cultivation, is the boll-weevil control (Ben Guerrero et al., 2020; Coelho et al., 2016; Grigolli et al., 2017). Boll weevil (*Anthonomus grandis grandis*) is an insect that feeds on the squares (a part of the developing cotton fruit) and bolls of the cotton plant, and causes huge losses in cotton crops (Ellis & Horton, 1997). Recent work has demonstrated the positive effects of IPM to control boll weevil in cotton cultivation (Alves et al., 2020). Toscano-Miranda et al. (2022) showed a research opportunity related to boll weevil and smart traps to determine how the pest spreads and to place geo-spatially the traps optimally.

Cognitive architectures have been applied to the domain of agriculture. As an example, Agostini et al. (2017) presented a cognitive architecture for automatic gardening, which is composed of a decision-making framework with robotics techniques for sensing and acting to autonomously treat plants. However, nowadays, to the best of our knowledge, there is no metacognitive architecture applied to the prediction and prescription of an adequate control of boll weevil in cotton crops. Therefore, a contribution of this paper is to define a metacognitive architecture to make predictions and prescriptions for boll weevil.

The aim of this paper is to specify a metacognitive architecture to improve cotton agriculture. The contributions of this article are the following:

- A new architecture, called *Metacognitive Architecture for Smart-Pest Management of Cotton* (MASMC), which:
    - Provides autonomous interactions among its three levels and the cognitive and metacognitive tasks to solve complex problems;
    - Uses all metacognitive tasks (meta-memory, meta-learning, meta-reasoning, meta-comprehension, meta-knowledge);
    - Includes metacognitive mechanisms for meta-reasoning that allows autonomy in decision making;
    - Includes a metacognitive cycle to detect failures in problem-solving and to comprehend what happened;
    - Makes an introspection to understand where a failure is located in case of an unexpected failure of reasoning strategies;
    - Allows the interaction of different technologies —such as data analytics and metacognition— to improve cotton agriculture.
- A case study that shows the application of MASMC to predict and prescribe an adequate control of the boll weevil in cotton crops.

The rest of this paper is organized as follows. Section 2 presents the related works. Section 3 describes a case study for the management of the boll weevil in cotton crops using MASMC. Section 4 shows a comparison with previous works. Finally, Section 5 presents the conclusions and future work directions.

2. Related works

This section presents a brief state-of-the-art of cognitive and metacognitive architectures. There are several metacognitive architectures defined before, some of them are the *Metacognitive Integrated Dual-Cycle Architecture* (MIDCA), the *Connectionist Learning with Adaptive Rule Induction On-line* (CLARION), *Meta-Asking Questions and Understanding Answers* (Meta-AQUA) and *State, Operator, And Result* (SOAR). CLARION

(Sun, 2006; Sun et al., 2006) is a theoretical framework with meta-cognitive mechanisms to (1) monitor, (2) control, and (3) regulate cognitive processes. Meta-AQUA (Cox & Ashwin, 1999) is one of the first cognitive architectures. SOAR (Laird et al., 2012) is the first cognitive architecture integrated with real robots, and supports adaptive cognitive robotic agents. SOAR provides representations of memories for the short-term and long-term knowledge for decision making.

MIDCA (Cox et al., 2016) is a metacognitive, integrated, dual-cycle (object and meta-levels) architecture that has the capacity to act in a dynamic environment and manage unexpected events, planning and executing in unexpected situations. MIDCA integrates characteristics of Meta-AQUA. The metacognitive module of MIDCA uses a higher-level symbolic representation, which is opposed to CLARION that uses a sub-symbolic mechanism.

After MIDCA, three other architectures appeared. Agostini et al. (2017) presented a cognitive architecture for automatic gardening, which can autonomously treat plants. *Cognitive ARchitecture for IntelligeNt Agent* (CARINA) (Caro et al., 2019) is a metacognitive architecture to monitor and control reasoning failures in AI agents. In Particular, CARINA is an instance of *Metamodel Supporting Metacognition* (MISM) (Caro et al., 2014). MISM covers a broad range of commonly referenced concepts of metacognitive models in AI. Finally, *Abductive-Deductive Cognitive Architecture System Planned and Reactive* (AD-CASPAR) (Longo & Santoro, 2020) is a cognitive architecture leveraging natural-language processing and first-order logic inference. AD-CASPAR can reason on queries using abduction as a pre-stage of deduction.

The architecture introduced in this article, MASMC, is an extension of CARINA; therefore, MASMC inherits most of its characteristics. There are three main differences between MASMC and CARINA. First, MASMC includes a ground level to deal with in-site and remote sensing. Second, MASMC's ground level is designed —specifically— to deal with insect pests and diseases in cotton crops. Third, meta-comprehension is defined in detail in MASMC, and it was not considered in CARINA.

3. Proposed architecture

This section presents the *Metacognitive Architecture for Smart-Pest Management of Cotton* (MASMC). MASMC implements cognition and metacognition through metacognitive tasks. There are three main levels in MASMC: (1) ground-level, (2) object-level, and (3) meta-level (see Figure 1). The ground level involves interactions with the environment, for which it is integrated by several sensors. Object and meta-levels correspond to the information processing involved in metacognition, according to (Nelson & Narens, 1990). MASMC allows autonomous interactions among the three cognitive levels and the metacognitive tasks to solve complex problems.

MASMC is an AI framework of a cognitive agent. Using MASMC, a farmer can define a goal for the agent (its object-level) to solve a problem. This problem will be solved using cognitive tasks. A cognitive task is the skill of an intelligent agent to process new information (i.e., acquire and use knowledge). *Cognitive tasks* allow agents to recall information from memory to be used in the same or similar situations (Kester & Kirschner, 2012). Meta-cognitive tasks are monitoring and controlling, all the time, what the object-level does. Therefore, the meta-level improves the cognitive tasks of the object-level.

The levels of MASMC are explained with examples for the management of boll weevil. In this case study, a farmer has a cotton crop affected by Boll weevil. The farmer wants to know the behavior of the pest, monitor it, and control it. The farmer may ask the cognitive agent about some of its tasks. Tasks can be to describe, diagnose, predict or prescribe (knowledge models). The cognitive agent sends the information to the object-level through the cognitive sensors. When the meta-level and object-level finish the cognitive cycle, the results are communicated to the farmer through the cognitive sensors (see Figure 1).

MASMC is used to make predictions and prescriptions to control boll weevil in cotton crops. To describe the case study, we used the *Methodology for the development of Data Mining applications based on an organizational analysis* (MIDANO) methodology (Pacheco et al., 2014). MIDANO is useful to implement autonomous cycles of data analysis (Aguilar et al., 2022; Vizcarrondo et al., 2017). MIDANO contains three phases. (1) *Identification of sources for knowledge extraction in an organization*. In this case study, we obtained data of pheromone traps of boll weevil from the *Colombian Agricultural Institute* (ICA) and data of climatic variables from the *Colombian Institute of Hydrology, Meteorology and Environmental Studies* (IDEAM). (2) *Data preparation and processing*. In this stage, we cleaned and combined the datasets, and analyzed them. Finally, (3) *Development of tools for data mining*. In this stage, for our case study, we instanced the metacognitive architecture.

Below we describe the levels of MASMC using as case study the definition of a prescription model for boll weevil. A prescription model for boll weevil must comprehensively manage the following controls: (1) soka destruction, (2) micro and macro application of chemicals, (3) pheromone traps and, (4) boll-weevil killing tubes, among other aspects. An adequate prescription configuration is a goal that a smart autonomous system must face. In our case, it does so through the communication of the three levels —ground level, object-level and meta-level— until a solution is delivered to the farmer.

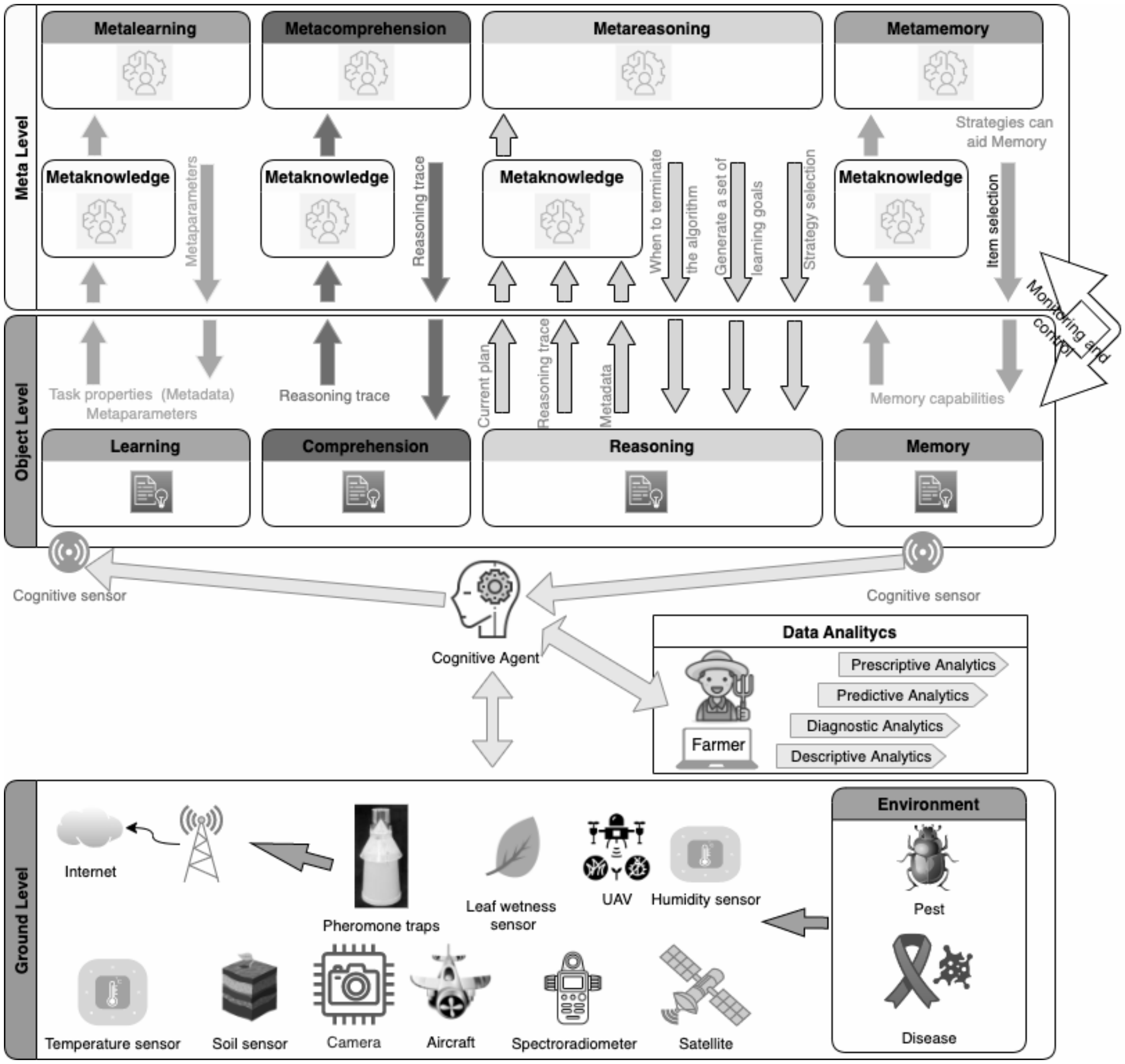


Fig 1. Metacognitive Architecture for Smart-Pest Management of Cotton (MASAC).

3.1 Ground Level

MASMC receives the information of the environment, from the *ground-level*, using pheromone traps, sensors, and cameras, or any combination of them. Thus, the *ground-level* takes the information from the environment about the pests or diseases using sensors, such as humidity sensors, leaf-wetness sensors, soil sensors, temperature sensors, cameras, pheromone traps or spectroradiometers. After, the ground level sends the information through the Internet to the cognitive agent. Finally, the cognitive agent waits for the requirements of the farmer. As an example, when the farmer needs to make a diagnosis of the boll weevil, MASMC needs to (1) count them, (2) classify them (red or black) and, (3) differentiate them from other species. In what follows, these tasks are explained:

- **Count**: Determines the quantity of boll weevils that has been captured with pheromone traps.
- **Classify**: When the boll weevils are captured, they are classified into two categories: red or black. The reds are young and the black are in the procreation stage.
- **Differentiate**: Determines if other species have migrated from other regions.

Finally, since the farmer has asked the cognitive agent to diagnose the insect pest, and then make a prediction and prescription model, these processes continue at the object level, which is explained in the next section.

Also, this level has actuators. An actuator is a device that modifies the state of the environment as a result of a decision of the cognitive system. Therefore, sensors and actuators allow monitoring and control at the ground level (Abioye et al., 2021; Ojha et al., 2015; Talavera et al., 2017).

3.2 Object-Level

Object-level refers to specific components of cognitive functions —such as cognitive tasks related to learning, reasoning, knowledge, memory and comprehension. The object-level uses the information from the environment to generate behaviors, rationally, to achieve goals for problem-solving (Caro et al., 2015; Cox & Raja, 2011).

Figure 2 shows the general procedure of the object-level. This procedure is called the *cognitive cycle*. The object-level acquires new information from the environment through the ground level (its sensors). To organize this information and the new knowledge generated, the object level uses the following memories: (1) sensory memory, (2) procedural memory, (3) semantic memory, (4) episodic memory and (5) working memory. Sensory memory can store information from the environment for a very short time (Atkinson & Shiffrin, 1968; Carlson et al., 2009). Working memory stores information to perform operations with that information (e.g., reasoning, planning) (Baddeley, 2003). Long-term memory (LTM) stores information over time (Atkinson & Shiffrin, 1968; Unsworth, 2010). LTM is subdivided into episodic memory, semantic memory, and procedural memory. Episodic memory stores specific past events, situations, and past experiences in a person' s life or in an agent's action history (Cox, 2007; Tulving, 2002). Semantic memory processes ideas and concepts that do not come from personal experience but from knowledge of the world and facts (Binder & Desai, 2011; Tulving, 1986). Procedural memory entails how to do things (Cohen & Bacdayan, 1994).

The *cognitive cycle* is defined by the following components: a perception component to get information from sensory memory; a situation assessment component to process the information perceived, which is then recognized and categorized; a reasoning component to make conclusions about the problem; a problem-solving component to solve problems related to manage a pest in cotton. This component includes planning strategies for diagnostic, and prediction and prescription models, for the best management of a pest. Finally, an action component gives an output to the environment to reach a goal.

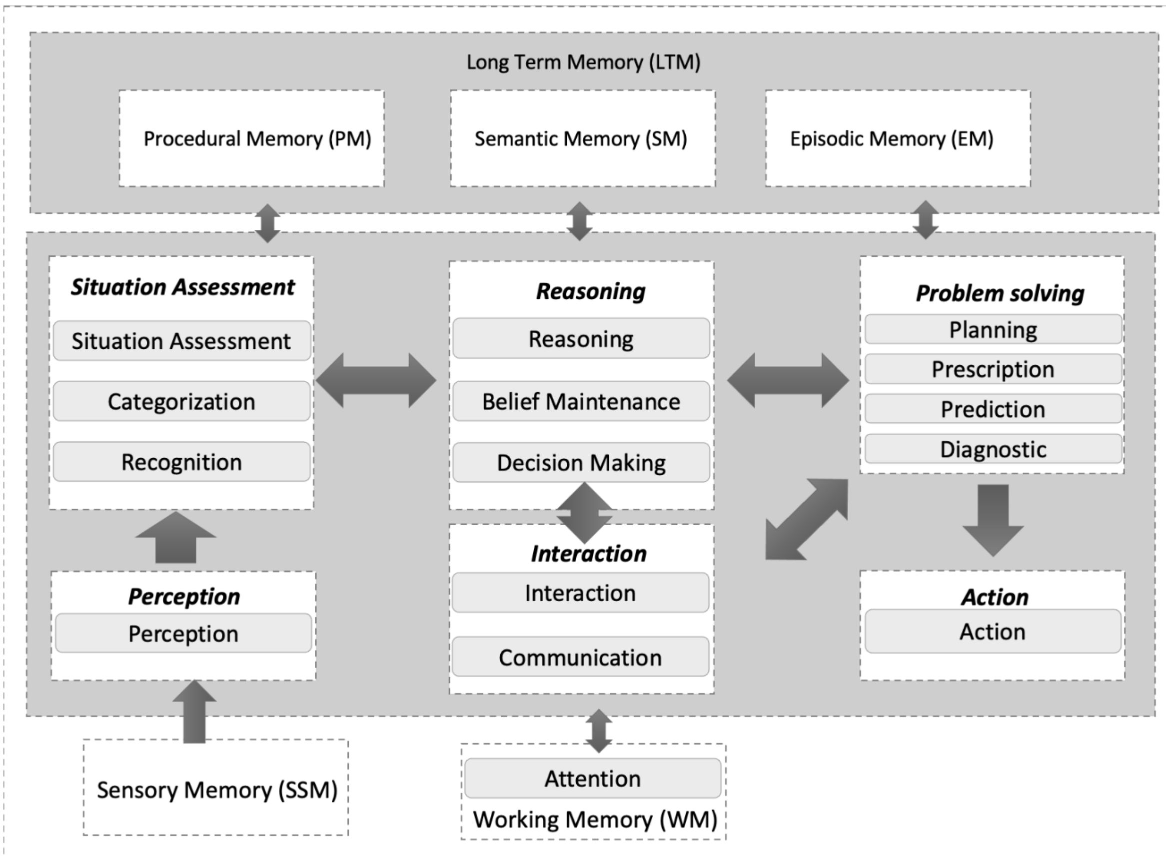


Fig. 2. Object-level components of MASMC.

In the case study, a cognitive task performed at the object-level is related to the prediction and prescription to control boll weevil. According to this case study, the farmer requests to the cognitive system a prescription model for boll weevil control (to decrease the quantity of boll weevils) in the cotton crop. Figure 3 shows the object-level cognitive process to perform cognitive tasks in the case study.

To generate the strategies for insect-pest control, in cotton crops, the following steps are taken. First, the cognitive system reads the *Current-Situation Dataset*, which contains boll weevil features (e.g., number of boll weevils and location where they were found), and applies a descriptive analysis to generate the *Current-Situation Learning Model*. The Current-Situation Learning Model is an overview of the current situation of the insect pest in the field. Second, the cognitive system makes a prediction about the behavior of the insect pest in the crop and generates a *Future-Situation Learning Model*. A Future-Situation Learning Model is a prediction about the behavior of the insect pest.

The last situation model is used for a prescriptive analysis, where a *planner* makes one or several plans. For example, a planner can make a plan with the setting to apply the most suitable agrochemical. Afterward, the *recommender* selects the best plan, which is the output of the prescription recommended to the farmer. The output is used by the *Dataset Generator* to generate a prescriptive-process dataset. With this dataset, the agent generates a Prescription Learning Model. A Prescription Learning Model is a knowledge model with the best prescription to manage the pest in cotton. Finally, the cognitive process begins a new cycle.

The cognitive tasks in this level can be defined as learning, reasoning, knowledge, memory and comprehension tasks. The *learning task* uses an algorithm (called *Learning Algorithm*) to optimize/create the knowledge models (e.g., prescriptive, descriptive and predictive). As an example, a genetic algorithm or a neural network can be used to optimize a prescription model. The Learning Algorithm is the main algorithm that solves the problem (i.e., it is used to reach the goal of the cognitive system). The function of the *reasoning task i*s to execute an *anytime algorithm*, all the time, to detect failures in the learning task. An anytime algorithm is an algorithm that works parallelly to monitor other algorithms. The *comprehension task* makes a reasoning trace (line by line) of the learning task as a story-understanding task, this leads to comprehension and understanding of the information (Cox & Raja, 2007). A *reasoning trace* records what the cognitive system did, and detects possible execution failures (e.g., the cognitive system made two prescriptions but recommended the least feasible). The *knowledge task* stores information related to other cognitive tasks.

The *memory task* aims to read the physical-memory capabilities of the system and determine if the algorithm can continue (if it has enough memory) or stop (to avoid a memory overflow). The information that generates the memory task is stored in the *Algorithmic Knowledge Profiles* (APK) (Caro et al., 2017). The APK are also called the knowledge of the world (Caro et al., 2019). APK are used for introspective monitoring in cognitive agents. The cognitive tasks are carried out through cognitive functions such as perception, recognition, reasoning, decision making, planning, prediction and communication (Madera-Doval et al., 2018).

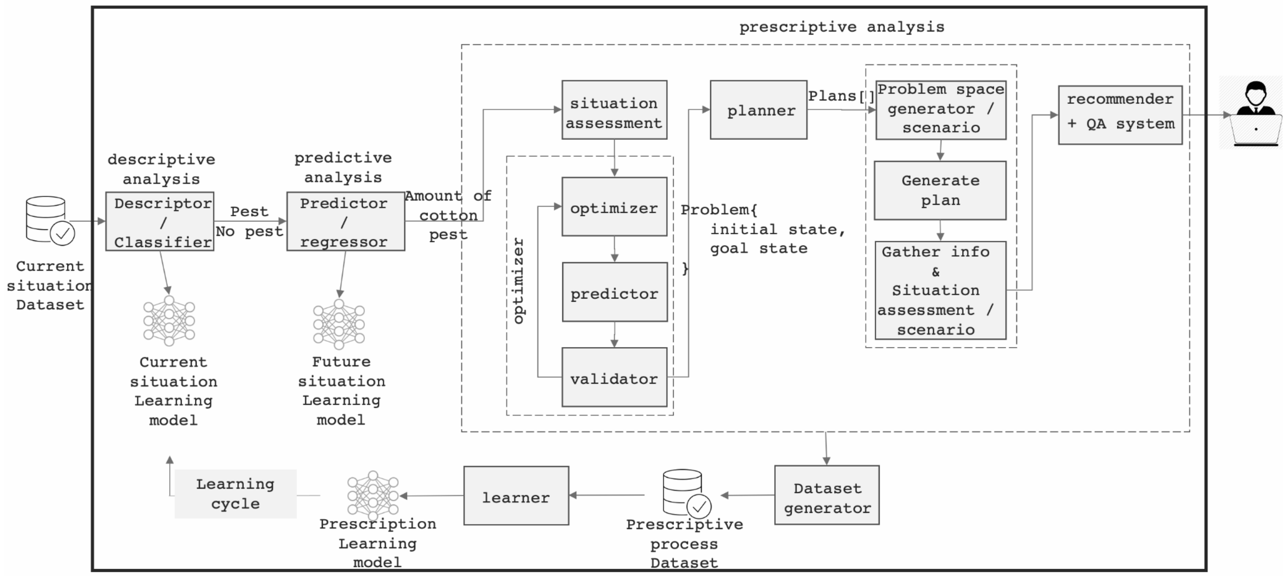


Fig. 3. Cognitive process of the object-level for prescriptive cognitive tasks.

With this preview analysis, the system can respond to what the farmer should do with the situation described.

3.3 Meta Level

This section contains the main contributions of this metacognitive architecture. The *meta-level* controls and monitors the object-level. The meta-level contains precise instructions to modify the processes that take place at the object-level. The Meta-level performs the process monitoring of the object-level, and for this, the Meta-level collects information on the cognitive tasks executing at the object-level.

The rest of this section describes the metacognitive tasks of the meta-level (meta-learning, meta-reasoning, meta-knowledge, meta-memory, and meta-comprehension), and how they communicate with the object-level cognitive tasks. The metacognitive tasks are carried out through metacognitive functions such as identification, and detection, among others (Madera-Doval et al., 2018).

Meta-learning tasks monitor the learning tasks (see Figure 4). This information is stored in the meta-knowledge of each cognitive task. *Meta-knowledge tasks* store the knowledge about the task that monitors and controls. *Meta-reasoning* tasks monitor the reasoning trace (Cox & Raja, 2011). Meta-reasoning tasks determine when an algorithm that is running in the learning task will be finished, generate a set of learning goals, explain why the learning goals, and generate a strategy of selection of the algorithms available for solving a given problem. *Meta-memory tasks* generate strategies for managing of the memory in the process of acquisition, retention, and retrieval of information (Nelson & Narens, 1990).

Finally, the meta-*comprehension tasks* make an introspection to understand a failure in the reasoning.

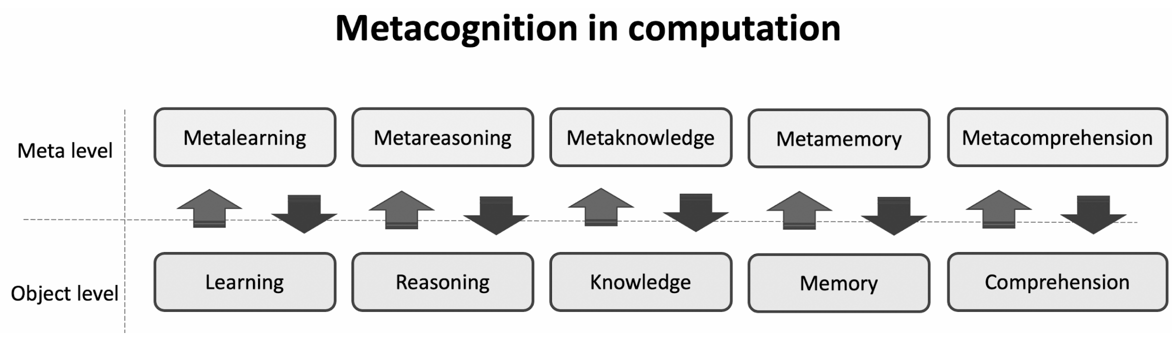


Fig. 4. A Cognitive cycle among the object-level cognitive and the meta-level.

3.3.1 Meta-learning tasks

A meta-learning task constantly monitors and controls a learning task; that is, its parameters must be changed to improve its performance (e.g., the algorithm can need to define the optimal iteration number, the optimal layer number, among other things). Figure 5 shows the interaction between a learning task, a meta-knowledge task, and a meta-learning task. Particularly, the meta-level monitors the accuracy and runtime of the learning algorithms. Also, meta-knowledge collects information about the learning task (meta-data). Some of this information is its hyper-parameters and properties. Task properties are meta-data about the learning task (e.g., type of learning algorithm required, the goal of the learning task). Hyper-parameters are the parameters of the learning algorithm of the learning task, which will be optimized by the meta-learning task. A meta-learning task uses methods that can be task-dependent or task-independent. An example of a task-dependent method defines a prescription model specific for the farmer requirement. An example of a task-independent method is to generate a general procedure to optimize the hyper-parameters of a given learning algorithm without importance where it will be applied.

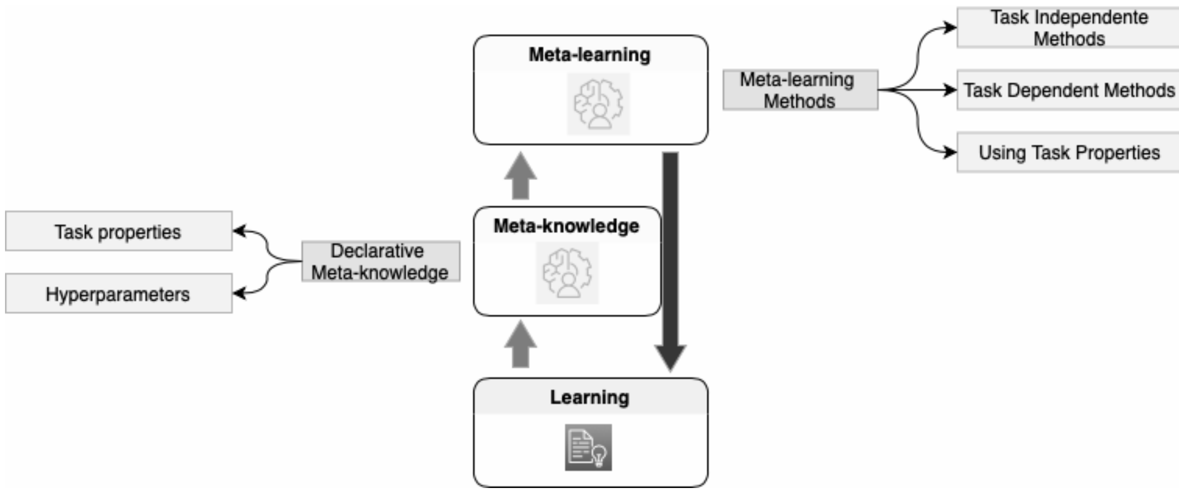


Fig. 5. Interactions among a learning task, a meta-knowledge task and a meta-learning task

The meta-knowledge task collects information about the learning task (properties and hyper-parameters, among others), which are the inputs for the meta-learning. This is declarative or semantic knowledge (Crowder & Friess, 2011; Tulving, 1986).

In the case study, the task properties can be the information of the algorithm types to solve a learning task. An example of these algorithms are: neural network, rule-based, decision tree, random forest and support vector machines (Kandalkar et al., 2014; Shahzadi et al., 2016; Xavier et al., 2019; Yang et al., 2014). Examples of the goal of the learning task are classification, prediction or prescription of insect pests.

Regarding hyper-parameters, some examples are, in the case of genetic algorithms, they need initialization parameters such as initial population, crossover probability, mutation probability, population size and the number of generations (Pongcharoen et al., 2002). In the case of neural networks, their initialization parameters are the number of layers and the number of nodes by layers, among others.

These are examples of information about the learning task that is stored in the meta-level. This knowledge is stored in the meta-knowledge of the learning task. Thus, a meta-learning task can select the best learning algorithm for a given moment. Also, it can choose the best hyper-parameters of a selected learning algorithm. For example, a meta-learning task can determine the best hyper-parameters of a neural network and command the object level to execute it. The aim is to determine the best learning algorithm, with its adequate hyper-parameters, for a learning task.

This represents two important aspects: The cognitive system (1) can choose the best algorithm for a problem, and (2) decreases the training time, due to the fact that the cognitive system does not work with training examples, only with the meta-data of the learning algorithms.

3.3.2 Meta-reasoning tasks

Meta-knowledge gathers computational performance data to build a profile of the algorithms to be used by a meta-reasoner. Particularly, when the meta-reasoner has several strategies (algorithms) to solve a problem (learning task), it takes the meta-data of each algorithm and chooses the best strategy. Meta-reasoner could involve generating explanations about the choices for the object-level, and about the effect on ground-level performance. Facing novel situations, the agent must learn from experience and create new strategies based upon its self-perceived strengths and weaknesses.

Figure 6 shows the meta-reasoner of the architecture for monitoring and controlling the object-level. The meta-reasoner is used when the reasoning task fails and is required to enrich it (Cox & Raja, 2011; Schmill et al., 2011). Meta-reasoner generally analyzes how the system makes decisions, and returns a set of learning goals after analyzing the reasoning trace, to improve previous goals. Also, the meta-reasoner must explain why reasoning fails, and why that result and not others.

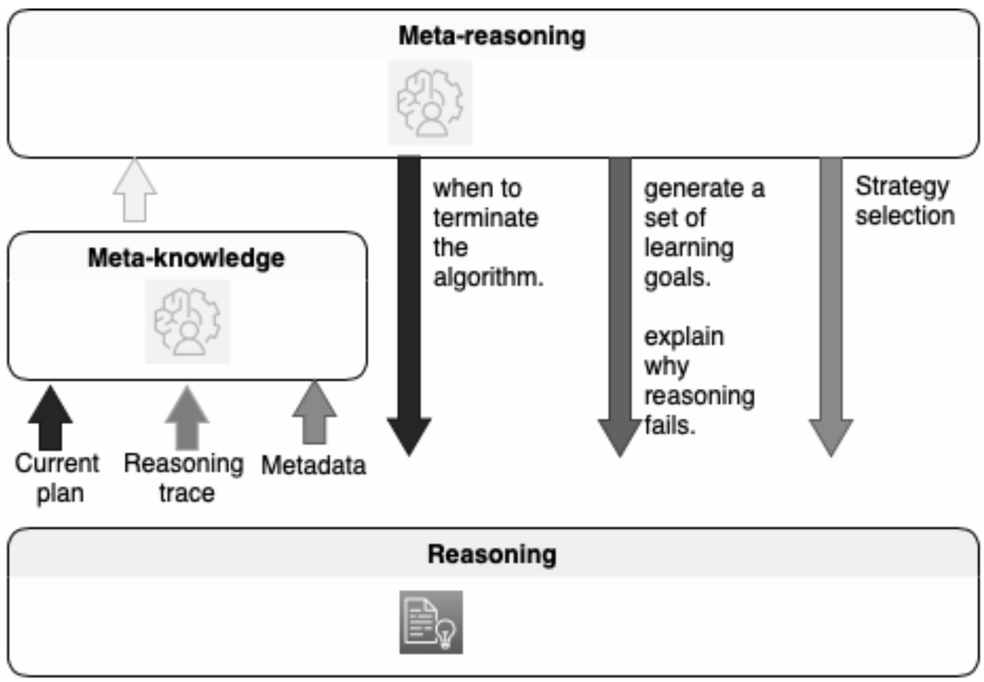


Fig. 6. Interaction among a reasoning task, a meta-knowledge task and a meta-reasoning task.

In general, a Meta-reasoning task must monitor (1) the solution quality, (2) the prospect for further improvement in solution quality, (3) the cost of doing nothing to continue reasoning in the search for a new solution, and (4) the expected utility of a solution, among other things.

For this case study, we consider that the reasoning algorithm, at the object-level, found a failure in the process of the learning task. The learning task did not provide adequate recommendations to control boll weevil with the expected accuracy. Therefore, the reasoning task generates a plan with several strategies to solve the problem in the learning task. For example, it includes another learning algorithm or try another combination of hyper-parameters for the current learning algorithms. In case of failure of the plan generated by the reasoning task, a meta-reasoning task analyzes how the reasoning task made the decision. Due to the unexpected failure of the reasoning strategy, meta-comprehension makes an introspection to understand line-by-line where the failure was.

3.3.3 Meta-comprehension tasks

Meta-comprehension is inside a meta-reasoning task, and it is executed when there is a failure in the reasoning (it is not the expected one) (Cox & Raja, 2007). After the reasoning trace is sent to the meta-level, meta-comprehension must understand the reasoning trace to help the meta-reasoning task understand the failure (see Figure 7). An example of a reasoning trace is: If there is a failure in the prescription then (1) record the failure information, (2) record the task information.

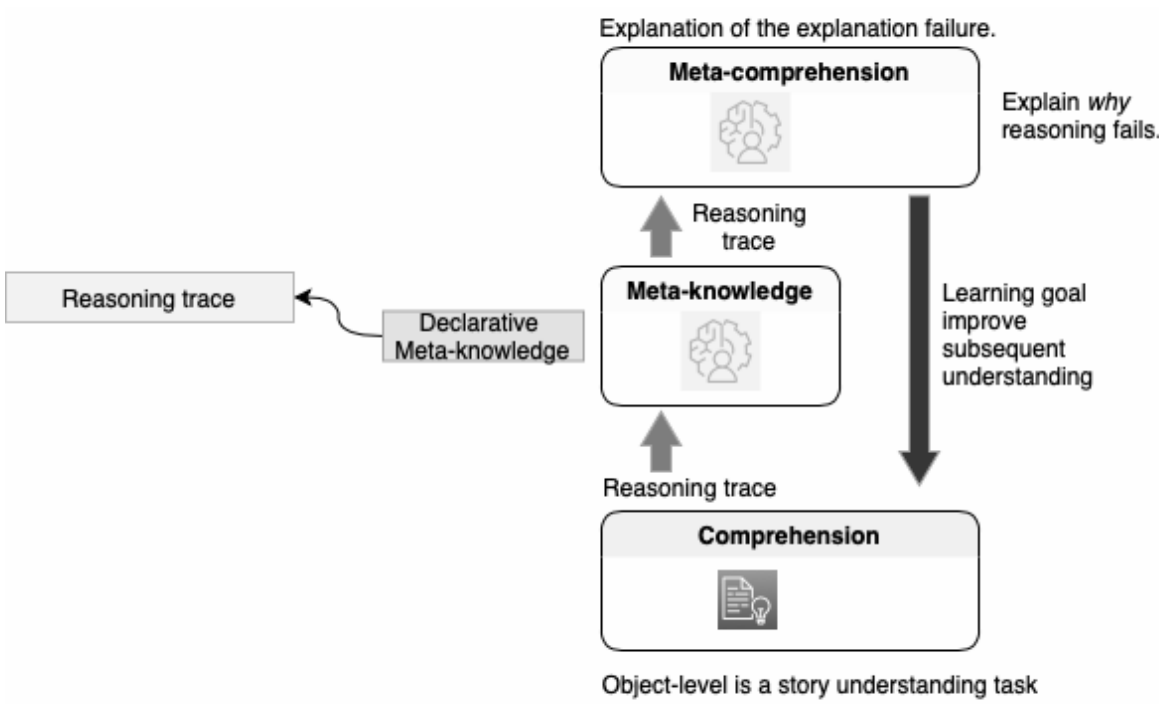


Fig. 7. Interactions among a comprehension task, a meta-knowledge task, and a meta-comprehension task.

Meta-comprehension makes an introspection to understand, line-by-line, where the failure was. It performs an introspection of the trace to obtain an explanation of the failure, called an *Introspective Meta-eXplanation Pattern* (IMXP). According to Cox & Raja (2007), an IMXP explains why reasoning fails. The output of meta-comprehension is a new learning goal to learn from the failure.

As an example, in our case study, one of the limitations of genetic algorithms is how to choose hyper-parameters such as the size of the population, mutation rate, crossover rate and the selection method. A meta-comprehension task makes an introspection to understand, line-by-line, where the genetic algorithm failed and generate a new learning goal. This learning goal defines a new configuration of parameters, such that the object-level must run again. Meta-comprehension allows a continuous improvement process.

3.3.4 Meta-memory tasks

Meta-memory refers to any judgment that is made about memory. Memory monitoring refers to assess the use—during a learning task—of allocated resources. A meta-memory task judges whether the object-level has successfully used the assigned resources. Thus, meta-memory is responsible for resource allocation decisions and the selection of relevant assignment strategies (Fairfield et al., 2015). Figure 8 shows the components of the meta-memory of the architecture to monitor and control the memory in the object-level.

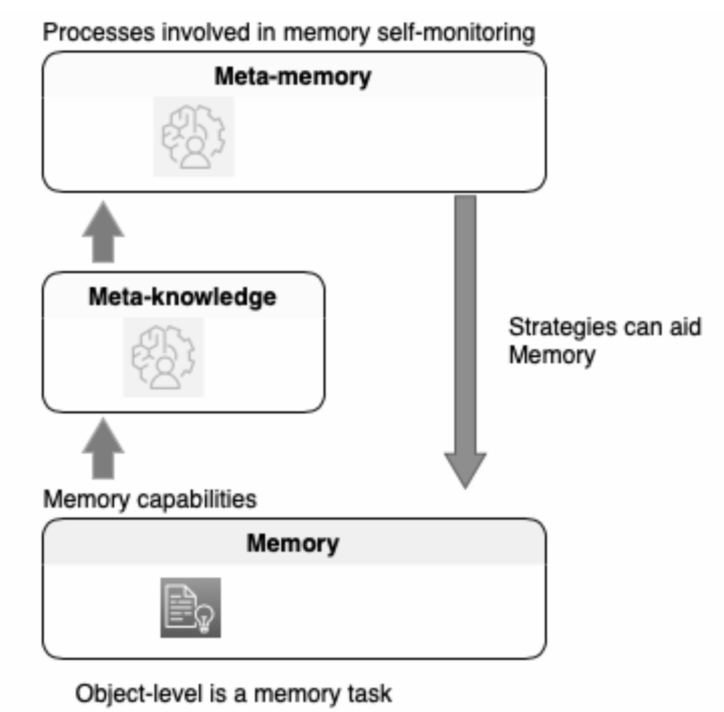


Fig. 8. Interaction among a memory task, a meta-knowledge task and a meta-

memory task.

A meta-memory task makes judgments about a learning task and defines strategies to improve its memory utilization. For this case study, the meta-memory task uses a *semantic memory* with an ontology of the insect pests. The memory task (at the object-level) determines that the learning task has taken a long time to get the results of the boll-weevil prescription. The learning task did not find the data; therefore, the meta-memory task defines a strategy; for example, to build an index in memory, for the most frequent cases. Next time, the learning task will take less time to do the data retrieval.

In general, meta-knowledge tasks are used for the rest of meta-levels. They manage the information of each metacognitive task. So, meta-learning, meta-reasoning, meta-comprehension, and meta-memory tasks have their own meta-knowledge tasks.

4. Comparison with previous works

In this section, we make a comparison with previous works to show the advantages of MASMC. Ground-level has the sensors and the mechanisms to send the information to the object-level. Object-level has cognitive tasks like learning, knowledge, reasoning, comprehension and memory. Finally, meta-level monitors and controls the object-level with meta-cognitive tasks such as meta-learning, meta-knowledge, meta-reasoning, meta-comprehension, and meta-memory. Table 1 shows a comparison of cognitive architectures with meta-levels. In Table 1, the only architectures that use all the meta-cognitive tasks are CARINA and this work.

Table 1. Comparison of MASMC with previous architectures about meta-cognitive tasks.

Cognitive Architectures	**ML**	**MK**	**MR**	**MC**	**MM**
Aqua (Cox & Ashwin, 1999)	✗	✗	✓	✓	✓
CLARION (Sun, 2006; Sun et al., 2006)	✗	✗	✗	✓	✓
MCL (Schmill et al., 2011)	✗	✗	✓	✗	✓
ONE (Singh, 2005)	✗	✗	✓	✗	✗
DMF (Kennedy & Sloman, 2003)	✗	✗	✓	✗	✗
MIDCA (Cox et al., 2016)	✗	✗	✓	✓	✓

MAMID (Hudlicka, 2010)	✗	✗	✗	✗	✓
SOAR (Laird, 2008)	N/A	N/A	✓	N/A	✓
CARINA (Caro et al., 2019)	✓	✓	✓	✓	✓
MASMC	✓	✓	✓	✓	✓

Abbreviations: Aqua= Meta-AQUA, MCL= The Meta-Cognitive Loop, ONE = EM-ONE Architecture, DMF = Distributed Metacognition Framework, Green = Cognitive Vision System, CASPAR = AD-CASPAR, MCAF = Multi-agent Cognitive Architecture Framework, MAMID = MAMID Cognitive-Affective Architecture, CASC = Cognitive Architecture for Smart Cities, MASMC = Meta-Architecture for Smart-Management of Cotton, ML= Meta-learning, MK = Meta-knowledge, MR = Meta-reasoning, MC = Meta-comprehension, MM = Meta-memory, N/A = Not Available.

Table 2 shows a comparison of cognitive architectures and their cognitive tasks at the object-level. Table 2 shows that Aqua, Clarion, MCL, MIDCA, CARINA and this work use the five cognitive tasks at the object-level. The other architectures use them partially. On the other hand, cognitive architectures such as MCAF (Shah, 2018), Green (Boissard et al., 2008), DMF (Kennedy & Sloman, 2003), CASC (Pranaya et al., 2017) have cognitive tasks (see Table 2), but not meta-cognitive tasks (see Table 1).

Table 2. Comparison of MASMC with previous architectures about cognitive tasks.

Cognitive Architectures	**L**	**K**	**R**	**C**	**M**
Aqua (Cox & Ashwin, 1999)	✓	✓	✓	✓	✓
CLARION (Sun, 2006; Sun et al., 2006)	✓	✓	✓	✓	✓
MCL (Schmill et al., 2011)	✓	✓	✓	✓	✓
ONE (Singh, 2005)	✓	✓	✓	N/A	✓
DMF (Kennedy & Sloman, 2003)	✓	✓	✓	N/A	✓
MIDCA (Cox et al., 2016)	✓	✓	✓	✓	✓
Green (Boissard et al., 2008)	✓	N/A	N/A	N/A	N/A
CASPAR (Longo & Santoro, 2020)	N/A	✓	✓	N/A	N/A
MCAF (Shah, 2018)	N/A	N/A	✓	N/A	N/A

MAMID (Hudlicka, 2010)	N/A	✓	N/A	N/A	✓
CASC (Pranaya et al., 2017)	N/A	N/A	N/A	N/A	✓
SOAR (Laird, 2008)	N/A	✓	✓	N/A	✓
CARINA (Caro et al., 2019)	✓	✓	✓	✓	✓
MASMC	✓	✓	✓	✓	✓

Abbreviations: Aqua= Meta-AQUA, MCL= The Meta-Cognitive Loop, ONE = EM-ONE Architecture, DMF = Distributed Metacognition Framework, Green = Cognitive Vision System, CASPAR = AD-CASPAR, MCAF = Multi-agent Cognitive Architecture Framework, MAMID = MAMID Cognitive-Affective Architecture, CASC = Cognitive Architecture for Smart Cities, MASMC = Meta-Architecture for Smart-Management of Cotton, L= Learning, K = Knowledge, R = Reasoning, C = Comprehension, M = Memory, N/A = Not Available.

Of the architectures presented above, only MLC and Meta-AQUA define a ground level. Thus, the main differences between our approach with the previous works are the following: (1) it uses the ground, object and meta levels in its design, (2) it integrates the five meta-cognitive tasks (meta-learning, meta-comprehension, meta-reasoning, meta-memory and meta-knowledge), (3) it has been used in a case study in agriculture to show the autonomy in decision making for the prescription of pest management strategies and better assistance to the farmer.

Now, all these cognitive architectures have specific requirements for their implementation that can be facilitated by using the multi-agent systems paradigm (Terán et al., 2017). Agent theory has developed mechanisms, methodologies, that are reusable for the implementation of cognitive architectures, which in the case of MASMC can easily be reused because it can be seen as a means of managing services for agents (e.g., its cognitive and metacognitive tasks), something that from the other architectures is more difficult to identify.

5. Conclusions

MASMC is a cognitive architecture extended from CARINA. MASMC integrates five cognitive functions at the meta-level: meta-learning, meta-knowledge, meta-reasoning, meta-comprehension, and meta-memory. The integration is designed to monitor and control the functions at the object-level to find solutions to problems. Since MASMC is an extension of CARINA, it inherits most of its characteristics. We recall from the introduction that there are two main differences between MASMC and CARINA. First, MASMC includes a ground-level to deal with sensing. Second, meta-comprehension tasks were defined.

To show the use of MASMC, we defined a case study to analyze the interactions among the three levels and their cognitive and meta-cognitive tasks to improve the results in the learning tasks for pest management (i.e., situation diagnosis, and prediction and prescription of proper pest management). Thus, we proved the applicability of the architecture in integrated pest management (IPM) with the description of the case study. We showed how to apply descriptive, predictive and prescriptive analysis for the case study. In particular, we proposed treatments for boll weevil using prescriptive models.

In future works, we plan to implement a part of the architecture in order to define an autonomous cognitive system for agriculture. For example, we will define a meta-learning task to define models of weevil behavior for different regions, using the transfer learning paradigm. Also, we will define a general rule-based diagnostic model that can be adapted to the context using appropriate metacognitive tasks. Finally, we will integrate our cognitive architecture with the multi-agent systems paradigm, to take advantage of the existing modeling capabilities and implementations in agent theory.

Acknowledgments

We thank the Colombian Agricultural Institute (ICA), Colombian Cotton Confederation (CONALGODON) and Colombian Institute of Hydrology, Meteorology and Environmental Studies (IDEAM).

Funding

This work was supported by the Colombian Science, Technology and Innovation Fund (FCTeI) of the General Royalty System (SGR); Universidad EAFIT; and Universidad de Córdoba.

Conflicts of Interest

The authors declare there are no conflicts of interest.

Ethical Approval

Not applicable

References

Abioye, E. A., Abidin, M. S. Z., Mahmud, M. S. A., Buyamin, S., AbdRahman, M. K. I., Otuoze, A. O., Ramli, M. S. A., & Ijike, O. D. (2021). IoT-based monitoring and data-driven modelling of drip irrigation system for mustard leaf cultivation experiment. *Information Processing in Agriculture*, *8*(2), 270–283. https://doi.org/10.1016/j.inpa.2020.05.004

Agostini, A., Alenyà, G., Fischbach, A., Scharr, H., Wörgötter, F., & Torras, C. (2017). A cognitive architecture for automatic gardening. *Computers and Electronics in Agriculture*, *138*, 69–79. https://doi.org/10.1016/j.compag.2017.04.015

Aguilar, J., Buendia, O., Pinto, A., & Gutiérrez, J. (2022). Social learning analytics for determining learning styles in a smart classroom. *Interactive Learning Environments*, *30*(2), 245–261. https://doi.org/10.1080/10494820.2019.1651745

Alves, A. N., Souza, W. S. R., & Borges, D. L. (2020). Cotton pests classification in field-based images using deep residual networks. *Computers and Electronics in Agriculture*, *174*(April), 105488. https://doi.org/10.1016/j.compag.2020.105488

Aqeel-Ur-Rehman, & Shaikh, Z. A. (2009). Smart agriculture. In *Applications of modern high performance networks* (Issue June 2014, pp. 120–129). Bentham Science Publishers. https://doi.org/10.2174/978160805077210901010120

Atkinson, R. C., & Shiffrin, R. M. (1968). Human Memory: A proposed system and its control processes. *The Psychology of Learning and Motivation*, *2*(5), 89–195. https://doi.org/10.1016/S0079- 7421(08)60422-3

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Ben Guerrero, E., Salvador, R., & Talia, P. (2020). Evaluation of hydrolytic enzyme activities from digestive fluids of Anthonomus grandis (Coleoptera: Curculionidae). *Archives of Insect Biochemistry and Physiology*, *July*, 1–8. https://doi.org/10.1002/arch.21730

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

Boissard, P., Martin, V., & Moisan, S. (2008). A cognitive vision approach to early pest detection in greenhouse crops. *Computers and Electronics in Agriculture*, *62*(2), 81–93. https://doi.org/10.1016/j.compag.2007.11.009

Carlson, N. R., Heth, D., Miller, H., Donahoe, J., & Martin, G. N. (2009). *Psychology: the science of behavior.* Pearson.

Caro, M., Gomez, A. A., & Giraldo, J. C. (2017). Algorithmic knowledge profiles for introspective monitoring in artificial cognitive agents. *Proceedings of 2017 IEEE 16th International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC 2017*, 475–481. https://doi.org/10.1109/ICCI-CC.2017.8109792

Caro, M., Josyula, D., Madera, D., Kennedy, C., & Gómez, A. (2019). The CARINA metacognitive architecture. *International Journal of Cognitive Informatics and Natural Intelligence*, *13*(4), 71–90. https://doi.org/10.4018/IJCINI.2019100104

Caro, M., Josyula, D. P., Cox, M. T., & Jiménez, J. A. (2014). Design and validation of a metamodel for metacognition support in artificial intelligent systems. *Biologically*

Inspired Cognitive Architectures, *9*, 82–104. https://doi.org/10.1016/j.bica.2014.07.002

Caro, M., Josyula, D. P., Jiménez, J. A., Kennedy, C. M., & Cox, M. T. (2015). A domain-specific visual language for modeling metacognition in intelligent systems. *Biologically Inspired Cognitive Architectures*, *13*, 75–90. https://doi.org/10.1016/j.bica.2015.06.004

Coelho, R. R., de Souza Júnior, J. D. A., Firmino, A. A. P., de Macedo, L. L. P., Fonseca, F. C. A., Terra, W. R., Engler, G., de Almeida Engler, J., da Silva, M. C. M., & Grossi-de-Sa, M. F. (2016). Vitellogenin knockdown strongly affects cotton boll weevil egg viability but not the number of eggs laid by females. *Meta Gene*, *9*, 173–180. https://doi.org/10.1016/j.mgene.2016.06.005

Cohen, M., & Bacdayan, P. (1994). Organizational routines are stored as procedural memory: Evidence from a laboratory study. *INFORMS*, *5*(4), 554–568. https://doi.org/10.1287/orsc.5.4.554

Cox, M. T. (2005). Metacognition in computation: A selected research review. *Artificial Intelligence*, *169*(2), 104–141. https://doi.org/10.1016/j.artint.2005.10.009

Cox, M. T. (2007). *Metareasoning, Monitoring, and Self-Explanation*.

Cox, M. T., Alavi, Z., Dannenhauer, D., Eyorokon, V., Munoz-Avila, H., & Perlis, D. (2016). MIDCA: A metacognitive, integrated dual-cycle architecture for self-regulated autonomy. *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 3712–3718.

Cox, M. T., & Ashwin, R. (1999). Introspective multistrategy learning: On the construction of learning strategies. *Artificial Intelligence*, *112*(1), 1–55. https://doi.org/10.1016/S0004-3702(99)00047-8

Cox, M. T., & Dannenhauer, Z. A. (2017). Perceptual Goal Monitors for Cognitive Agents in Changing Environments. *Advances in Cognitive Systems*, 1–16.

Cox, M. T., & Raja, A. (2007). *Metareasoning : A Manifesto*. 1–4.

Cox, M. T., & Raja, A. (2011). Metareasoning: An Introduction. In M. T. Cox & A. Raja (Eds.), *Metareasoning: Thinking about thinking* (Issue 3, pp. 3–14). MIT Press. https://doi.org/10.7551/mitpress/9780262014809.003.0001

Crowder, J. A., & Friess, S. (2011). Metacognition and metamemory concepts for AI systems. *Proceedings of the 2011 International Conference on Artificial Intelligence, ICAI 2011*, *2*(July 2011), 419–424.

Ellis, H. C., & Horton, D. L. (1997). *Boll Weevil*. https://web.archive.org/web/20150219082851/http://www.bugwood.org/factsheets/99-001.html

Fairfield, B., Mammarella, N., Palumbo, R., & Di Domenico, A. (2015). Emotional meta-memories: A review. *Brain Sciences*, 509–520. https://doi.org/10.3390/brainsci5040509

FAO. (2017). Integrated pest management. In *Integrated Pest Management* (Vol. 4). https://doi.org/10.1007/978-94-007-7802-3

Grigolli, J. F. J., Souza, L. A., Fernandes, M. G., & Busoli, A. C. (2017). Spatial Distribution of Adult Anthonomus grandis Boheman (Coleoptera: Curculionidae) and Damage to Cotton Flower Buds Due to Feeding and Oviposition. *Neotropical Entomology*, *46*(4), 442–451. https://doi.org/10.1007/s13744-016-0471-1

Hudlicka, E. (2010). Modeling Cultural and Personality Biases in Decision-Making. *Proceedings of the 3rd Intl. Conference on Applied Human Factors and Ergonomics*, *May*, 550–559. https://doi.org/10.1201/ebk1439834954-c55

Kandalkar, G., Deorankar, A. V, & Chatur, P. N. (2014). Classification of Agricultural Pests Using DWT and Back Propagation Neural Networks. *International Journal of Computer Science and Information Technologies (IJCSIT)*, *5*(3), 4034–4037.

Kennedy, C. M., & Sloman, A. (2002). *Reflective Architectures for Damage Tolerant Autonomous Systems*.

Kennedy, C. M., & Sloman, A. (2003). Autonomous recovery from hostile code insertion using distributed reflection. *Cognitive Systems Research*, *4*(2), 89–117. https://doi.org/10.1016/S1389-0417(02)00096-7

Kester, L., & Kirschner, P. (2012). Encyclopedia of the Sciences of Learning. In *Encyclopedia of the Sciences of Learning*. Springer. https://doi.org/10.1007/978-1-4419-1428-6

Laird, J. E. (2008). Extending the soar cognitive architecture. *Frontiers in Artificial Intelligence and Applications*, *171*(1), 224–235.

Laird, J. E., Kinkade, K. R., Mohan, S., & Xu, J. Z. (2012). Cognitive robotics using the soar cognitive architecture. *AAAI Workshop - Technical Report*, *WS-12-06*, 46–54.

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*(2), 141–160. https://doi.org/10.1016/j.cogsys.2006.07.004

Lieto, A. (2021). *Cognitive design for artificial minds*. Routledge.

Longo, C. F., & Santoro, C. (2020). AD-CASPAR: Abductive-Deductive Cognitive Architecture based on Natural language and first order logic Reasoning. *CEUR Workshop Proceedings*, *2735*(November), 73–86.

Madera-Doval, D. P., Caro, M., Gómez-Salgado, A. A., Cardozo-Soto, A. M., & Jiménez-Builes, J. A. (2018). Design of metacognitive expectations of cognitive functions through ontological representations. *DYNA (Colombia)*, *85*(206), 194–201.

Nelson, T. O., & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings. *Psychology of Learning and Motivation - Advances in Research and Theory*, *26*(C), 125–173. https://doi.org/10.1016/S0079-7421(08)60053-5

Ojha, T., Misra, S., & Raghuwanshi, N. S. (2015). Wireless sensor networks for agriculture: The state-of-the-art in practice and future challenges. *Computers and Electronics in Agriculture*, *118*, 66–84. https://doi.org/10.1016/j.compag.2015.08.011

Pacheco, F., Rangel, C., Aguilar, J., Cerrada, M., & Altamiranda, J. (2014). Methodological Framework for Data Processing based on the Data Science Paradigm. *XL Latin American Computing Conference*. https://doi.org/10.1109/CLEI.2014.6965184

Paisner, M., Cox, M. T., Maynord, M., & Perlis, D. (2014). Goal-Driven Autonomy for Cognitive Systems. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Pongcharoen, P., Hicks, C., Braiden, P. M., & Stewardson, D. J. (2002). Determining optimum Genetic Algorithm parameters for scheduling the manufacturing and assembly of complex products. *International Journal of Production Economics*, *78*(3), 311–322. https://doi.org/10.1016/S0925-5273(02)00104-4

Pranaya, Y. C., Himarish, M. N., Baig, M. N., & Ahmed, M. R. (2017). Cognitive architecture based smart grids for smart cities. *3rd International Conference on Power Generation Systems and Renewable Energy Technologies, PGSRET 2017*, *2018-Janua*(April), 44–49. https://doi.org/10.1109/PGSRET.2017.8251799

Schmill, M., Anderson, M., Fults, S., Josyula, D., Oates, T., Perlis, D., Shahri, H., Wilson, S., & Wright, D. (2011). The metacognitive loop and reasoning about anomalies. In Cox, Michael T & A. Raja (Eds.), *Metareasoning: Thinking about thinking* (pp. 183–198). MIT Press.

Shah, V. S. (2018). Multi-agent cognitive architecture-enabled IoT applications of mobile edge computing. *Annales Des Telecommunications/Annals of Telecommunications*, *73*(7–8), 487–497. https://doi.org/10.1007/s12243-018-0648-1

Shahzadi, R., Ferzund, J., Tausif, M., & Asif, M. (2016). Internet of Things based Expert System for Smart Agriculture. *International Journal of Advanced Computer Science and Applications*, *7*(9), 341–350. https://doi.org/10.14569/ijacsa.2016.070947

Simberloff, D., & Rejmanek, M. (2011). Encyclopedia of biological invasions. In *Choice Reviews Online* (Vol. 48, Issue 11). Univ of California Press. https://doi.org/10.5860/choice.48-6030

Singh, P. (2005). *EM-ONE: An Architecture for Reflective Commonsense Thinking* (Issue 1998). Massachusetts Institute of Technology.

Sun, R. (2006). The CLARION cognitive architecture: Extending cognitive modeling to social simulation. In *Cognition and multi-agent interaction* (pp. 79–99).

Sun, R. (2009). Theoretical status of computational cognitive modeling. *Cognitive Systems Research*, *10*(2), 124–140. https://doi.org/10.1016/j.cogsys.2008.07.002

Sun, R., Zhang, X., & Mathews, R. (2006). Modeling meta-cognition in a cognitive architecture. *Cognitive Systems Research*, *7*(4), 327–338. https://doi.org/10.1016/j.cogsys.2005.09.001

Talavera, J. M., Tobón, L. E., Gómez, J. A., Culman, M. A., Aranda, J. M., Parra, D. T., Quiroz, L. A., Hoyos, A., & Garreta, L. E. (2017). Review of IoT applications in agro-industrial and environmental fields. *Computers and Electronics in Agriculture*, *142*(118), 283–297. https://doi.org/10.1016/j.compag.2017.09.015

Terán, J., Aguilar, J., & Cerrada, M. (2017). Integration in industrial automation based on multi-agent systems using cultural algorithms for optimizing the coordination mechanisms. *Computers in Industry*, *91*, 11–23. https://doi.org/10.1016/j.compind.2017.05.002

Toscano-Miranda, R., Toro, M., Aguilar, J., Caro, M., Marulanda, A., & Trebilcok, A. (2022). *Artificial Intelligence and Sensing Techniques for the Management of Insect Pests and Diseases in Cotton Crops: A Systematic Literature Review*.

Tulving, E. (1986). What kind of a hypothesis is the distinction between episodic and semantic memory? *American Psychological Association*.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, *53*(1).

Unsworth, N. (2010). On the division of working memory and long-term memory and their relation to intelligence: A latent variable approach. *Acta Psychologica*, *134*(1), 16–28. https://doi.org/10.1016/j.actpsy.2009.11.010

Vizcarrondo, J., Aguilar, J., Exposito, E., & Subia, A. (2017). MAPE-K as a service-oriented Architecture. *IEEE Latin America Transactions*, *15*(6), 1163–1175.

Wang, Y., Howard, N., Kacprzyk, J., Frieder, O., Sheu, P., Fiorini, R. A., Gavrilova, M., Patel, S., Peng, J., & Widrow, B. (2018). Cognitive informatics: Towards cognitive machine learning and autonomous knowledge manipulation. *International Journal of Cognitive Informatics and Natural Intelligence*, *12*(1), 1–13.

https://doi.org/10.4018/IJCINI.2018010101

Xavier, T. W. F., Souto, R. N. V., Statella, T., Galbieri, R., Santos, E. S., Suli, G. S., & Zeilhofer, P. (2019). Identification of Ramularia Leaf Blight Cotton Disease Infection Levels by Multispectral, Multiscale UAV Imagery. *Drones*, *3*(2), 33. https://doi.org/10.3390/drones3020033

Yang, C., Odvody, G. N., Fernandez, C. J., Landivar, J. A., Minzenmayer, R. R., & Nichols, R. L. (2014). Evaluating unsupervised and supervised image classification methods for mapping cotton root rot. *Precision Agriculture*, *16*(2), 201–215. https://doi.org/10.1007/s11119-014-9370-9

Appendix C

A Classification Model of Cotton Boll-Weevil Population

A Classification Model of Cotton Boll-Weevil Population

Raul Toscano-Miranda
Department of Educational Informatics
Universidad de Córdoba, Universidad Eafit
Montería, Colombia
ORCID: 0000-0003-0260-9234

William Hoyos
Department of Bacteriology
Universidad de Córdoba, Universidad Eafit
Montería, Colombia
ORCID: 0000-0002-9165-8208

Manuel Caro
Department of Educational Informatics
Universidad de Córdoba
Montería, Colombia
ORCID: 0000-0001-6594-2187

Jose Aguilar
CEMISID, Universidad de Los Andes, Mérida, Venezuela;
GIDITIC, Universidad Eafit, Medellin, Colombia
Dpto de Automática Universidad de Alcalá, Alcalá, Spain
ORCID: 0000-0003-4194-6882

Anibal Trebilcok
Department of Agronomic Engineering
Universidad de Córdoba
Montería, Córdoba
atrebilcok@correo.unicordoba.edu.co

Mauricio Toro
GIDITIC
Universidad Eafit
Medellín, Colombia
ORCID: 0000-0002-7280-8231

***Abstract*— Integrated pest management (IPM) seeks to minimize the environmental impact of pesticide application. IPM is based on two important aspects —prevention and monitoring of diseases and insect pests— which today are being assisted by sensing and artificial-intelligence (AI). Particularly, AI helps to identify, monitor, control and make decisions about pests in crops. In this paper, we present a comparison among five machine-learning models to classify the population of the boll weevil in cotton into three classes: low, medium and high. Weather data (average daily rainfall, humidity and temperature) were used to classify the population of the boll weevil in the department of Córdoba, Colombia. The results showed that XGBoost obtained the highest accuracy (88%). Results showed that it is possible to classify boll-weevil populations using weather data.**

Keywords— Data analysis, pest control, insect pest management, cotton crop, machine learning, XGBoost, weather

I. INTRODUCTION

Cotton (*Gossypium hirsutum L.)* is the main source of natural textile fiber and one of the most important oil crops [1]. Cotton contains 49 species distributed throughout most tropical and subtropical regions of the world. The world's cotton industry represents a multibillion-dollar enterprise, from the production of raw fiber to finished textile products [2]. Between 2016 and 2017, 32.4 million hectares were planted in more than 80 countries [3].

Pests and diseases —in cotton crops— generate large economic losses. If they are not controlled in time, that is, at an early stage, they can cause an infestation, and decrease the production yield and quality of the harvested product [4]. As an example, in Brazil, annual losses, in agricultural production, due to pests, can reach an average of 7.7%, equivalent to, approximately, US$ 17.7 billion [5]. Entomological and pathogenic problems are one of the causes of low yields and economic losses in cotton crops [6], [7].

Toscano-Miranda et al. [8] showed a research opportunity related to boll weevils and smart traps to determine how boll weevil spreads and where to place the traps. Boll weevil is a pest directly affecting cotton production [9]–[11]. Adults feed on fruiting forms, leaf petioles and terminal growth [12]. Since 2000, Colombian Agricultural Institute (ICA in Spanish) implemented a strategy to monitor boll weevil (*Anthonomus grandis Boheman*)[13]. This strategy consists of periodically checking the number of boll weevils in each trap to know the population fluctuation in the country. Weevil captures are made using pheromone traps, which have a specific location with global positioning system (GPS) coordinates. The updated information on the population of the boll weevil helps timely decision-making for the management of the pest.

There is a need to develop computational strategies to help detect and classify the presence of boll-weevil in cotton crops, to take measures to avoid economic losses. Therefore, the objective of this study is to develop a machine-learning model to automatically classify the boll-weevil population using data collected from pheromone traps and weather data. Cotton cultivation in Córdoba, Colombia generates an income for a significant number of families [14]. For this reason, Córdoba was chosen as a case study for this article.

II. STATE-OF-THE-ART

Alves et al. [15] used *convolutional neural networks* (CNN) to classify 13 insect pests, including boll weevil. They used a

978-1-6654-7671-3/22/$31.00 ©2022 IEEE

modified deep residual learning (RestNet34*) with images of insects. The developed model performed excellently with an accuracy of 97.8%.

Several studies have included AI and traps for insect pests. Cho et al. [16] implemented image processing to detect pests in a controlled environment like a greenhouse. Martin et al. used sticky traps for whiteflies, aphids and trips [17]. Martin et al. developed a decision-support system for early pest-detection based on video analysis and scene interpretation from multi-camera data to reduce pesticide use [17]. Martin et al. studied whiteflies and aphids in a rose greenhouse. Xia et al. [18] used sticky traps to monitor and take samples of whiteflies, aphids and thrips in tomato crops.

Using weather data and its relationship with insect population, Skawsang et al. [19] applied artificial neural networks (ANN), random forest (RF) and multiple linear regression (MLR) to forecast the brown planthopper (*Nilaparvata lugens*) population using weather and host-plant phenology factors in rice paddy fields. Nyabako et al. [20] developed models to predict P. truncatus infestation and maize-grain damage. Nyabako et al. considered DT, k-nearest neighbors, multi-layer perception, support vector machines (SVM) and MLR with weather data, in their study.

Unfortunately, so far, no studies have included the use of pheromone traps for boll weevils, nor the use of weather data and ML models to classify the population of boll weevils. Therefore, this paper presents a classification model for the boll-weevil population in cotton.

III. METHODOLOGY

Data were obtained from Cordoba, Colombia. In particular, from the cities of Cereté, Lorica, Ciénaga de Oro, Montería, Cotorra and Valencia. In these cities, there are reports of boll-weevil captured by pheromone traps. The readings of pheromone traps are done every 15 days. The records from the six cities were used; however, after data preprocessing, the records from Valencia and Cotorra were removed because of missing values.

The dataset has records from May 2013 to May 2016. Data was gathered by the ICA. The dataset has 13,585 samples (Table 1). Each record has the following features: city, route, trap code, name, GPS, date of reading, the entity performing the reading, crop stage, red boll weevils, black boll weevils, category, average daily rainfall, humidity, maximum temperature and comments. The red boll weevils are the youngest and the black weevils are the ones that can procreate. Weather data (rainfall, humidity, and maximum temperature) were obtained from the *Institute of Hydrology, Meteorology and Environmental Studies* (IDEAM in Spanish). In data processing, trap captures were merged with weather data.

According to [13], [21], [22], the ideal conditions for the boll weevil are temperatures between 24 °C and 28 °C, and humidity between 60% and 90%. According to Villarreal et al. (2005), a heavy rainy day can also negatively affect the weevil population. In the dataset, the values for rainfall were between 0 mm and 17.12 mm (mean of 3.3 with a standard deviation of 3.38), relative humidity was between 68.39% - 89.84%, (mean of 79.5 with a standard deviation of 4.7) and maximum temperature was between 31.30 °C and 37.25 °C (mean of 33.5 with a standard deviation of 1.2).

TABLE I. QUANTITY OF SAMPLES FOR EACH CITY OF CÓRDOBA

City	Total
Cereté	6,015
Lorica	1,800
Ciénaga de oro	985
Montería	1,162
Cotorra	1,928
Valencia	1,693
Total in Córdoba	13,585

Data processing removed rows that did not have all complete data, and removed outliers for humidity, rainfall, temperature, and the number of red and black boll weevils. After processing, the number of samples was 6,785.

The number of red and black boll weevils were converted into classes using data ranges. The boll-weevil population was grouped as follows: low (0 to 4), medium (5 to 20), and high (> 20). These ranges were given by the ICA. The distribution of classes for boll-weevil risk was unbalanced (Table 2). For this reason, SMOTE oversampling was used [23]. In addition, data were standardized. However, Ciénaga de Oro and Montería were not oversampled because they had few samples of the high class of red weevils.

TABLE II. DISTRIBUTION OF CLASSES FOR BOLL WEEVIL IN CORDOBA

Class	Red boll weevils	Black boll weevils
Low (0 to 4),	6,456	4,701
Medium (5 to 20)	304	1,244
High (> 20).	83	808

To select meaningful features, the feature importance of random forest was used. The following features were considered: maximum temperature, humidity, rainfall and city. The results showed that maximum temperature was the most relevant and the least relevant was the city (Figure 1).

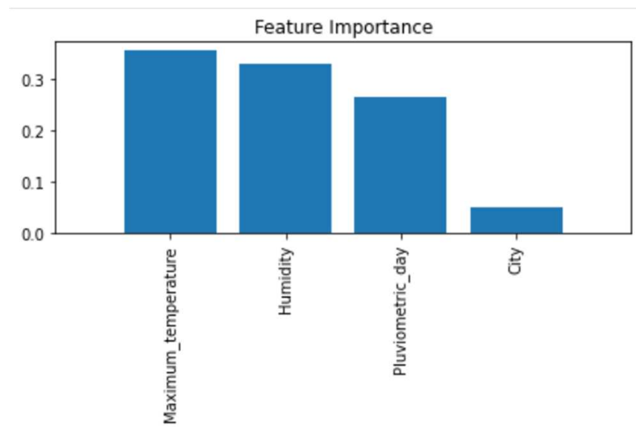


Fig. 1. Results of feature importance using random forest.

Finally, the dataset was divided into 80% for training and 20% for testing. We used cross-validation to find the best hyperparameters in each model developed. The evaluation metrics used were accuracy and F1. 80%-20% has been used due to the low number of data.

IV. Results

In this section, experimental studies are explained and a study case is presented. Five techniques were tested: Extreme Gradient Boosting (XGBoost), SVM, ANN, RF and DT. They were selected because (1) they are techniques that have shown good performance on structured data, and (2) according to the literature review [8], they are the most widely used techniques for the classification of structured data. The classification models were tested separately (for red and black boll weevils). The experiments were carried out with three weather features: rainfall, maximum temperature and humidity.

A. Experimental studies

The results showed that the best algorithm was XGBoost, outperforming SVM, ANN, RF and DT (see Table IV). Table III shows the comparison of the best hyperparameters for each model. The hyperparameters were calculated using 10-fold cross-validation. The evaluation metrics used were accuracy and F1.

TABLE III. THE BEST HYPERPARAMETERS FOR EACH ALGORITHM FOR RED AND BLACK BOLL WEEVILS

Algorithm	Best hyperparameters
XGBoost	Mtry = 1 Minimum n = 39 Tree depth = 13 Learn rate = 0.0459 Loss reduction = 0.0189 Sample size = 0.973
SVM	Kernel: radial-based kernel C = 1000 gamma = 0.01
ANN	512 units in hidden layer. Alfa = 0.01 Relu activation function Adam as optimizer
RF	Number of estimators = 1522 Minimum samples to split = 5
DT	Maximum depth = 15 Minimum samples to split = 2

Abbreviations: ANN = Artificial neural network (multilayer perceptron), SVM = Support Vector Machines, RF = Random Forest, DT = Decision Trees, XGBoost = Extreme Gradient Boosting (trees).

B. Study case

Of the models tested for red boll weevils, XGBoost obtained the best results with an 82% of accuracy and ANN was the worst (70%) (see Table IV). The results of RF and DT were very similar to XGBoost (probably because XGBoost is based on trees). All the models, except ANN, had over 80% of accuracy; however, the accuracy to predict black boll weevils was lower, always below 60%.

TABLE IV. RESULTS OF FIVE CLASSIFICATION MODELS USING RAINFALL, HUMIDITY AND MAXIMUM TEMPERATURE

Model	Red boll weevils				Black boll weevils			
	Accuracy		*F1-Score*		*Accuracy*		*F1-Score*	
	Training	*Test*	*Training*	*Test*	*Training*	*Test*	*Training*	*Test*
XGBoost	0.82	**0.82**	0.82	**0.82**	0.60	**0.60**	0.59	**0.59**
SVM	0.80	0.80	0.80	0.80	0.51	0.51	0.51	0.51
ANN	0.70	0.70	0.70	0.70	0.47	0.47	0.47	0.47
RF	0.81	0.81	0.81	0.81	0.58	0.58	0.58	0.58
DT	0.81	0.81	0.81	0.81	0.58	0.58	0.58	0.58

Abbreviations: XGBoost = Extreme Gradient Boosting (trees), RF = Random Forest, SVM = Support Vector Machines, ANN = Artificial Neural Networks, DT = Decision Trees.

Experiments were also conducted using only rainfall for the whole department of Cordoba and for each of the cities. The model was applied to all Cordoba and to each of the cities. Results showed that the model had less accuracy using a single feature than using the three features (see Tables V and VI).

TABLE V. RESULTS OF THE MODEL OF CLASSIFICATION USING XGBOOST ALGORITHM AND RAINFALL

City	Red boll weevil				Black boll weevil			
	Accuracy		*F1-Score*		*Accuracy*		*F1-Score*	
	Training	*Test*	*Training*	*Test*	*Training*	*Test*	*Training*	*Test*
Córdoba	0.75	0.74	0.75	0.73	0.57	0.56	0.57	0.55
Cereté	0.67	0.65	0.67	0.65	0.52	0.49	0.52	0.49
Lorica	0.78	**0.73**	0.78	0.73	0.60	0.56	0.60	0.56
Ciénaga	FoO	FoO	FoO	FoO	0.69	0.64	0.69	0.64
Monteria	FoO	FoO	FoO	FoO	0.82	**0.70**	0.82	0.70

Abbreviations: XGBoost = Extreme Gradient Boosting, FoO= Fail on oversample.

Since XGBoost gave the best results, and maximum temperature was the most significant feature, new experiments were carried out using only maximum temperature (see Table VI). Accuracy and F1 for the red boll weevils —on the training dataset— were improved from 82% using the three features to 83% using only maximum temperature. This means there is overfitting.

TABLE VI. RESULTS OF XGBOOST, FOR RED AND BLACK BOLL WEEVILS, USING THE MAXIMUM TEMPERATURE

Model	Red boll weevils				Black boll weevils			
	Accuracy		*F1-Score*		*Accuracy*		*F1-Score*	
	Train	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*
XGBoost	**0.83**	0.79	**0.83**	0.79	0.62	0.59	0.62	0.59

Abbreviations: XGBoost = Extreme Gradient Boosting (trees).

XGBoost was also used with the three features for each city (see Table VII). In this case, results showed that accuracy was better for Lorica, Ciénaga de Oro and Cereté with black boll weevils, and for Lorica with red boll weevils. Similar results were found for Cereté for red boll weevil. A model trained with data from all Córdoba (including the samples of Cereté, Lorica and Ciénaga de Oro) gave less accuracy for red and black boll weevils. Ciénaga de Oro failed due to oversampling (the number of captures was mainly in the low class). Montería did not have the three features: it only had available maximum temperature and rainfall.

TABLE VII. XGBOOST MODELS FOR CLASSIFICATION USING MAXIMUM TEMPERATURE, RAINFALL AND HUMIDITY

Model	Red boll weevil				Black boll weevil			
	Accuracy		*F1-Score*		*Accuracy*		*F1-Score*	
	Train	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*
*Córdoba	0.82	0.82	0.82	0.82	0.60	0.60	0.59	0.59
Cereté	0.78	0.77	0.78	0.77	0.57	0.52	0.57	0.52
Lorica	0.88	**0.88**	0.88	0.88	0.66	0.58	0.66	0.58
Ciénaga de Oro	FoO	FoO	FoO	FoO	0.71	**0.69**	0.71	0.69
Monteria	NH	NH	NH	NH	NH	NH	NH	NH

*Córdoba (included Cereté, Lorica, and Ciénaga de Oro). Abbreviations: XGBoost = Extreme Gradient Boosting, FoO = Fail on oversample, NH = No humidity.

Taking into account previous results, a final experiment was made. In this case, using models with the highest accuracy: Lorica for red boll weevils and Montería for black boll weevils. These models were tested for all the other cities to determine if the best model of one city could make a better classification for other cities. Unfortunately, accuracy decreased: Cereté decreased from 77% to 48% for red boll weevils, and from 52% to 29% for black boll weevils.

V. Conclusions and Future Work

Weather data (rainfall, humidity, and temperature) were used to classify the population of the boll weevil by classes for Córdoba, Colombia. The model that had the best accuracy was XGBoost compared to the other five algorithms. The temperature was the feature with the highest importance (analyzed using RF). Results showed that it is possible to classify a boll-weevil population using weather data.

A first simulation was carried out with the objective to determine which was the algorithm with the best performance. The algorithms tested were XGBoost (the best), SVM, ANN, RF, and DT. A second simulation was carried out to determine the precision using the rainfall variable, data from the department and cities. In this case, it was observed that the precision decreased. A third simulation was done to determine the accuracy using the maximum temperature variable. The results were better, and therefore, the maximum temperature was more significant than the rainfall. A fourth simulation was carried out to compare the results by cities using the three features (rainfall, humidity, and maximum temperature). The combination of three features improved the previous results in three cities (Lorica, Ciénaga de Oro, and Cereté with black boll weevil). In another city (Cereté with red boll weevil), they were similar, and in another (Montería), the results could not be verified because this city has not these variables. A fifth simulation was carried out, in which the objective was to determine how the best city model (with the highest precision) behaved when entering the data of the cities with less precision. Unfortunately, the accuracy dropped drastically.

Finally, it is concluded that 1) the combination of the three variables of weather data provided greater precision to the classification model, and 2) the information of the worst cities cannot predict using the model of the best city.

In future works, we plan to integrate this work with an autonomous cognitive system for agriculture as part of a metacognitive architecture. We will define a meta-learning task to define models of weevil behavior for different regions, using the transfer learning paradigm. Also, we will define a general rule-based diagnostic model that can be adapted to the context using appropriate metacognitive tasks. Finally, these models will integrate our cognitive architecture with the multi-agent systems paradigm, to take advantage of the existing modeling capabilities and implementations in agent theory.

Acknowledgments

This work is financed by a scholarship "Doctoral Excellence of the Bicentennial 2019-2020 Colciencias" supported by the Colombian Science, Technology and Innovation Fund (FCTeI) of the General Royalty System (SGR); Universidad EAFIT and Universidad de Córdoba. We would also like to thank the Colombian Agricultural Institute (ICA), Colombian Cotton Confederation (CONALGODON) and Colombian Institute of Hydrology, Meteorology and Environmental Studies (IDEAM).

References

[1] R. Zhang, Z. Ren, J. Sun, W. Tang, D. Ning, and Y. Qian, "Method for monitoring the cotton plant vigor based on the WSN technology," *Computers and Electronics in Agriculture*, vol. 133, pp. 68–79, 2017, doi: 10.1016/j.compag.2016.12.009.

[2] C. W. Smith and J. T. Cothren, *Cotton: Origin, History, Technology, and Production*. Texas: John Wiley & Sons, 1999.

[3] J. Carvalho, L. Ruiz, F. and A. O.- FAO, and Gobierno de Brasil, *Cooperación sur-sur trilateral estudio nichos de mercados del algodón*. Santiago de Chile: FAO and ABC, 2018.

[4] N. El-Wakeil and A. M. M. Abdallah, "Cotton pests and the actual strategies for their management control," in *Cotton: Cultivation, varieties and uses*, vol. 7, no. September 2012, B. Giuliano and E. Vinci, Eds. UK: Nova Science Publishers, 2012, pp. 161–218.

[5] C. M. Oliveira, A. M. Auad, S. M. Mendes, and M. R. Frizzas, "Crop losses and the economic impact of insect pests on Brazilian agriculture," *Crop Protection*, vol. 56, pp. 50–54, 2014, doi: 10.1016/j.cropro.2013.10.022.

[6] M. Anees and S. A. Shad, "Chapter 11 Insect pests of cotton and their management," in *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies*, Multan, Pakistan: Springer, 2020, pp. 177–212. doi: 10.1007/978-981-15-1472-2_11.

[7] S. Chohan, R. Perveen, M. Abid, M. Nouman, and M. Sajid, "Chapter 13 Cotton Diseases and Their Magement," in *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies*, Multan, Pakistan: Springer, 2020, pp. 239–270.

[8] R. Toscano-Miranda, M. Toro, J. Aguilar, M. Caro, A. Marulanda, and A. Trebilcok, "Artificial Intelligence and Sensing Techniques for the Management of Insect Pests and Diseases in Cotton Crops: A Systematic Literature Review," 2022.

[9] E. Ben Guerrero, R. Salvador, and P. Talia, “Evaluation of hydrolytic enzyme activities from digestive fluids of Anthonomus grandis (Coleoptera: Curculionidae),” *Archives of Insect Biochemistry and Physiology*, no. July, pp. 1–8, 2020, doi: 10.1002/arch.21730.

[10] J. F. J. Grigolli, L. A. Souza, M. G. Fernandes, and A. C. Busoli, “Spatial Distribution of Adult Anthonomus grandis Boheman (Coleoptera: Curculionidae) and Damage to Cotton Flower Buds Due to Feeding and Oviposition,” *Neotropical Entomology*, vol. 46, no. 4, pp. 442–451, Aug. 2017, doi: 10.1007/s13744-016-0471-1.

[11] R. R. Coelho *et al.*, “Vitellogenin knockdown strongly affects cotton boll weevil egg viability but not the number of eggs laid by females,” *Meta Gene*, vol. 9, pp. 173–180, Sep. 2016, doi: 10.1016/j.mgene.2016.06.005.

[12] H. C. Ellis and D. L. Horton, “Boll Weevil,” 1997. https://web.archive.org/web/20150219082851/http://www.bugwood.org/factsheets/99-001.html (accessed Mar. 10, 2020).

[13] N. Villarreal, V. Lobatón, J. Galindo, and V. Mondragón, *Monitoreo del Picudo del Algodonero Anthonomus grandis Boheman en Colombia*. 2005.

[14] A. Trabilcok, “El cultivo del algodonero en Córdoba,” Monteria, Colombia, 2020.

[15] A. N. Alves, W. S. R. Souza, and D. L. Borges, “Cotton pests classification in field-based images using deep residual networks,” *Computers and Electronics in Agriculture*, vol. 174, no. April, p. 105488, 2020, doi: 10.1016/j.compag.2020.105488.

[16] J. Cho *et al.*, “Automatic identification of whiteflies, aphids and thrips in greenhouse based on image analysis,” *Iinternational Journal of Mathematics and Computers in Simulation*, vol. 1, no. 1, pp. 46–53, 2007.

[17] V. Martin, S. Moisan, B. Paris, and O. Nicholas, “Towards a video camera network for early pest detection in greenhouses,” *ENDURE International Conference on Diversifying crop protection, France*, no. October, pp. 12–15, 2008.

[18] C. Xia, T. S. Chon, Z. Ren, and J. M. Lee, “Automatic identification and counting of small size pests in greenhouse conditions with low computational cost,” *Ecological Informatics*, vol. 29, no. P2, pp. 139–146, 2014, doi: 10.1016/j.ecoinf.2014.09.006.

[19] S. Skawsang, M. Nagai, N. K. Tripathi, and P. Soni, “Predicting rice pest population occurrence with satellite-derived crop phenology, ground meteorological observation, and machine learning: A case study for the Central Plain of Thailand,” *Applied Sciences (Switzerland)*, vol. 9, no. 22, pp. 1–19, 2019, doi: 10.3390/app9224846.

[20] T. Nyabako, B. M. Mvumi, T. Stathers, S. Mlambo, and M. Mubayiwa, “Predicting Prostephanus truncatus (Horn) (Coleoptera: Bostrichidae) populations and associated grain damage in smallholder farmers’ maize stores: A machine learning approach,” *Journal of Stored Products Research*, vol. 87, p. 101592, May 2020, doi: 10.1016/j.jspr.2020.101592.

[21] CONALGODÓN, SENA, SAC, and FFA, “Biología y manejo del picudo del algodonero en la región Costa-Llanos,” *Confederación Colombiana de Algodón, Bogotá*, 2008.

[22] O. Burbano-Figueroa, A. Sierra-Monro, L. Grandett-Martinez, C. Borgemeister, and E. Luedeling, “Estrategias de control del picudo del algodón (Anthonomus grandis Boheman) en el Caribe Colombiano: Un modelo conceptual.,” *AgriXiv*, 2019, doi: 10.31220/osf.io/db8nu.

[23] A. Gosain and S. Sardana, “Handling class imbalance problem using oversampling techniques: A review,” *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 79–85, 2017, doi: 10.1109/ICACCI.2017.8125820.

Appendix D

A fuzzy classification system to analyze the yield of cotton production

A fuzzy classification system to analyze the yield of cotton production

Raul Toscano-Miranda[a,c], Jose Aguilar[b,c,e,f], Manuel Caro[a], Anibal Trebilcok[d], Mauricio Toro[c]

[a]Department of Educational Informatics, Universidad de Córdoba, Colombia
[b]CEMISID, Universidad de Los Andes, Mérida, Venezuela
[c]GIDITIC, Universidad Eafit, Medellín, Colombia
[d]Department of Agronomic Engineering, Universidad de Córdoba, Colombia
[e]Universidad de Alcalá, Departamento de Automática, Spain
[f]IMDEA Networks Institute, Leganés, Madrid, Spain

Abstract:

CONTEXT

Properly managing the cultivation of cotton is essential because it directly impacts the amount of cotton that is produced.

OBJECTIVE

The aim of this work is the proposal of a fuzzy classification system for diagnosis-prediction tasks of the cotton crop yield.

METHODS

We used a soft computing method to handle/describe experts' knowledge. Seven input variables (attack level of the red boll weevil, attack level of the black boll weevil, crop stage, rainfall, fertilizer, pheromone traps, and boll-weevil killing tube) were considered in the system to analyze the cotton production.

RESULTS AND CONCLUSIONS

System tests were carried out on different agricultural scenarios, to determine their robustness and adaptability. According to the results, the fuzzy system has the capability to generate outputs that correspond with the experts' evaluations, which can be used to help farmers select the best practices in cotton crop management, in order to obtain the best yield in a specific context.

SIGNIFICANCE

The developed models enhance our capacity to predict crop yields based on climate data, the soil and pest behaviors, a valuable indicator for decision-making and overall sustainability.

Keywords: Fuzzy System; Classification System; Cotton; Yield; Diagnostic model; Predictive model

1. Introduction

Precision agriculture faces significant difficulties in accurately forecasting and diagnosing crop yield, making it one of the most demanding tasks. Currently, extensive research in agriculture is being carried out to enhance crop yield predictions and diagnoses, with the use of machine learning (ML) algorithms (Lobell et al., 2013; Obsie et al., 2020; Wang et al., 2018). The significance of predicting/diagnosing cotton yield in agricultural areas lies in its potential to enhance cotton crop management practices. Several techniques of Artificial Intelligence (AI) have been used to face the challenge to predict/diagnose crop yield. We focus on *fuzzy classification systems* (FCS). The use of FCS, a soft computing technique that replicates the way humans reason and make decisions, has become prevalent in modeling complex systems. To obtain FCS, experts are consulted to determine the input variables of the system and define the cause-and-effect relationship among the variables using the 'if-then' rules (Cerrada et al., 2005).

FCS has been used in many different scientific fields for modeling and decision-making. For example, recently in transportation, for predicting the speed limits on Brazilian highways (Lanzaro & Andrade, 2022); in medicine, for diagnosing child anemia (Boadh et al., 2022); for predicting chronic kidney diseases (Hamedan et al., 2020); in business, for recommendation about consumer preference (Mandal et al., 2021); in social science, for improving students' learning performances (Hwang et al., 2020). On the other hand, expert systems have been proposed in agriculture for decision-making and decision-support tasks (Toscano-Miranda, Toro, et al., 2022). More specifically, expert systems have been developed and applied in different fields of agriculture to give advice and make management decisions (Mansour & Abu-Naser, 2019; Mendes et al., 2019; Salman & Abu-Naser, 2019).

On the other hand, for crop yield prediction, some studies have demonstrated the validity of using variables such as fertilizers (Mourhir et al., 2017; Prabakaran et al., 2018), and climatic data (Holzman & Rivas, 2016; Maskey et al., 2019; Obsie et al., 2020). Also, the variables rainfall, humidity, and temperature are useful to predict the attack level of the boll weevil (Toscano-Miranda, Hoyos, et al., 2022). However, so far there are no studies that integrate the above variables with the level of pest attack, crop stages, and techniques for pest monitoring and control. In this work, we included the variables pheromone traps and boll-weevil killing tube that are used to monitor and control the boll weevil, and the variable fertilizer that helps with soil conditions. All the previous variables are related to the right management at a specific crop stage, in this way, these variables together help to the diagnosis/prediction of crop yield.

This work aims to apply a soft computing technique, specifically FCS, to build a diagnosis/predictive model useful to infer the cotton yield in the smart agriculture context. The contributions of this work are the following:

- An adaptive model based on an FCS for the diagnosis/prediction of cotton yield, validated by experts, with a very low error rate;
- An approach for the management of uncertainty based on the fuzzy variables of the context;

- An integration scheme with a previous work (Toscano-Miranda, Hoyos, et al., 2022) that determines the classification of the population of pests (boll weevil) as the level attack ate the cotton crop, used as an input variable in this investigation;
- The simultaneous utilization of climatic variables, level attack of pests, information on fertilizers and crop stage, and techniques for the monitoring and control of pests. Our results show that the use simultaneous of these variables leads to good yield prediction/diagnosis.

The rest of the paper is organized as follows: Section 2 surveys the works related to crop yield prediction/diagnosis using AI. Section 3 outlines the development of the FCS for analyzing cotton yield. Section 4 presents the results of the proposed system, and Section 5 concludes this work by highlighting some of the future directions.

2. Related works

This section explores research that is relevant to the purpose of this study, namely, the use of AI for predicting/diagnosing crop yields. Papageorgiou et al. (2013) proposed a yield prediction in apples using a knowledge-based approach with a dynamic influence graph of *Fuzzy Cognitive Maps* (FCMs). The authors used main soil factors such as soil texture (clay and sand content), soil electrical conductivity, potassium, phosphorus, organic matter, calcium, and zinc. They classified apple yield using an efficient FCM learning algorithm, the non-linear Hebbian learning, and compared it with the conventional FCM tool and benchmark machine learning algorithms.

Maskey et al. (2019) investigated the correlation among various weather parameters related to strawberry yield at the field level, and evaluated yield forecasts using the predictive principal component regression (PPCR) and two ML techniques: (a) a single-layer neural network (NN) and (b) a generic random forest (RF). They used eight attributes: two wetness counts, two wetness minutes, the ambient, canopy and soil temperatures, and the volumetric moisture. Correlation analysis showed that all parameters were significantly correlated with strawberry yield, and provided the potential to develop weekly yield forecasting models. In general, the ML technique showed better skills in predicting strawberry yields when compared to the principal component regression. More specifically, the NN provided the most skills in forecasting strawberry yield.

The climate data also were used by Ali et al. (2018) and Lobell et al. (2013). Ali et al. (2018) proposed a hybrid genetic programming model integrated with the Markov Chain Monte Carlo (MCMC) based Copula technique. The climate data included rainfall, mean monthly temperature, and mean monthly relative humidity, for the years 1981 to 2013. In a similar way, Lobell et al. (2013) used non-linear regression for the prediction of maize yield. The authors showed that there was a strong negative yield response to temperatures above 30 ºC, and a relatively weak response to the seasonal rainfall.

In addition to climatic data, Obsie et al. (2020) used data generated by the Wild Blueberry Pollination Model to predict the best yield. Obsie and colleagues used Multiple linear regression (MLR), boosted decision trees (BDT), RF, and extreme gradient boosting (XGBoost). They found that clone size, honeybee, bumblebee, Andrena bee species, Osmia bee species, maximum upper-temperature ranges, and the number of days with precipitation, were the best predictor variables. The results showed that the XGBoost outperformed other algorithms in all measures of model performance for predicting the yield of wild blueberries.

On the other hand, Ranjan & Parida (2019) employed Sentinel-based both optical (Sentinel-2B) and SAR (Sentinel-1A) sensors data for paddy acreage mapping. The authors used RF classification technique for the paddy acreage mapping. A simple linear regression yield model was developed for predicting yields. Wang et al. (2018) also used remote sensing data. They applied Deep Transfer Learning for predicting soybean crop yields. The transfer learning was applied from Argentina's data to Brazil's data because Brazil had a smaller amount of data. They used the imagery of the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite, which was processed with the long short-term memory (LSTM) recurrent NN.

Thus, there is a lot of interest in this area, as has been pointed out by (Klompenburg et al., 2020). In the scope of this work, which seeks to develop an FCS for crop yield prediction/diagnosis, there are not many works. The closest work is that of Papageorgiou et al. (2011), which used FCM with variables related to the soil. However, our proposal shows the possibility of integrating other contextual variables that have not been included so far for crop yield prediction/diagnosis. In addition, for the yield management of cotton crops, no papers were found.

Therefore, to our knowledge, there are no studies a) that simultaneously perform a diagnosis/prediction of cotton crop yields; b) that integrate different models of knowledge to characterize the context (in this work, our FCS with the work (Toscano-Miranda, Hoyos, et al., 2022)); c) that simultaneously use different types of variables (climate, behavior of pests, fertilizers, among others); d) that use adaptive models for the diagnosis/prediction of crop yields. In this way, our study focuses on these gaps.

3. Process for the Development of the FCS

This section presents the process to build the FCS for the diagnostic/prediction of cotton yield. For the development of the FCS was used the methodology proposed by (Buchanan, 1983) (see Fig. 1). This methodology has been previously used for implementing expert systems (Brüngel et al., 2019; Ele et al., 2014; Grüger et al., 2022).

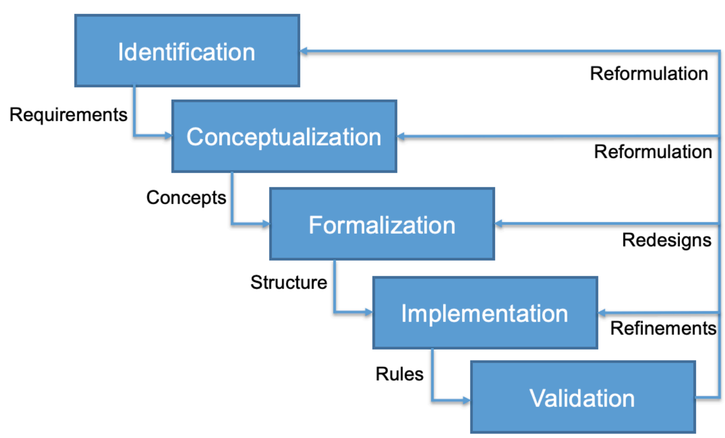


Fig. 1 Adaptation of Buchanan's model

First, let us introduce the concept of fuzzy logic, as it is used in this work. Fuzzy logic is one of the most useful AI techniques (Navinkumar et al., 2020), used to build FCS with the skill to simultaneously handle numerical data and linguistic knowledge (Mendel, 1995).

To build an FCS, it is necessary to define a rule set that represents the business logic. These rules are written in the next format: IF <x> THEN <y>. With these rules, in our case, we can determine the conditions of the cotton crop to estimate the possible crop yield. We can carry out a diagnosis using the conditions of the cotton crops, and the estimation of the yield of the crop would be the prediction of its behavior. To define the fuzzy rules, it is necessary to define the fuzzy variables. Each fuzzy variable is defined by its fuzzy sets, with the boundaries given by the specific context.

Using these key concepts, the five stages of the methodology of C(Buchanan, 1983) in our case are: 1) Identification: In this stage, the characteristics of the cotton yield problem are determined, and the input/output variables were identified for the prediction/diagnosis tasks (this is described in detail in the Input Variable Analysis subsection of the next section). 2) Conceptualization: The fuzzy sets for each input/output variable are established (this is described in detail in the Membership Features subsection of the next section). 3) Formalization: The fuzzy rules for the prediction/diagnosis tasks are designed (this is described in detail in the Fuzzy Rules subsection). 4) Implementation: The results of the previous stages (conceptualizations and designs) are computationally implemented (in our case, we have used the scikit-fuzzy library of Python. This is described in the Defuzzification subsection). 5) Validation: Finally, at this stage, the FCS (its rules) are validated against expert opinion (this is described in detail in the Results section).

4. Case study

4.1 Contextualization

Cotton crops located in various regions of Córdoba, Colombia, were considered as the case study to validate our FCS. These regions involve cities that make up the Sinú Valley (High Sinú, Middle Sinú and Bajo Sinú) (Trebilcok, 2020). This area is located at ~8°55'33.6"N, 75°48'16.5"W.

Independently of any other circumstance, the water supply associated with the planting date is determining to a high degree its effect on yields and the general behavior of the plant. However, the Sinú area has a very extended planting date due to different circumstances, being the main crop rotation of the area with the cultivation of corn, which directly and indirectly is causing an effect on the population dynamics of pests, mainly for the boll-weevil, which makes it pertinent to study the collateral effects that this situation implies. The Sinú Valley area is characterized by a high relative humidity sustained during almost every month of the year, a situation that, together with the high temperature, planting date, rainfall, and population density, creates the conditions for the development of pathogens (Trebilcok, 2020).

This work uses the results of the previous work of (Toscano-Miranda, Hoyos, et al., 2022), which gives the classification of the population of pests (boll weevil) with the level attack at the cotton crop. The predictor variables are temperature, humidity, and rainfall. The prediction (i.e., the population of red and black boll weevils) is used as input variables of the FCS. In addition, in the current investigation, we use another 5 input variables: crop stage, rainfall, fertilizer, pheromone traps, and boll-weevil killing tube. In total, are 7 input variables to estimate the output variable (crop yield). Also, for the regions that make up the area to be planted, the space of time is assumed to be more than 60 days (Trebilcok, 2020).

The crop stages included in the FCS were vegetative, flowering, fruiting, harvesting, destruction of soca, and closing), and the date of planting the cotton in the study region is around August-September. There are several insect pests in the cotton crops, however, for this case study, we used data from boll weevil because of the attack severity in this region (Trebilcok, 2020).

Next, experts were surveyed to gather their viewpoints on various factors linked to the input variables and cotton yield. The information of the survey was used for the creation of the membership functions and fuzzy rules (see survey design subsection). Finally, the system's effectiveness was assessed using multiple performance metrics.

4.2 Analysis of input variables

A group of professionals was requested to assess the significance of multiple factors for the diagnosis/prediction of the cotton yield. Their evaluations were documented on a Likert Scale with five categories that range from weak influence to strong influence. So, the FCS was programmed with seven selected variables as inputs. These variables include the attack level of the red boll weevil, the attack level of the black boll weevil, crop stage, rainfall, fertilizer, pheromone traps, and the boll-weevil killing tube.

For each input variable, the experts defined numerical ranges for the membership functions. The selection of these variables was based on experience in cotton crop management. The information for these variables was obtained by 1) A network of monitoring of boll weevils implemented by the *Colombian Agricultural Institute* (ICA), 2) Pheromone traps used in each cotton crop implemented by the owner, 3) Records of climate data from the *Institute of Hydrology, Meteorology and Environmental Studies* (IDEAM), 4) Records of climate data obtained in each cotton crop implemented by the owner, 5) Records of the management of each cotton crop implemented by the owner. 6) Crop yields in the area according to the *Colombian Cotton Confederation* (CONALGODON). The characterization of each input variable is as follows:

- Attack level of the red boll weevil: Population of the red boll weevil in the cotton crop. This data is processed and is the output of the previous work of (Toscano-Miranda, Toro, et al., 2022), which used an XGBoost algorithm for the classification of the population of the boll weevil into low (0 to 4), medium (5 to 20), and high (> 20). For this, they used climate data (temperature, humidity, and rainfall). The presence of the boll weevil must be monitored and controlled so that the attack level is zero.
- Attack level of the black boll weevil: It is similar to the red boll weevil, but in this case with blacks.
- Crop stage: Crop stage in the year. The cotton crop has the next stages: vegetative, flowering, fruiting, and harvesting. Then, it is necessary the destruction of soca, and during the closed season, another rotation crop is grown (in the context of the case study, it is usually corn).
- Rainfall: Amount of rain. Three linguistic variables were established: Low, Medium, and High. The desirable is Medium.
- Fertilizer: Amount of fertilizer used in the crop. Nitrogen, phosphorus, and potassium (NPK) fertilizers are used. For this case study, the measure used was bulks applied per hectare.
- Pheromone traps: Number of traps used in the crop. The traps are proportional to the area

of the lot monitored by an agronomist. The ICA has a minimum size of 1 trap per 250 hectares.

- Boll-weevil killing tube: Number of tubes used in the crop. The tubes are proportional to the area of the lot monitored by an agronomist.

4.3 Survey design

A set of questions was designed to collect expert assessments of the state of the crops and the probable cotton yield per hectare. The expert analyzed several scenarios about the crop yield of cotton. Scenarios are a combination of data (i.e., input variables) describing a real context and are used to validate the FCS. The scenarios were presented to the experts, and they adjusted the scenarios using linguistic variables. Each input variable has between two and six linguistic variables. Table 1 displays an overview of the input variables, their fuzzy sets, descriptions, and ranges. The next section contains a comprehensive explanation of the variables and ranges. To classify crop yields, experts in cotton cultivation were asked to sort them into low, medium, or high fuzzy sets, and specify a numeric value between 0 and 6 ton/ha, as is shown in Appendix I.

Table 1

Summary of the input variables.

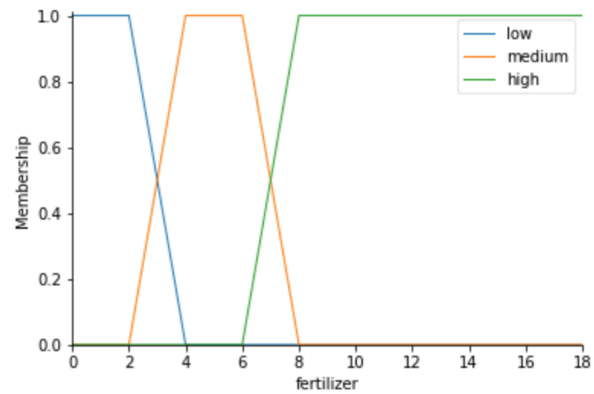
Input variable	Description	Fuzzy sets	Range
Attack level of the red boll weevil	Population of the red boll weevil in the cotton crop	Low, Medium, and High	[0, 150]
Attack level of the black boll weevil	Population of the black boll weevil in the cotton crop	Low, Medium, and High	[0, 200]
Crop stage	Crop stage in the year	Vegetative, Flowering, Fruiting, Harvesting, Destruction of soca, and Closing	[0, 12]
Rainfall	Amount of rain	Low, Medium, and High	[0, 17]
Fertilizer	Amount of fertilizer used in the crop	Low, Medium, and High	[0, 18]
Pheromone traps	Number of traps used in the crop	Absent, Adequate	[0, 1]
Boll-weevil killing tube	Number of tubes used in the crop	Absent, Adequate	[0, 1]

Nine scenarios were designed, which are defined by specific values for the input/output variables. Each variable was defined by a fuzzy value (e.g., low, medium, and high). Thus, each scenario has a different combination of situations for a crop yield waited (i.e., low, medium, or high). For example, one of the scenarios is: 15 red boll weevils (medium attack level), 15 black boll weevils (medium attack level), 0.5 in crop stage (vegetative), 17 mm of rainfall (high), 1 bulk of fertilizer (low), absent pheromone trap (0), and absent boll-weevil killing tube (0). This scenario triggers a specific set of rules from our FCS.

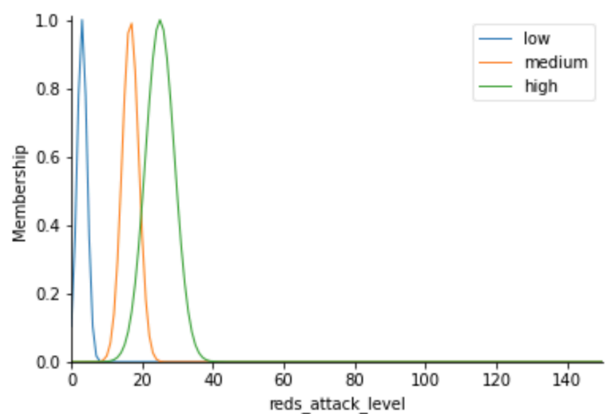
4.4 Membership functions

Each variable's fuzzy set was defined using triangular, trapezoidal, and Gaussian membership functions. The triangular function is used for the categorical variables, and the trapezoidal functions are used for the rests. Incorporating Gaussian functions into fuzzy systems enhances the ability to capture uncertainty and nonlinearity.

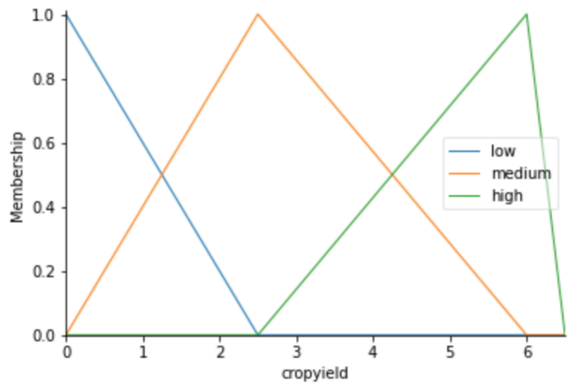
Linguistic variables, their corresponding fuzzy sets, and membership functions were constructed based on the responses from experts. The membership functions were generated for all seven input variables and the output variable, using the same principles as for the input variables. The same considerations for the input variables were used for the output variable. Fig. 2 shows examples of different membership functions for several of the fuzzy variables of our problem.



A) Trapezoidal Member Function for the Fertilizer variable.



B) Gaussian Member Function for the reds_attack_level variable



C) Triangular Member Function for the Crop Yield variable

Fig. 2 Examples of membership functions for different variables of our model.

Finally, 13 membership functions were designed for the input and output variables (see Table 2). For the crop stage, pheromone trap, and boll-weevil killing tube variables, the Gaussian membership function was not used. This design decision is due to different reasons. In the case of the crop stage variable, the planting dates are established by the ICA, and therefore, this question was not asked of the experts. In the case of the pheromone trap and the boll-weevil killing tube, the options are absent or present. Thus, these three variables were not included in the survey, whence they do not have a mean and standard deviation to generate the Gaussian membership function.

Table 2

Membership functions for each fuzzy variable.

Variable	Membership functions	
Red attack level	Trapezoidal	Gaussian
Black attack level	Trapezoidal	Gaussian
Crop stage	Trapezoidal	
Rainfall	Trapezoidal	Gaussian
Fertilizer	Trapezoidal	Gaussian
Pheromone trap	Triangular	
Boll-weevil killing tube	Triangular	
Crop yield	Triangular	Gaussian

Particularly, the membership functions were defined using the viewpoints of experts obtained from the questionnaire. One example of the fuzzy membership functions for the fertilizer variable, an antecedent input, is illustrated in Fig. 3. Thus, Fig. 3 shows the membership functions of the fertilizer variable mentioned in Table 2.

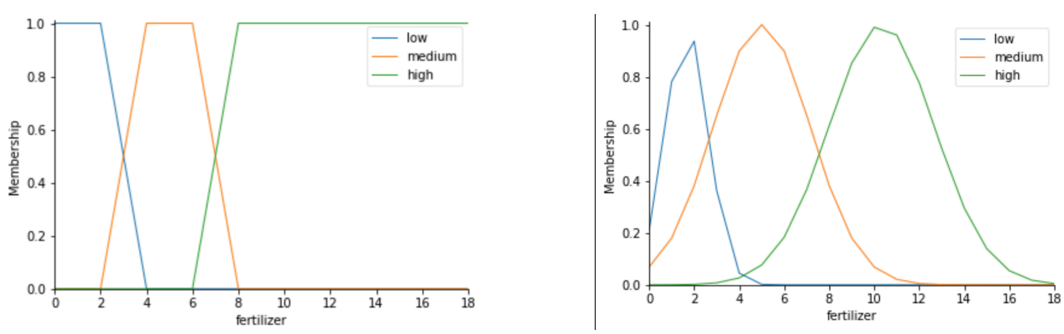


Fig. 3. Example of trapezoidal/Gaussian membership functions for one of the input variables (fertilizer).

4.5 Fuzzy rules

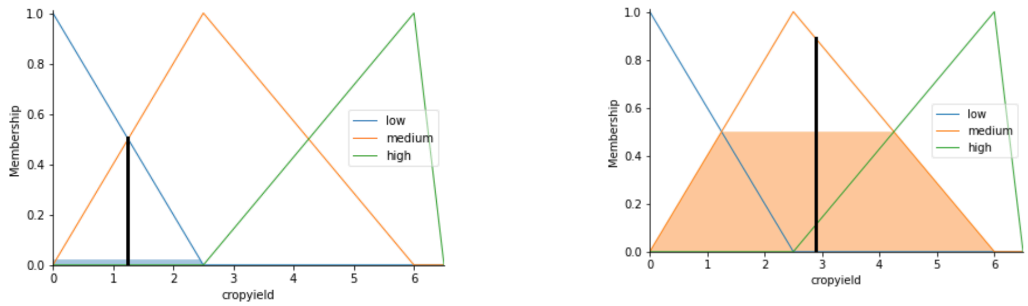
The fuzzy rules were generated using the responses of the experts. The rules are represented as IF-THEN. The input variables are antecedents and the crop yield is the consequent. According to the information in the antecedent, the consequent has a result. Table 3 shows two examples of the rules designed, which we use to show their behavior in the inference process using the different membership functions defined in the previous section. For example, Rule number 1 is: IF the red attack level is High AND the black attack level is High AND the crop stage is Vegetative AND the rainfall is High AND the fertilizer is Low AND the pheromone trap is Absent AND the boll-weevil killing tube is Absent THEN the crop yield is Low. Similarly, rule 2 has a different combination in the antecedent, and as result the crop yield is Medium. Finally, thirty-eight rules were designed for the system.

Table 3
Rule structure (Example of two of them).

Rule	If							Then
	Red attack level	Black attack level	Crop stage	Rainfall	Fertilizer	Pheromone trap	Boll-weevil killing tube	Crop yield
1	High	High	Vegetative	High	Low	Absent	Absent	Low
2	High	High	Flowering	Low	High			Medium

4.6 Defuzzification

The process of obtaining precise outputs (crop yield values) is known as defuzzification. In this subsection, we show two examples of the defuzzification process using the rules shown before, once the yield of the crop with said inputs is inferred. The center of gravity (CoG), also known as the centroid method, was adopted for defuzzification in this study (Cerrada et al., 2005). This approach uses the value of the output variables weighted by their membership functions. For example, Fig. 4.B shows the result of the defuzzification for one of the rules presented in Table 3, for a given set of inputs, with a result of prediction of Medium (2.88 ton/ha).



A. Crop yield of 1.23 ton/ha

B. Crop yield of 2.88 ton/ha

Fig. 4. Examples of defuzzification of the output variable (crop yield).

A standard fuzzy Mamdani system was developed using the scikit-fuzzy in Python, with a total of 38 if-then rules. Multiple sets of membership functions were examined during the FCS design process. The variables were set with triangular/trapezoidal/Gaussian membership functions according to Table 2.

5. Results

This section shows the experiments, the use of the surveys, and the scenarios to evaluate the FCS. The results of the FCS were then compared with the crop yield given by the experts.

5.1 Determination of the optimal membership functions for each scenario

The experts were surveyed about the specific values on the low, medium, and high scales for some variables. The specific value is a number representing the scale. The mean and standard deviation were calculated for each one. Table 4 shows the survey results for each fuzzy variable.

Another group of experts was used to validate the information, and they were surveyed about the specific values on the low, medium, and high scales of some variables.

The specific value is a number representing the scale. The mean and standard deviation were calculated for each one. Table 4 shows the survey results for each fuzzy variable.

Table 4

Survey Results: Experts' Assessments

Variable	Low		Medium		High	
	Mean	Std	Mean	Std	Mean	Std
Attack level of the red boll weevil	3	1.41	16.66	2.35	25	4.08
Attack level of the black boll weevil	2.66	1.69	15	4.08	25	7.07
Rainfall	2.66	0.47	6	0.81	12.33	1.69
Fertilizer	1.66	0.94	5	2.16	10.33	2.35
Crop yield	1.16	0.23	2.33	0.23	3.83	0.23

Abbreviation: Std= standard deviation

The mean and standard deviation were used to generate the Gaussian shape in the membership function. For each scenario, a triangular/trapezoidal or Gaussian membership function combination was used. These combinations were for five variables: red weevil attack level, black weevil attack level, precipitation, fertilizer, and crop yield. The other three variables (crop stage, pheromone trap, and weevil elimination tube) only have a triangular/trapezoidal membership function. In this sense, 32 possibilities (2^5) were generated for each scenario. In total were generated 288 combinations (9 scenarios x 32 possibilities). Table 5 shows the best performance using the best combination of the membership function for each scenario. In some cases, the trend was triangular/trapezoidal (e.g., scenarios 1 and 6), while in other cases it was Gaussian (e.g., scenario 9). In general, the FCS results are very close to expert opinion (see the last two columns).

Table 5

Evaluation of the best combination of membership functions.

Scenario	Membership Function								Expert System	Mean Expert
	Input							Output		
1	T	T	T	G	T	T	T	T	1,236	1,366
2	T	T	T	G	T	T	T	T	1,236	1,633
3	G	G	T	T	G	T	T	T	2,820	2,666
4	G	G	T	T	G	T	T	G	3,831	4
5	G	G	T	T	G	T	T	G	3,831	4
6	T	T	T	T	T	T	T	T	2,880	2,766
7	G	G	T	T	G	T	T	G	1,655	1,5
8	G	G	T	T	G	T	T	G	3,831	4
9	G	G	T	G	G	T	T	G	1,917	1,833

The input variables are red attack level, black attack level, crop stage, rainfall, fertilizer, pheromone trap, and boll-weevil killing tube. The output variable is crop yield. T = triangular / trapezoidal membership function; G = Gaussian membership function

Finally, the results of the validation experts were compared with those of the experts who defined the initial rules for the FCS. The comparison is explained in the following subsection.

5.2 Evaluation of the estimation capabilities of our FCS

A test was carried out with the best models (formats of the membership functions, see Table 5) for each scenario, to check whether the results (crop yield) matched the crop yield established by the experts. For this test, different values of the input variables describing the different scenarios were evaluated (more than 50.000 entries). The results (i.e., FCS outputs) were then compared to the crop yield provided by the experts. To do this, the mean of the responses for each scenario was calculated to arrive at a sole crop yield for each scenario, which was then compared to the crop yield for each scenario established by the experts that defined the initial rules for the FCS. Thus, the difference between the answers of the experts and the results of the FCS was evaluated.

Three measures were used to evaluate the performances of the estimations of our FCS (see Table 6). We used the *Coefficient of Determination* (R^2), which refers to the percentage of the variability in the dependent variable that can be explained by the independent variables. On the other hand, the *Mean Squared Error* (MSE) was used to evaluate the dissimilarity between the expert and predicted values. In addition, the *Mean Absolute Error* (MAE) was used to calculate the mean absolute difference between the experts and predicted values. R^2 ranges from 0 to 1. R2 values that approach 1 are desirable, it means that the FCS obtained a very good result. For the MSE and MAE, the values are always positive and a value closer to 0 or a lower value is better. In Table 6, we can see that the error metrics MSE and MAE are very small, which means very good results.

Table 6

Evaluation of estimates

R^2	MSE	MAE
0.9374	0.0661	0.2154

6. Discussion of results

This study defines a FCS with adaptive models (membership functions) for the diagnosis/prediction of cotton yield, validated by experts, with a very low error rate. In this study, we take into account the factors that influence crop yields; for example, the presence of insect pests related to the climate (Toscano-Miranda, Hoyos, et al., 2022; Trebilcok, 2020; Villarreal et al., 2005), the soil conditions related to the fertilizers (Mourhir et al., 2017; Papageorgiou et al., 2013), and other techniques to monitoring and control the insect pests (pheromone traps, boll weevil killing tube). Mendes et al. (2019) and Navinkumar et al. (2021) used AI techniques to predict the level of water needed to improve crop yields. Thus, we include the analysis of rainfall levels (low, medium, and high) according to the needs of cotton crops at a specific crop stage. As for fertilizers, we include rules that consider their efficient use according to soil analysis and crop stage. Moreover, according to Prabakaran et al. (2018), seasonal factors and pest incidence determine yield and their study can be used to maximize crop production. Thus, we use also these variables. Our results showed that the integration of all these variables is useful for the diagnosis/prediction of cotton crop yield.

Papageorgiou et al. (2013) used FCM and soil variables to predict crop yield. Thus, we found that fuzzy logic can be used to model the experts' knowledge and thus manages the uncertainty of the variables of the context. Uncertainties are unavoidable, such as, for example, climatic data, which we took into account in our proposal.

Unlike Papageorgiou et al. (2013), we use the data generated from expert opinion to fit the model (particularly, the membership function). Also, compared with variables, the approach proposed in the current study uses multi variables that cover different aspects of the crop (pests, stage of the crop, etc.). Other studies used techniques like single-layer neural networks (Maskey et al., 2019), extreme gradient boosting (Obsie et al., 2020), and Random Forest (Ranjan & Parida, 2019), but no one of them use adaptative models. Nor do they consider the data generated from expert opinion to fit the model.

In summary, in Table 7 we present four aspects of comparison. First, whether they proposed adaptive models for diagnosis/prediction. Second, whether they managed the uncertainty from the variables of the model or the context. Third, whether they considered the data that is generated from the opinion of the experts to adapt the model. Fourth, whether they simultaneously used variables of the climate, pests, fertilizers, and crops. As we discussed before, our approach is the first to combine these criteria to propose a diagnostic/predictive approach, which can be extended with the multi-agent systems paradigm (Aguilar et al., 2007; Terán et al., 2017) to give it more adaptability, extendibility and autonomy (Vizcarrondo et al., 2017) to the system.

Table 7

Comparison with other works.

Work	Adaptive model	Uncertainty	Data expert	CLFCT
Papageorgiou et al. (2013)	X	X	X	
Maskey et al. (2019)				
Ali et al. (2018)		X		
Lobell et al. (2013)				
Obsie et al. (2020)		X		
Ranjan & Parida (2019)				
This work	X	X	X	X

Abbreviation: CLFCT= Simultaneous use of Climatic, of pests, of Fertilizer, and of Crop variables.

Lastly, environmental variables such as soil organic matter, weed coverage percentage, and tillage system management, which were not explored in this study, might also explain the diagnosis/prediction of cotton yield observed in this study.

7. Conclusions

An FCS for the diagnosis/prediction of cotton crop yield is proposed in this work. The results showed that the system manages the experts' knowledge very well. This FCS has integrated with the results of previous work and different input variables were considered in the system (climates, pests and crops, among others). Results show that the fuzzy system can generate outputs that align with the assessments made by experts, and the metrics to evaluate the FCS show small errors. Our FCS can be used to help farmers select the best practices in crop management, to obtain the best yield in a specific context.

We used the opinion of experts in the crop, management, and marketing of cotton, which allowed defining good rules that achieve a diagnosis/prediction of the yield of the cotton crop, relating some characteristics of the cotton with the study area. In addition, the model can adapt the fuzzy rules and handle uncertainty. A limitation of the current work refers to the general use of the amount of fertilizer. We did not include each specific fertilizer class. This could be overcome by specifying the amount of each fertilizer, and adding them to the model.

We have plans to integrate this study and other knowledge models into a metacognitive architecture for smart agriculture that operates as an autonomous cognitive system. Furthermore, we will use the transfer learning paradigm to define models of boll-weevil behavior for various regions through a meta-learning task. In addition, we will extend the FCS with the skill of adaptability using data in real-time. This study did not include other environmental variables such as soil organic matter, weed coverage percentage, and tillage system management, which can be considered in future works. Finally, these models will be integrated into our cognitive architecture using the multi-agent systems approach, leveraging the modeling abilities and implementations within agent theory.

Acknowledgments

We thank the Colombian Agricultural Institute (ICA), Colombian Cotton Confederation (CONALGODON) and Colombian Institute of Hydrology, Meteorology and Environmental Studies (IDEAM).

Funding

This work was supported by the Colombian Science, Technology and Innovation Fund (FCTeI) of the General Royalty System (SGR); Universidad EAFIT; and Universidad de Córdoba.

Conflicts of Interest

The authors declare there are no conflicts of interest.

Ethical Approval

Not applicable

References

Aguilar, J., Cerrada, M., & Hidrobo, F. (2007). A Methodology to Specify Multiagent Systems. *Lecture Notes in Computer Science*, *4496*, 92–101.

Ali, M., Deo, R. C., Downs, N. J., & Maraseni, T. (2018). Cotton yield prediction with Markov Chain Monte Carlo-based simulation model integrated with genetic programing algorithm: A new hybrid copula-driven approach. *Agricultural and Forest Meteorology*, *263*, 428–448. https://doi.org/10.1016/j.agrformet.2018.09.002

Boadh, R., Chaudhary, K., Dahiya, M., Dogra, N., Rathee, S., Kumar, A., & Rajoria, Y. K. (2022). Analysis and investigation of fuzzy expert system for predicting the child anaemia. *Materials Today: Proceedings*, *56*, 231–236. https://doi.org/10.1016/j.matpr.2022.01.094

Brüngel, R., Rückert, J., Wohlleben, W., Babick, F., Ghanem, A., Gaillard, C., Mech, A., Rauscher, H., Hodoroaba, V. D., Weigel, S., & Friedrich, C. M. (2019). NanoDefiner e-tool: An implemented decision support framework for nanomaterial identification. *Materials*, *12*(19). https://doi.org/10.3390/ma12193247

Buchanan, B. G. (1983). Constructing an expert system. *Building Expert Systems*, 127–167.

Cerrada, M., Aguilar, J., Colina, E., & Titli, A. (2005). Dynamical membership functions: An approach for adaptive fuzzy modelling. *Fuzzy Sets and Systems*, *152*(3), 513–533. https://doi.org/10.1016/j.fss.2004.10.004

Ele, S., Umoh, E., & Adesola, W. (2014). An overview of the development principles, stages and building blocks of Expert System. *West African Journal of Industrial & Academic Research*, *11*(1), 44.

Grüger, J., Geyer, T., Bergmann, R., & Braun, S. A. (2022). CGK4PM: Towards a Methodology for the Systematic Generation of Clinical Guideline Process Models and the Utilization of Conformance Checking. *BioMedInformatics*, *2*(3), 359–374. https://doi.org/10.3390/biomedinformatics2030023

Hamedan, F., Orooji, A., Sanadgol, H., & Sheikhtaheri, A. (2020). Clinical decision support system to predict chronic kidney disease: A fuzzy expert system approach. *International Journal of Medical Informatics*, *138*. https://doi.org/10.1016/j.ijmedinf.2020.104134

Holzman, M. E., & Rivas, R. E. (2016). Early Maize Yield Forecasting from Remotely Sensed Temperature/Vegetation Index Measurements. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(1), 507–519. https://doi.org/10.1109/JSTARS.2015.2504262

Hwang, G. J., Sung, H. Y., Chang, S. C., & Huang, X. C. (2020). A fuzzy expert system-based adaptive learning approach to improving students' learning performances by considering affective and cognitive factors. *Computers and Education: Artificial Intelligence*, *1*. https://doi.org/10.1016/j.caeai.2020.100003

Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. In *Computers and Electronics in Agriculture* (Vol. 177). Elsevier B.V. https://doi.org/10.1016/j.compag.2020.105709

Lanzaro, G., & Andrade, M. (2022). A fuzzy expert system for setting Brazilian highway speed limits. *International Journal of Transportation Science and Technology*, *xxxx*. https://doi.org/10.1016/j.ijtst.2022.05.003

Lobell, D. B., Hammer, G. L., McLean, G., Messina, C., Roberts, M. J., & Schlenker, W. (2013). The critical role of extreme heat for maize production in the United States. *Nature Climate Change*, *3*(5), 497–501. https://doi.org/10.1038/nclimate1832

Mandal, M., Mohanty, B. K., & Dash, S. (2021). Understanding consumer preference through fuzzy-based recommendation system. *IIMB Management Review*, *33*(4), 287–298. https://doi.org/10.1016/j.iimb.2021.03.015

Mansour, A. I., & Abu-Naser, S. (2019). Expert System for the Diagnosis of Wheat Diseases. *International Journal of Academic Information Systems Research (IJAISR)*, *3*(4), 19–26. www.ijeais.org/ijaisr

Maskey, M. L., Pathak, T. B., & Dara, S. K. (2019). Weather based strawberry yield forecasts at field scale using statistical and machine learning models. *Atmosphere*, *10*(7). https://doi.org/10.3390/atmos10070378

Mendel, J. M. (1995). Fuzzy logic systems for engineering: a tutorial. *Proceedings of the IEEE*, *83*(3), 345–377. https://doi.org/10.1109/5.364485

Mendes, W. R., Araújo, F. M. U., Dutta, R., & Heeren, D. M. (2019). Fuzzy control system for variable rate irrigation using remote sensing. *Expert Systems with Applications*, *124*, 13–24. https://doi.org/10.1016/j.eswa.2019.01.043

Mourhir, A., Papageorgiou, E. I., Kokkinos, K., & Rachidi, T. (2017). Exploring precision farming scenarios using fuzzy cognitive maps. *Sustainability (Switzerland)*, *9*(7). https://doi.org/10.3390/su9071241

Navinkumar, T. M., Kumar, R. R., & Gokila, P. v. (2021). Application of artificial intelligence techniques in irrigation and crop health management for crop yield enhancement. *Materials Today: Proceedings*, *45*, 2248–2253. https://doi.org/10.1016/j.matpr.2020.10.227

Navinkumar, T. M., Ranjith Kumar, R., & Gokila, P. V. (2020). Application of artificial intelligence techniques in irrigation and crop health management for crop yield enhancement. *Materials Today: Proceedings*, *xxxx*. https://doi.org/10.1016/j.matpr.2020.10.227

Obsie, E. Y., Qu, H., & Drummond, F. (2020). Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. *Computers and Electronics in Agriculture*, *178*. https://doi.org/10.1016/j.compag.2020.105778

Papageorgiou, E. I., Aggelopoulou, K. D., Gemtos, T. A., & Nanos, G. D. (2013). Yield prediction in apples using Fuzzy Cognitive Map learning approach. *Computers and Electronics in Agriculture*, *91*, 19–29. https://doi.org/10.1016/j.compag.2012.11.008

Papageorgiou, E. I., Markinos, A. T., & Gemtos, T. A. (2011). Fuzzy cognitive map based approach for predicting yield in cotton crop production as a basis for decision support system in precision agriculture application. *Applied Soft Computing Journal*, *11*(4), 3643–3657. https://doi.org/10.1016/j.asoc.2011.01.036

Prabakaran, G., Vaithiyanathan, D., & Ganesan, M. (2018). Fuzzy decision support system for improving the crop productivity and efficient use of fertilizers. *Computers and Electronics in Agriculture*, *150*, 88–97. https://doi.org/10.1016/j.compag.2018.03.030

Ranjan, A. K., & Parida, B. R. (2019). Paddy acreage mapping and yield prediction using sentinel-based optical and SAR data in Sahibganj district, Jharkhand (India). *Spatial Information Research*, *27*(4), 399–410. https://doi.org/10.1007/s41324-019-00246-4

Salman, F. M., & Abu-Naser, S. (2019). Expert System for Castor Diseases and Diagnosis. *International Journal of Engineering and Information Systems (IJEAIS)*, *3*(3), 1–10.

Terán, J., Aguilar, J., & Cerrada, M. (2017). Integration in industrial automation based on multi-agent systems using cultural algorithms for optimizing the coordination mechanisms. *Computers in Industry*, *91*, 11–23. https://doi.org/10.1016/j.compind.2017.05.002

Toscano-Miranda, R., Hoyos, W., Caro, M., Aguilar, J., Trebilcok, A., & Toro, M. (2022). A Classification Model of Cotton Boll-Weevil Population. *2022 XVLIII Latin American Computer Conference (CLEI)*, 1–5.

Toscano-Miranda, R., Toro, M., Aguilar, J., Caro, M., Marulanda, A., & Trebilcok, A. (2022). Artificial-intelligence and sensing techniques for the management of insect pests and diseases in cotton: a systematic literature review. *The Journal of Agricultural Science*, *160*, 16–31. https://doi.org/10.1017/ S002185962200017X

Trebilcok, A. (2020). *El cultivo del algodonero en Córdoba*.

Villarreal, N., Lobatón, V., Galindo, J., & Mondragón, V. (2005). *Monitoreo del Picudo del Algodonero Anthonomus grandis Boheman en Colombia*.

Vizcarrondo, J., Aguilar, J., Exposito, E., & Subia, A. (2017). MAPE-K as a service-oriented Architecture. *IEEE Latin America Transactions*, *15*(6), 1163–1175.

Wang, A. X., Tran, C., Desai, N., Lobell, D., & Ermon, S. (2018, June 20). Deep transfer learning for crop yield prediction with remote sensing data. *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS 2018*. https://doi.org/10.1145/3209811.3212707

624

Appendix I. Surveys used for the development of the FCS.

The purpose of this study is to implement an FCS to diagnose/predict the cotton-crop yield in Córdoba, Colombia. This is useful to determine the actions to implement in the future to maximize production. Your viewpoints on various cotton crop scenarios should be expressed in this survey. Nine scenarios will be presented, related to the following variables:

- Attack level of red boll weevil (Low, Medium, and High)
- Attack level of black boll weevil (Low, Medium, and High)
- Crop stage (Vegetative, Flowering, Fruiting, Harvesting, Destruction of soca, and Closing)
- Rainfall (Low, Medium, and High)
- Fertilizer (Low, Medium, and High)
- Pheromone traps (Absent, Adequate)
- Boll-weevil killing tube (Absent, Adequate)

At the conclusion, it's essential to declare whether the suggested output should be classified as low, medium, or high, and designate an integer between 0 and 6 ton/ha as the yield. For example:

Scenario 1

VEGETATIVE STAGE Case 1 - low yield

- Attack level of red boll weevil: High
- Attack level of black boll weevil: High
- Crop stage: Vegetative
- Rainfall: High
- Fertilizer: Low
- Pheromone traps: Absent

- Boll-weevil killing tube: Absent

Note 1: The structure is recurrent in all 9 scenarios, but the variable values are modified for each individual case.

Ending comments

Are you interested in adding another scenario, not previously mentioned, what is essential for diagnosing the yield?

Appendix E

Precision farming using autonomous data analysis cycles for integrated cotton management

Precision farming using autonomous data analysis cycles for integrated cotton management

Raul Toscano-Miranda[a,c], Jose Aguilar[b,c,e], Manuel Caro[a], Anibal Trebilcok [d], Mauricio Toro[c]

[a]Department of Educational Informatics, Universidad de Córdoba, Colombia
[b]CEMISID, Universidad de Los Andes, Mérida, Venezuela
[c]GIDITIC, Universidad Eafit, Medellín, Colombia
[d]Department of Agronomic Engineering, Universidad de Córdoba, Colombia
[e]IMDEA Networks Institute, Leganés, Madrid, Spain

Abstract

Precision farming (PF) allows the efficient use of resources such as water, fertilizers, among others; as well, it helps to analyze the behavior of insect pests, in order to increase production and decrease the cost of crop management. This paper introduces an innovative approach to integrated cotton management, involving the implementation of an Autonomous Cycle of Data Analysis Tasks (ACODAT). The proposed autonomous cycle is composed by a classification task of the population of pests (boll weevil) (based on XGBoost), a diagnosis-prediction task of cotton yield (based on a fuzzy system), and a prescription task of strategies for the adequate management of the crop (based on genetic algorithms). The proposed system can evaluate several variables according to the conditions of the crop, and recommend the best strategy for getting increase the cotton yield. In particular, the classification task has an accuracy of 88%, the diagnosis/prediction task obtained a 98% of accuracy, and the genetic algorithm recommends the best strategy for the context analyzed. Focused on integrated cotton management, our system offers flexibility and adaptability, which facilitates the incorporation of new tasks.

Keywords: Precision Farm, Artificial Intelligence, Data Analysis. Autonomous Systems, Integrated Cotton Management.

1. Introduction

Precision Farm (PF) involves technologies for data collection, data analysis, and decision-making (Say et al., 2018). Data collection technologies are used to know the environment (e.g., sensors and images) (Cui et al., 2022). Data processing technologies use data models for the interpretation tasks (Kong et al., 2019). Decision-making technologies also use data models and actuators for planning tasks and changing the environment (Singh & Sharma, 2022).

On the other hand, there is a need to improve cotton production (Ghaffar et al., 2020) and PF technologies can help with this task (Coulibaly et al., 2022). According to Ghaffar et al. (2020), there is a great challenge in the management of cotton cultivation in which factors such as proper management of nutrients, pests, diseases, irrigation, etc. play an important role. In this paper, a PF approach based on autonomous data analysis cycles for integrated cotton management has been used.

There are some works for integrated management based on the PF. For example, Tribouillois et al. (2022) built integrated modeling of crop and water management to optimize irrigation. They used a combination of techniques to reduce water usage while also diversifying the types of crops grown in irrigated watersheds. Hajimirzajan et al., (2021) made a proposition for large-scale crop planning, which involves a comprehensive strategic framework that employs a decision support system to determine the sustainable use of water, as well as optimal crop selection, timing, and cultivation practices. Aggarwal and colleagues (2022) developed a system of geospatial analysis to preserve land fertility, optimize agricultural revenue, and minimize agricultural pollution and water consumption. The system allows land use planning with rotating crops. Wu et al. (2020) developed a model for integrated nutrient management that included four factors: chemical fertilizers, domestic livestock manure, large-scale livestock manure, and cultivated area. The authors found that there is a need to improve integrated nutrient management, expand livestock manure, and control cultivated areas of certain crops. However, to the best of our knowledge, there is no autonomous systems based on classification, diagnosis/prediction, and prescription tasks for the integrated management of crops. Therefore, in this paper, we focus on PF based on Autonomous Cycle of Data Analysis Tasks (ACODAT) for integrated cotton management.

We use ACODAT, which has two advantages. First, ACODAT allows automating the entire process, the phases of monitoring, analysis and decision making. Second, it does so from the process data. According to Sanchez et al. (2016), ACODAT makes use of diverse succeeding data analysis tasks interacting with one another to obtain the necessary knowledge to introduce process improvements. ACODAT has been utilized in various fields, including telecommunications, smart cities, industry 4.0, education, and medicine, as evidenced by different works (Aguilar et al., 2022; Morales et al., 2019; Sánchez et al., 2020). Morales et al. (2019) focused on the telecommunications sector, where they developed an ACODAT to manage quality of service in Internet of Things (IoT) platforms, utilizing classification and clustering tasks. It has been employed in smart cities for the purpose of regulating and monitoring heating, ventilation, and air conditioning systems (Aguilar et al., 2019). The efficiency of production processes in Industry 4.0 has been enhanced through the use of ACODAT. For instance, Sánchez et al. (2020) introduced an architecture that resolves the issues of heterogeneity and actor integration in manufacturing processes. The outcomes demonstrated that ACODAT facilitated interaction among actors such as things, data, people, and services, resulting in the definition of a self-optimization and self-configuration plan. In the educational domain, ACODAT has been implemented to identify learning styles in smart classrooms, demonstrating its usefulness. Aguilar et al. (2022) utilized ACODAT to study social network and web data, creating knowledge models about students to facilitate ongoing monitoring of their learning process. The findings underscored ACODAT's capacity to generate practical knowledge that can improve the learning experience, particularly in smart classrooms. Finally, the ACODAT approach has been used in the domain of medicine for clinical disease management (Hoyos et al., 2022).

This work aims to define an ACODAT for integrated cotton management. The contributions of this work are the following:

- The definition and implementation of an autonomous system based on ACODAT for integrated cotton management;
- A task of classifying the pest population (boll weevil) according to the level of attack on the cotton crop, based on the work (Toscano-Miranda et al., 2022);
- An adaptive model for the management of uncertainty based on a fuzzy system (FS) for the prediction/ diagnosis of cotton yield;
- The use simultaneous of information of fertilizers and crop stage, climatic variables, and level attack of pests, for the monitoring and control of pests, which improves the yield of prediction/diagnosis;
- A prescription task for the generation of strategies for the adequate management of the crop based on the previous tasks of the autonomous cycle.

The paper is structured in the following manner: Section 2 presents the works related to integrated crop management using computational techniques. Section 3 presents the theoretical framework of this paper. Section 4 outlines our integrated cotton management approach based on PF using ACODAT. Section 5 presents a case study to evaluate our proposal, and Section 6 describes the results. Finally, Section 7 shows the conclusions and highlights some of the future directions of this work.

2. Theoretical framework

This section presents concepts about PF for integrated production management, ACODAT and the Methodology for data analytics based on organizational characterization through a user-center design (MIDANO).

3.1. PF for integrated production management

PF aims to reduce costs, increase yield, using the right resources, being friendly to the environment. According to Gandonou (2005), PF is a set of technologies that help the farmer manage the agricultural process. In addition, it aids in production risk management (e.g., through the variable nutrient application), and reduces water consumption (e.g., through drip irrigation).

Say et al. (2018) grouped the PF technologies in three: a) Data collections technologies (e.g., soil sampling and mapping, yield monitoring and remote sensing); b) Data analysis technologies (e.g., geographic information system, economic analysis and modelling); c) and decision-making technologies (e.g., variable rate application, agricultural robots). Next, some examples:

a) Data collection technologies: These technologies detect insects and diseases in crops using field sensors, and remote sensors (Khattab et al., 2019; Toscano-Miranda et al., 2022a). In addition, using images for the same tasks (Alves et al., 2020; Caldeira et al., 2021).

b) Data analysis technologies: for predicting the behavior of insects (Hudgins et al., 2017; Toscano-Miranda et al., 2022) and crop yield (Maskey et al. (2019), expert systems for decision-making about diseases in crops (Mansour & Abu-Naser, 2019), etc.
c) Decision-making technologies: Automated crop management and treatment using PF (Vulpi et al., 2022), such as irrigation control using robots (Agostini et al., 2017), and spray control for insects or diseases (Song et al., 2017). For this, it is useful the unmanned vehicles in rural farm areas (Mammarella et al., 2021; Saha et al., 2022), geospatial analysis to decision support (Aggarwal et al., 2022), etc.

Our work integrates data collection, data analysis and decision-making technologies in an ACODAT.

3.2. ACODAT

Due to the significant increase in data generation, the development of new tools is essential to extract valuable knowledge. ACODAT is useful for this and is based on the autonomic computing paradigm. ACODAT involves a series of interconnected data analysis tasks that must be carried out in conjunction to achieve a desired objective within a given system or context. The tasks perform distinct roles within the cycle and interacts with one another (Sanchez et al., 2016; Terán et al., 2017; Vizcarrondo et al., 2017): they observe the process, analyze, and interpret events, and make appropriate decisions. The responsibility of observation tasks is to gather information and data about the environment or system, while analysis tasks interpret and diagnose the system using this data. Knowledge models are constructed to understand the cycle's behavior. Decision-making tasks, on the other hand, are responsible for improving the process by carrying out activities.

The autonomic computing paradigm is oriented to define autonomic characteristics to systems based on a smart control loop, known as MAPE+K (Sterritt et al., 2005; Vizcarrondo et al., 2017). The letter K corresponds to the knowledge models (e.g., classification, diagnostic, prediction, and prescription models) within the autonomous cycle. An ACODAT collects, filters, processes and analyzes data of the supervised problem. Also, it analyzes complex situations and predicts forthcoming situations. Additionally, it establishes the actions that must be carried out to reach the system objectives and defines mechanisms to execute the plan. Because of this, the autonomous cycle requires managing a large amount of information. The design of the autonomous cycle must include all these aspects to achieve the objectives that give solution to the problem.

3.3. MIDANO

MIDANO is a methodology that allows gaining a deeper understanding of the data, which relies on organizational characterization as a key component to develop ACODATs (Pacheco et al., 2014). Fig. 1 shows the three primary phases of MIDANO. The initial phase seeks to familiarize with the organization to define the goal of the data analysis system. The focus of this stage is to recognize and frame the solution to a problem, from the viewpoint of developing data analysis-based applications. Also, it defines the ACODAT for the solution of the problem. The responsibility of Phase 2 is to prepare and treat the data, following the ETL paradigm (Extraction, Transformation, Loading). Its primary goal is to produce high-quality data that can be used to build knowledge models and specify the multidimensional data model of ACODAT. In Phase 3, data analysis tasks are implemented to generate various knowledge models such as descriptive, predictive, classification, and prescriptive (Aguilar et al., 2021).

Problem characterization and ACODAT definition were accomplished during the first phase of our work using MIDANO. The second phase, which involved data preparation and treatment, was incorporated into the ACODAT to enable real-time processing of data, and increase the autonomy of the process. Additionally, this phase identified the required data sources for ACODAT development. Our work provides a detailed explanation of how each MIDANO phase was applied to cotton crop management.

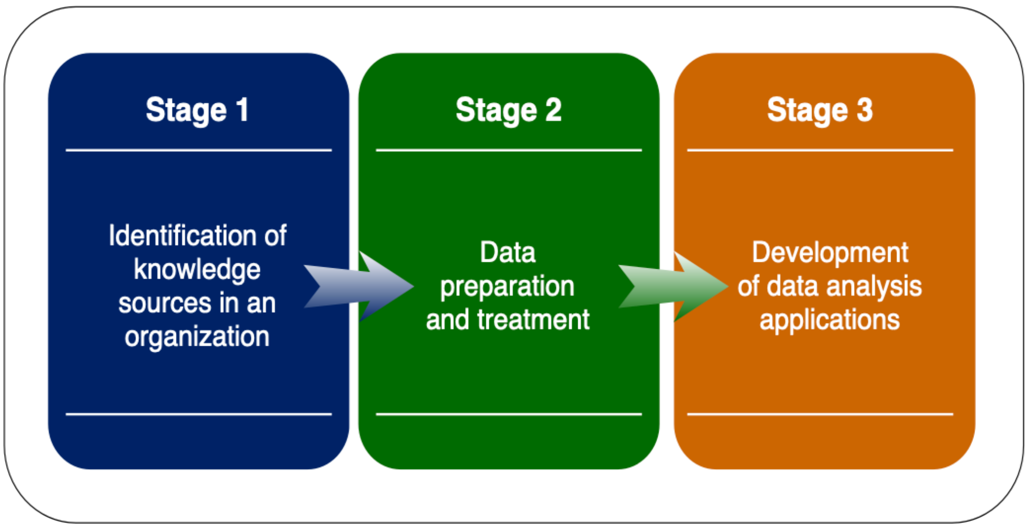


Fig. 1. MIDANO Methodology for Data Analysis from Organizational Characterization. Adapted from (Aguilar et al., 2021).

3. ACODAT for the integrated management of production processes

The process of creating an ACODAT for managing cotton crops is outlined in this section. Each task involved in ACODAT is detailed. This section describes in a general way the aspects necessary to implement this approach, which may be applicable to other crops and other pests with the variables selected by the expert. The specific variables are described in the Case study section in which the application and validation of the proposed approach is demonstrated.

4.1. Characterization of the management of cotton crop

The main objective of cotton cultivation is to obtain its valuable fiber (Trebilcok, 2020). There are several factors that influence production performance. For this reason, integrated crop management with the help of technologies seeks to improve yields with sustainable management and reduced environmental impacts (Abbas et al., 2020; Ghaffar et al., 2020). For example, fertilizer deficiencies slow plant growth and development, and consequently cotton yields are reduced (Ali et al., 2018; Ahmed, Ali, Danish, et al., 2020; Ahmed, Ali, Hussain, et al., 2020; Trebilcok, 2020). Cotton cultivation requires adequate nutrition, and its demand depends on various factors such as, stage of cultivation, genotype, environment (Trebilcok, 2020). Also, the water supply associated with the sowing date affects the yields and the general behavior of the plant. Regarding insect pests, it is recommended to control all types of cotton insect pests through integrated pest management techniques (Anees & Shad, 2020). Additionally, cotton production is more vulnerable to climate change. This produces a negative impact on cotton production (Ahmad et al., 2020).

Thus, there is a great challenge in the management of cotton cultivation in which factors such as proper management of nutrients, pests, diseases, irrigation, etc. play an important role (Ghaffar et al.; 2020). In this paper, we focus on a PF using ACODAT for integrated cotton management. Integrated cotton management includes several factors that, when used in a mixed manner, help to make better planning and decision-making. These factors are related to the right management of fertilizers, insect pests, diseases, irrigation, weed, etc. (Ghaffar et al., 2020). In this study, we included information related to fertilizer management, insect pests, irrigation, climate data and crop stages. These factors are related and were considered for planning and decision-making to assist the farmer in integrated cotton crop management.

4.2. MIDANO Application

We use the MIDANO methodology to design our ACODAT. Inside of our ACODAT are included data preparation and treatment data tasks.

4.2.1 ACODAT specification.

Fig. 2 shows our ACODAT approach for this purpose. ACODAT consists of a trilogy of steps that are linked together through a network of tasks to assist decision-making in cotton crop management. The first step, monitoring, is made up of two tasks: verifying and correcting data. The second step, analytics, involves classifying the population of boll weevils according to climate data and diagnosis/prediction of the cotton yield. The final step, decision-making, involves prescribing the best management strategy for cotton crops.

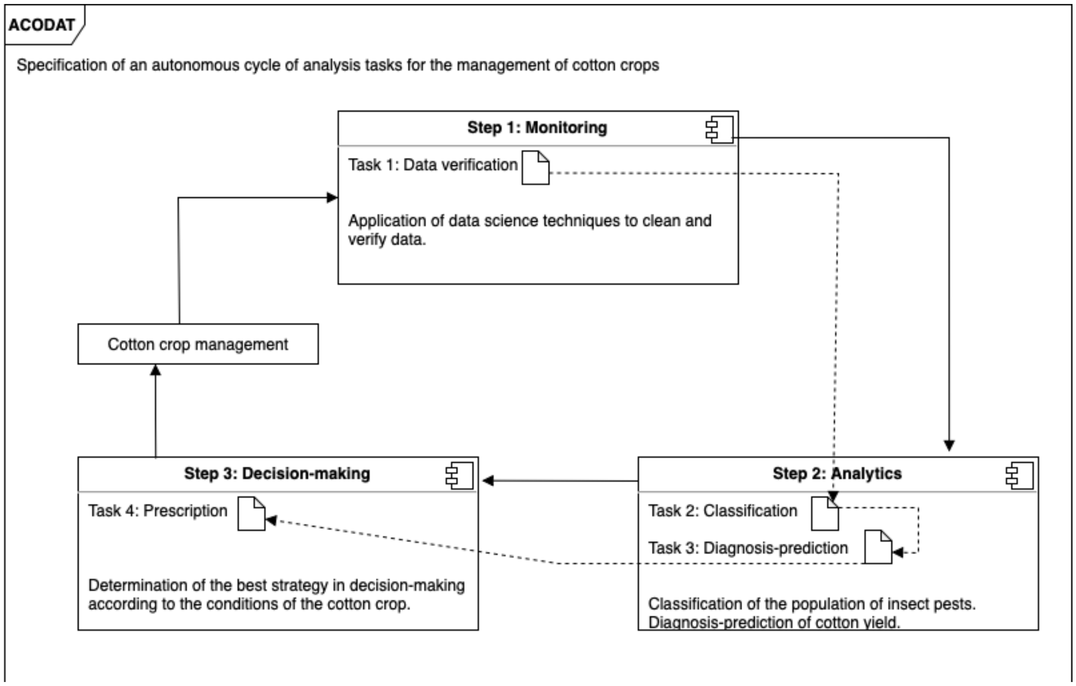


Fig. 2. ACODAT architecture for cotton crop management.

The techniques employed in the data analysis tasks belong to diverse domains of artificial intelligence (AI), including eXtreme Gradient Boosting (XGBoost), fuzzy systems, and genetic algorithm (GA). Therefore, the monitoring, analysis, and decision-making functionalities provided by ACODAT-based self-monitoring are as follows:

- Monitoring tasks: This process includes Task 1 to capture data, clean it and prepare it for the following tasks. In addition, relevant characteristics are extracted and preprocessed, and information about the behavior of insect pests is obtained. The selected features are used in the following steps.
- Analysis tasks: A set of tasks (tasks 2 and 3) to understand, interpret, and predict/diagnose what is happening in the cotton growing process.
- Decision-making tasks: This process includes Task 4 to prescribe the best strategy in the integrated management of cotton crops.

The complete cycle includes four integrated tasks, which communicate with each other and pass information from the first to the last. Each task used different techniques to achieve the objectives. Table 1 shows the interrelation between tasks, data sources and used techniques. The following subsections explain in detail each task in the autonomous cycle.

Table 1
Description of the ACODAT's tasks for integrated cotton management.

Functionalities	**Task name**	**Characteristics of the task**				
		Description	**Data source**	**Analytics type**	**Technique**	**Knowledge model**
Monitoring	Data verification	Verification of data (data processing) and	Datasets of monitoring of insects, Climate data	Description	Verification Oversampling / Statistical	Descriptive

		correction of errors			analysis	
Analysis	Classification	Classification of boll weevil population by climate data	Previous task	Classification/ Predictive	XGBoost	Predictive
	Diagnosis/prediction	Diagnosis/ prediction of cotton yield	Previous task, Dataset of cotton production	Diagnosis/ Predictive	Fuzzy logic	Diagnosis/ predictive
Decision-making	Prescription	Determination of the best strategy for the management of cotton crop	Previous task	Optimization	Genetic algorithm	Prescriptive

4.2.2. Monitoring tasks

Task 1 - Verification and data processing

Data Verification was designed as Task 1. This task includes a statistical analysis to evaluate the quality of the data. The modeling results are heavily influenced by the quality of the data. Thus, initially, our ACODAT identifies and fixes any potential data errors. Also, since missing data is common in this type of data, the dataset is purged of rows with missing data. Finally, an oversampling technique was used to balance the classes in the dataset.

In summary, the procedure for this task involves the subsequent actions: 1) extract the structured database about the insect pests, 2) verify if there are errors in the data, 3) delete rows with missing data, 4) Balance the dataset, where the number of samples from the minority class (the class with fewer examples) is increased by creating synthetic examples using the oversampling technique (Gosain & Sardana, 2017). Fig. 3 shows the steps in this task, while Table 1 lists its main features.

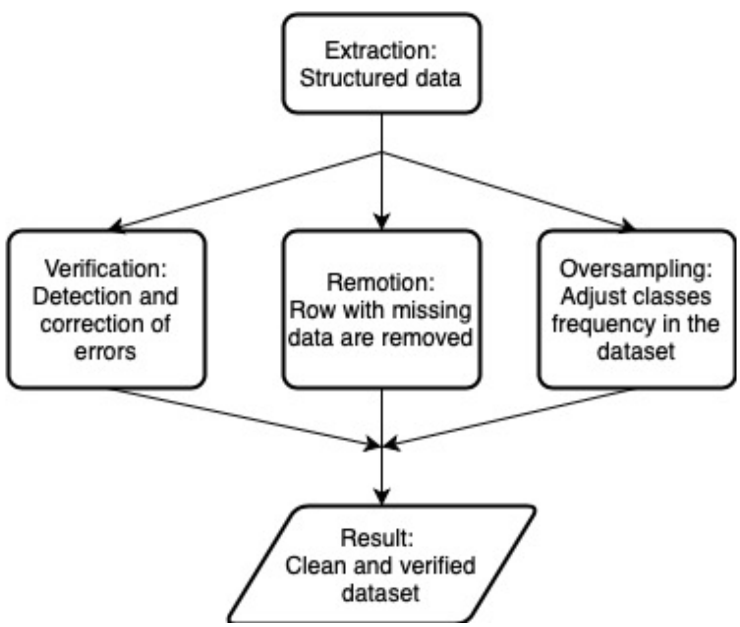


Fig. 3. Activities or sub-tasks related to task 1 (data verification and correction).

4.2.3. Analysis tasks

There are two analysis tasks. one of classification and another of diagnosis/prediction. The following is the description of each task:

Task 2 - Classification of the insect pest population:

The classification techniques are employed in this task to establish the population level of the insect pest. The classification is based on the climate variables of each city, considering their respective climate conditions. The XGBoost technique, which has demonstrated the highest accuracy in prior studies (Toscano-Miranda et al., 2022), was utilized. The main features of this task are detailed in Table 1.

Task 3 - Diagnosis/prediction of crop yield:

After the classification task, we develop the diagnosis/prediction task. This task uses a fuzzy model to diagnose/predict the cotton yield. We used expert opinions to build/define the fuzzy variables, their membership functions, and the fuzzy rules based on the work (Toscano-Miranda et al., 2023). The process involved in this task is illustrated in Fig. 4. The FS uses input variables that are passed to the fuzzification process. The inference engine uses the rule base and then the defuzzification process is performed to give a crisp output, which is the diagnosis/prediction of cotton yield.

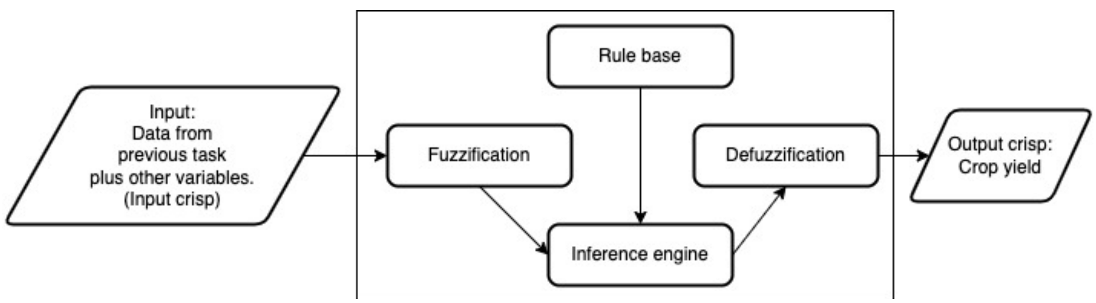


Fig. 4. Steps related to task 3 (Diagnosis/prediction).

Table 2 provides a summary of the input variables, including their descriptions, ranges, fuzzy sets and units of measure. Among these input variables, the attack level of red and black boll weevils is processed and categorized in Task 2 based on the count of boll weevils: Low (0 to 4), Medium (5 to 20), and High (greater than 20). The variable "Crop stage" indicates the phase of the crop in the year, providing insights into the ongoing activities during that phase.

The rainfall variable was acquired through sensor measurements conducted by the Institute of Hydrology, Meteorology, and Environmental Studies (IDEAM). In Task 1, boll weevil catches, and climatic data are consolidated into a unified dataset. The variable "Fertilizer" denotes the quantity of fertilizer utilized. Conventional pheromone traps are employed to capture red and black boll weevils, which are then monitored by ICA engineers every 15 days. The engineers manually record the boll weevil counts and subsequently enter this data into information systems databases.

Table 2

Summary of the input variables.

Input variable	Description	Fuzzy sets	Range	Units of measure
Attack level of the red boll weevil	Population of the red boll weevil in the cotton crop.	Low, Medium, and High	[0, 150]	Integer
Attack level of the black boll weevil	Population of the black boll weevil in the cotton crop.	Low, Medium, and High	[0, 200]	Integer
Crop stage	Crop stage in the year.	Vegetative, Flowering, Fruiting, Harvesting, Destruction of soca, and Closing	[0, 12]	Integer
Rainfall	Amount of rain that falls during the day.	Low, Medium, and High	[0, 17]	mm
Fertilizer	Amount of fertilizer used in the crop.	Low, Medium, and High	[0, 18]	Integer (Packages)
Pheromone traps	Number of traps used in the crop.	Absent, Adequate	[0, 1]	Integer
Boll-weevil killing tube	Number of tubes used in the crop.	Absent, Adequate	[0, 1]	Integer

This task uses fuzzy sets with membership functions Gaussian, triangular, and trapezoidal. The triangular function is used for the categorical variables, and the trapezoidal/Gaussian functions are used for the rests. Finally, 13 membership functions were defined for the input and output variables. Fig. 5 shows an example with a trapezoidal membership function.

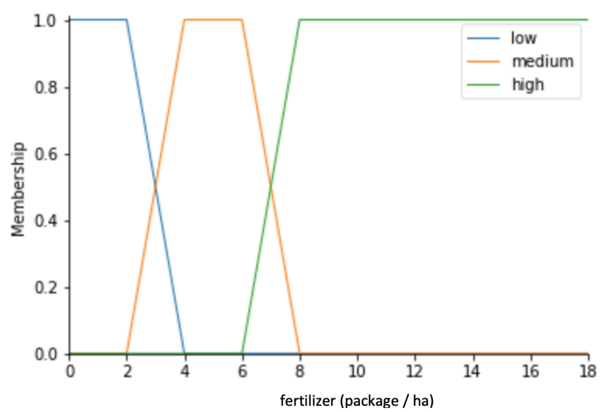


Fig. 5 Example of a trapezoidal membership function.

The experts' answers were also utilized to define the fuzzy rules. The rules are defined as IF-THEN. The antecedent are the input variables and the consequent is the crop yield. Table 3 presents two examples of the rules. For example, Rule number 1 is: IF the red attack level is High AND the black attack level is High AND the crop stage is Vegetative AND the rainfall is High AND the fertilizer is Low AND the pheromone trap is Absent AND the boll-weevil killing tube is Absent THEN the crop yield is Low. Rule 2 defines a different combination in the antecedent, and as a result the crop yield is Medium. Thirty-eight rules were defined for the system.

Table 3

Rule structure (Example of two of them).

Rule	If							Then
	Red attack level	Black attack level	Crop stage	Rainfall	Fertilizer	Pheromone trap	Boll-weevil killing tube	Crop yield
1	High	High	Vegetative	High	Low	Absent	Absent	Low
2	High	High	Flowering	Low	High			Medium

4.2.4. Decision-making tasks

Task 4 – Prescribing of strategies for crop management

For decision-making it was implemented a prescription task. The task was performed with a GA to determine the most efficient strategy for solving the problem. Experts' recommendations in crop management and marketing were identified as the starting point for this task. The crop management prescriptions in this task are based on expert opinion and compiled into a list. The GA optimizes the most efficient strategy for a specific scenario based on the previous task's findings. Table 1 outlines the task's characteristics. Thus, we use expert opinion to build a set of activities for each strategy. One strategy can be shaped by a combination of 13 activities. Specifically, our GA is based on the next procedure:

Algorithm 1: Training procedure of the Genetic Algorithm (GA)

Input: Data from the previous task, synthetic dataset
Output: Strategy recommended according to the best individual

1. Initialize the population
2. Evaluate the population
3. While (stopping condition not satisfied):
    (a) Select the population
    (b) Crossover the population
    (c) Mutate the population
    (d) Evaluate the population
    (e) Update the population
4. Return the best individual in the population

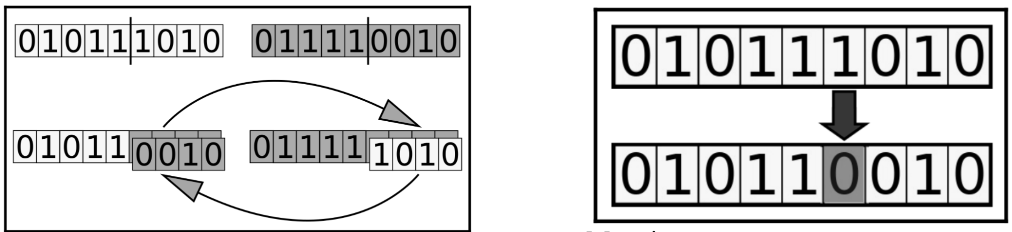
In this task, the result is the prescription of a strategy defined by a set of activities. Thus, an individual in a population is a strategy defined by a binary chain where each bit represents a gene (i.e., an activity). For example:

0	0	1	1	1	1	0	1	1	1	1	0	0

Fig. 6 An individual (prescription).

Therefore, the whole chromosome (individual) is a possible prescription. An activity should be used when a 1 appears, and not when it is 0. Thus, the population is a collection of candidate prescriptions for the context analyzed in cotton cultivation.

The following steps were taken to find the best strategy (see algorithm 1): 1) Initialize population: creating randomly a set of binary chromosomes that depict distinct solutions (possible prescriptions). 2) Evaluate: calculation of the fitness of each chromosome using the fitness function presented in the next paragraph, 3) Generating new individuals through genetic operators: In this stage, the chromosomes of the two fittest parents are selected, to which the crossover and mutation operators are applied (see Fig. 6). 4). Return the best individual in the population (i.e., the best strategy).



Crossover

Mutation

Fig. 6. Example of crossover and mutation processes in a GA.

The fitness function must evaluate possible solutions formulated on the analysis of the crop, in the before process (Task 2) and the context of the crop given as input in this task (Task 3). The fitness function returns values from 0 to 301. This output determines the best recommendation, being 0 as very adequate and the highest value as not adequate. The fitness function evaluates the context of the crop and the activities to be included in the recommendation. If the chromosome includes inappropriate activities, then those activities are penalized. For example, if the recommendation/prescription includes the activities "Conduct soil analysis" and "Apply the necessary amounts of fertilizer according to the soil analysis and the agronomist's recommendations" at the fruiting stage of the crop, this should be penalized. Experts in the management of cotton cultivation do not recommend this because at this stage of cultivation costs increase and it is not necessary. All equations were constructed based on the opinion of cotton crop management experts. The equations are:

$$C1 = A_9 * 100 + A_{10} * 100 + A_{11} * 100 + A_{12} * 100$$

Where C1 is the constraint 1, and 100 is a value that represents the penalization. The previous equation penalizes (i.e., it gives a higher value) in case the prescription includes the following activities in the flowering and fruiting stage: a) put pheromone traps (A_9), b) move the pheromone traps (A_{10}), c) put boll-weevils killing tube (A_{11}), and c) move the boll-weevils killing tube (A_{12}). This penalization is due to these activities are not recommended at these two stages and increasing the costs.

The next equation penalizes in case the prescription includes the following activities in the fruiting stage: soil analysis (A_7) and applying fertilizers (A_8). These activities are economically unfeasible at this stage and increase the costs.

$$C2 = A_7 * 100 + A_8 * 100$$

4. Case study

This section presents the experimental context and the instantiation of ACODAT in a case study for integrated cotton crop management using datasets from a region of Colombia. In this case study, we demonstrate how the ACODAT tasks are executed on particular datasets.

5.1. Context

We identified the data sources according to the MIDANO methodology (Aguilar et al., 2021). To identify the appropriate sources of knowledge, we engaged with experts in cotton cultivation for this case. For our purpose, we used the next data sources: 1) Net of monitoring of boll weevil (*Anthonomus grandis*) of the *Colombian Agricultural Institute* (ICA in Spanish), 2) Pheromone traps utilized in each cotton crop developed by the owners, 3) Registers of climate data from the *Institute of Hydrology, Meteorology and Environmental Studies* (IDEAM in Spanish), 4) Records of climate data of each cotton crop registered by the owners, 5) Registers of the management of each cotton crop registered by the owners, 6) Crop yields in the area according to the *Colombian Cotton Confederation* (CONALGODON in Spanish)

Our ACODAT was validated using cotton crops from different areas of Córdoba, Colombia, specifically, the cities comprising the Sinú Valley (High Sinú, Middle Sinú, and Low Sinú) (Trebilcok, 2020), located at ~8°55'33.6"N, 75°48'16.5"W. The data used for this implementation correspond to the Net of monitoring of the boll weevil operationalized by the ICA and records of climate data from the IDEAM. Data from the cities of Córdoba: Montería, Cereté, Lorica, and Ciénaga de Oro (from 2015 to 2021) were used for the experiments. We chose these regions because they are cultivated with cotton and have the records of the pheromone traps.

The dataset was composed by 13,585 samples of captures of boll weevils in pheromone traps. Of the 15 variables presented in the dataset, 11 were excluded from our study for not providing valuable information, for example: trap code, name of GPS, among others. Finally, six variables corresponding to the climatic data and related to the number of boll weevils were selected. Table 4 shows each of the variables, a brief description, and the task where it was used. The dataset containing climate variables was merged with the dataset comprising catches of boll weevils. The combination of these datasets was facilitated by using dates and cities as common identifiers. Given the absence of sensors within each crop, the location closest to the crop with climate sensors was used. In addition, the variables related to the stage of cultivation and fertilization were acquired from expert sources, ICA and CONALGODON.

Table 4
Variables and their descriptions, used in the cotton crop management.

Variable	Description	Units of measure	Task	Data source
Red boll weevils	The red boll weevils are the youngest. Quantity of captures of boll weevils.	Integer	1, 2	ICA

Black boll weevils	The black boll weevils are the ones that can procreate. Quantity of captures of boll weevils.	Integer	1, 2	ICA
Rainfall	Amount of rain that falls during the day.	mm	1, 2, 3, 4	IDEAM
Humidity	Hourly relative humidity (average of the day).	%	1, 2	IDEAM
Temperature	Maximum daily temperature, measured in degrees Celsius.	˚C	1, 2	IDEAM
City	City with records of boll-weevil attacks.		1, 2	ICA
Attack level of the red boll weevils	Low, medium, or high level as a result of the previous task.	Integer	2, 3, 4	Task 2
Attack level of the black boll weevils	Low, medium, or high level as a result of the previous task.	Integer	2, 3, 4	Task 2
Crop stage	Growth stage of cotton cultivation.	Integer	3, 4	ICA
Fertilizer	Amount of fertilizers used during growth stages.	Integer (Packages)	3, 4	CONALGODON, experts
Pheromone traps	The use of conventional pheromone traps in the cotton crop.	Integer	4	ICA
Boll-weevil killing tube	The use of boll-weevil killing tube in the cotton crop.	Integer	4	CONALGODON

5.2 Instantiation of ACODAT

5.2.1 Verification and data processing task

In the verification and data processing task, data about the boll-weevil captures were extracted. The dataset contained outlier data in the captures of the boll weevil, temperature, humidity, and rainfall. These samples were eliminated from the dataset. To identify these outliers, rigorous data validation techniques were applied. This involved performing data range checks and examining the distribution of values through visual inspection. Subsequently, identified outliers were carefully examined and removed from the data set to ensure accuracy and validity for subsequent analyses. For boll weevil catches, all data points determined to be significantly distant from most of the data were treated as outliers and excluded from the data set to maintain the reliability of the analysis. For example, values of 1200 catches (in 15 days) of boll weevil were considered outliers. Considering the regional climate conditions, specific thresholds were established for the variables of humidity, temperature, and rainfall. Humidity values above 68% and below 90% were considered appropriate for inclusion in the analysis, as they represented the relevant range of moisture levels in the region. Similarly, temperature values above 28 ºC and below 50 ºC were considered to encompass the typical temperature range of the area under investigation. In the case of rainfall, values ranging from 0 mm to less than 18 mm were selected as they represented the relevant spectrum of precipitation levels within the region. By defining these specific thresholds, we aimed to focus the analysis on the climatic conditions most pertinent to the study, ensuring the inclusion of meaningful data points.

In some periods of the year, the cities of Cereté, Lorica, and Montería experienced missing data in the climatic variables, including rainfall, temperature, and humidity. To ensure the integrity of the analysis and minimize potential biases caused by missing values, missing data processing was performed using a deletion method based on Mckinney (2010). Under this method, any individual in the dataset with missing data for any variable included in the analysis was excluded from further analysis. By removing individuals with missing data, we aimed to retain complete cases and maintain the reliability and validity of the analysis. This approach enabled a more robust examination of the available variables and their relationships, ensuring that only complete and reliable data were considered in our analysis. Additionally, we employed Synthetic Minority Oversampling Technique (SMOTE) (Gosain & Sardana, 2017) to even out the classes, given the low occurrence of categories of the boll weevil. Thus, for this first task, data were verified, corrected and balanced.

5.2.2 Classification task

The classification task used XGBoost as the classification technique to determine the population level of the boll weevil. In a previous work (Toscano-Miranda et al., 2022), this is the best technique for this task among Random Forrest, Support Vector Machine and Backpropagation Neural Networks. XGBoost gave an accuracy of 88%.

This task classified the attack level according to the boll-weevil population on the three labels of the dataset. The labels were low, medium, and high. The input for this task was a dataset that had been cleaned and validated in the previous task. The dataset was divided into 80% for training and validation, and 20% for testing. XGBoost was configured in different ways and 10-fold cross-validation was performed to determine the most optimal combination of hyperparameters. The hyperparameter settings for XGBoost are shown in Table 5.

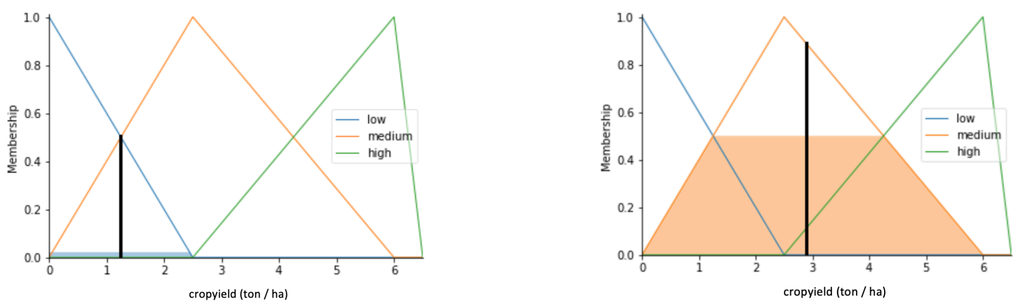
Table 5

Configuration of the hyperparameters of the XGBoost algorithm used to build the five models.

Algorithm	Best hyperparameters
XGBoost	Mtry = 1 Minimum n = 39 Tree depth = 13 Learn rate = 0.0459 Loss reduction = 0.0189 Sample size = 0.973

5.2.3 Crop yield diagnosis/prediction task

The analysis of cotton production involved the use of a soft computing method that incorporated the knowledge of experts. To perform the analysis, the system considered seven input variables, which were listed in Table 2. These variables included the level of attack from black and red boll weevils, the crop stage, the amount of rainfall, the amount of fertilizer applied, the use of pheromone traps, and the use of boll-weevil killing tubes. By considering these variables, the soft computing approach was able to generate insights into the factors that affect cotton production. This information could be used to improve the management practices of cotton farms and to increase the efficiency and profitability of cotton production. Four of these variables were reused of the previous task, including the classification of the boll-weevil population. As a result of this task, the diagnosis/prediction of cotton yield was obtained. To assess its robustness and adaptability, the system was subjected to tests using various agricultural scenarios. These scenarios are different combinations strategically made to reflect the practices of growers in the region described in the case study. In this sense, the scenarios allow us to evaluate the FS predictions and the strategies generated in the prescription task. The knowledge provided by experts was utilized to create the fuzzy rules (see Table 3). The FS was designed with a standard fuzzy Mamdani system that integrates 38 if-then rules. To determine the yield of the crop based on the inferred inputs and rules, the defuzzification process utilized the centroid method, which is also known as the center of gravity (CoG) (Cerrada et al., 2005). This process results in a single crisp value that represents the output of the fuzzy system. As an example, Fig. 7 illustrates the outcome of defuzzification for a given set of inputs using the rules presented in Table 3. The predicted result was medium, with a yield of 2.88 ton/ha (Fig. 7B). While Fig. 7A shows a low yield of 1.23 ton/ha.



A. Crop yield of 1.23 ton/ha

B. Crop yield of 2.88 ton/ha

Fig. 7. Examples of defuzzification of the output variable (crop yield).

To evaluate the performance of the FS, two measures were utilized as outlined in Table 6. Firstly, the *Coefficient of Determination* (R^2) was used to determine the proportion of the variance in the response variable that can be explained by independent variables. Secondly, the *Mean Squared Error* (MSE) was used to determine the difference between predicted and expert values. The R^2 score ranges between 0 and 1, and its high score represents a good result for the FS. On the other hand, the MSE should have a value lower or close to 0 for it to be considered good. The prediction was obtained by comparing the outputs of the FS with the predictions made by domain experts. The FS utilizes fuzzy reasoning, which activates fuzzy rules based on crisp input values such as fertilizer, crop stage, rainfall, pheromone trap data, black attack level, red attack level, and boll-weevil killing tube readings. These crisp values are first converted into fuzzy sets and then processed to generate both fuzzy and crisp output. The resulting output serves as the prediction, which is further utilized to calculate metrics such as R^2 and MSE. Additional details can be found in the Results section.

Table 6
Evaluation of estimates.

R^2	MSE
0.9374	0.0661

Findings indicate that the FS is capable of producing outputs that correspond with the evaluations of experts, thereby facilitating farmers in choosing the most effective cotton crop management practices to achieve optimal yield under specific circumstances.

5.2.4 Prescriptive task

The task of prescribing helps decision-making regarding the planning and management of cotton cultivation. The aim of this task was to establish the most effective strategy to manage cotton crops according to the context analyzed. It employs a series of prescriptions for the management of cotton crops according to experts in cotton cultivation, management, and marketing. Considering the results of the previous task (i.e., diagnosis/prediction of cotton yield), the GA optimizes the best strategy for a given scenario (it is an input). We use expert opinion to build a set of activities for each strategy. One strategy can be shaped by a combination of 13 activities. The activities considered in our case study are:

0. The cotton crop should be monitored more frequently.
1. The area where the boll weevil was found should be marked, according to the last inspection.
2. Cotton plant bolls that have fallen to the ground should be picked up daily.
3. The bolls affected by the boll weevil should be collected to prevent further feeding and propagation of the boll weevils.
4. The previously demarcated area should be fumigated.
5. Excessive rain must be evacuated using adequate drainage channels.
6. Implement an irrigation system.
7. Conduct soil analysis.
8. Apply the necessary amounts of fertilizer according to the soil analysis and the agronomist's recommendations.
9. Pheromone traps must be placed.
10. Move the pheromone traps frequently (use traps in the area recommended by the engineer and according to monitoring).
11. Place boll-weevil killing tube.
12. Frequently move the kill tubes (use tubes in the area recommended by the engineer and according to the monitoring).

According to Trebilcok (2020), Colombia employs various agricultural strategies to manage cotton crops from an entomological perspective. When the boll weevil infests the crop, specific activities are implemented accordingly. This involves distinguishing between two scenarios: when the boll weevil invades the crop in large numbers or when it appears in isolated foci. In the case of a mass invasion, where the weevils spread and establish themselves extensively throughout the lot, the most effective solution is to closely monitor the crop from day one until day 40, when fruiting begins. During this period, a comprehensive application of insecticide is conducted to eliminate the boll weevils before they have a chance to oviposit. As reproductive structures are not yet present, they cannot serve as a host for the boll weevil's eggs.

Alternatively, if the boll weevils appear in separate foci within the crop (one or multiple foci, depending on the crop area), the agronomist identifies and marks the locations during crop monitoring. By demarcating these foci, the agronomist signals to the farm administrator the presence of boll weevil infestation in those specific areas. Subsequently, the agronomist advises the farm manager to apply insecticide and collect the reproductive structures. Typically, one or two insecticide applications are carried out consecutively, with a time gap of one or two days between them. The objective is to suppress or minimize the boll weevils attempting to colonize the crop. During the colonization process, the boll weevils may have caused damage to the reproductive structures through feeding or oviposition. To address this, personnel (one, two, or three individuals depending on the size of the infestation focus) are assigned to collect the structures. The structures open their bracts within 48 hours and start falling to the ground. The staff can either pick them up from the ground or remove them from the plant before they naturally fall. Damaged structures exhibiting symptomatic open bracts can be easily detached from the plant. This unique strategy ensures a nearly absolute reduction in boll weevil colonization. Staff pick them up from the ground or take them from the plant without waiting for them to fall to the ground. Damaged structures are known for their open square symptomatology and can therefore be torn from the plant. This is a very special strategy to make an almost absolute reduction in the colonization of the boll weevil.

In this sense, this task used eight variables as input. The level attack of red and black boll weevil was the result of the classification in Task 2, rainfall was processed from the classification in Task 2, crop yield was result of the diagnosis/prediction in Task 3, and the variables crop stage, pheromone traps, boll-weevil killing tube and fertilizer is processed in this task. In this task, the result is the prescription of a set of activities (they form a strategy). The GA uses the fitness function that minimizes the cost defined in the previous section. When the farmer applies the best/optimal strategy increases the yield of cotton. Thus, the fitness function minimizes costs in the proper use of the irrigation system, the use of pheromone traps, the use of boll-weevil killing tubes and the use of fertilizer.

The crossover probabilities were set to 0.9 and mutation to 0.1. Previous research has indicated that the probability values used here have been successful in producing optimal results on comparable problems (Eiben et al., 1999; Hassanat et al., 2019). The crossover operator divides two chosen parents' chromosomes at a random point, resulting in two initial and two final gene subsets. These final subsets are then exchanged, generating two new chromosomes. The mutation operator randomly modifies each offspring's genes on a chromosome level.

5. Results

6.1 Results of Task 1 - Verification and data processing

The boll weevil population was categorized based on data ranges, with the low, medium, and high groups being defined as 0 to 4, 5 to 20, and greater than 20 respectively. These intervals were determined by the ICA. The distribution of the attack level classes was uneven and required SMOTE oversampling (Gosain & Sardana, 2017), as well as data standardization. Nonetheless, SMOTE was not used with Ciénaga de Oro and Montería due to their limited number of high-class red boll weevils.

Table 7
Distribution of classes for boll weevil in the Córdoba region.

Class	Red boll weevils	Black boll weevils
Low (0 to 4)	6,456	4,701
Medium (5 to 20)	304	1,244
High (> 20)	83	808

6.2 Results of Task 2 - Classification of the boll-weevil population

Extreme Gradient Boosting (XGBoost) was selected because (1) it is the technique that has shown good performance in this context (Toscano-Miranda, 2022), and (2) according to the literature review (Toscano-Miranda, Toro, et al., 2022a), it is the most frequently technique among structured data classification technique. The model for classification was evaluated independently for black and red boll weevils. Three weather features - temperature, humidity and rainfall - were tested in the experiments.

XGBoost achieved an 82% accuracy rate in detecting red boll weevils, the highest among the models tested, but its ability to predict black boll weevils was constantly below 60% (see Table 8).

Table 8
Outcomes of the classification model of black and red boll weevils using rainfall, humidity, and temperature.

Boll weevils	*Accuracy*		*F1-Score*	
	Training	*Test*	*Training*	*Test*
Reds	0.82	0.82	0.82	0.82
Blacks	0.60	0.60	0.59	0.59

Additionally, experiments were performed that solely used rainfall to encompass the entire Córdoba department as well as its cities. The results indicated that the accuracy of the model was lower when using just one attribute rather than all three (see Tables 9 and 10).

Table 9
Results of the model of classification using the XGBoost algorithm and rainfall.

City	Red boll weevil				Black boll weevil			
	Accuracy		*F1-Score*		*Accuracy*		*F1-Score*	
	Training	*Test*	*Training*	*Test*	*Training*	*Test*	*Training*	*Test*
Córdoba	0.75	0.74	0.75	0.73	0.57	0.56	0.57	0.55
Cereté	0.67	0.65	0.67	0.65	0.52	0.49	0.52	0.49
Lorica	0.78	0.73	0.78	0.73	0.60	0.56	0.60	0.56
Ciénaga	FoO	FoO	FoO	FoO	0.69	0.64	0.69	0.64
Monteria	FoO	FoO	FoO	FoO	0.82	0.70	0.82	0.70

Abbreviation: FoO= Fail on oversample.

Feature selection using Random Forest gave as result that temperature was the main feature. Then, new trials were executed solely using it (see Table 10). The performance of the red boll weevils' algorithm was improved in general for Córdoba through feature selection, resulting in an increase in Accuracy and F1 scores on the training dataset, from 82% (three features) to 83% (temperature only). However, not all cities obtained good results. For this reason, new tests were carried out including the three features as described later in this section.

Table 10

Outcomes of the classification model of red and black boll weevils using temperature.

Boll weevils	*Accuracy*		*F1-Score*	
	Training	*Test*	*Training*	*Test*
Reds	0.83	0.79	0.83	0.79
Blacks	0.62	0.59	0.62	0.59

XGBoost, was applied to the data, using three features for each city, as detailed in Table 11. The results showed that Lorica, Cereté and Ciénaga de Oro had better accuracy with black boll weevils, while Lorica and Cereté had better accuracy with red boll weevils. However, when a model was trained using data from all locations in Córdoba, including the samples from Ciénaga de Oro, Cereté, and Lorica, the accuracy for both black and red boll weevils was found to be lower. This decrease in accuracy could potentially be attributed to the unsuccessful oversampling technique applied in Ciénaga de Oro with data of red boll weevils, where the number of captures was predominantly in the low class. This skewed data distribution may have resulted in a biased model. That is, in Ciénaga de Oro, there were few captures of boll weevils; therefore, the categorization in the Medium and High classes was not sufficient to perform oversampling effectively. Specifically, the Low class had 946 records, the Medium class had 36 records, and the High class had only 3 records. This limited representation of the Medium and High classes in Ciénaga de Oro significantly impacted the oversampling process, as the dataset lacked a robust distribution across all classes. Additionally, it should be noted that Montería, another city included in the study, had limited available features, with only maximum temperature and rainfall being recorded.

Table 11

Classification model with XGBoost using temperature, rainfall, and humidity. The experiment included four cities of Córdoba.

Model	Red boll weevil	Black boll weevil

	Accuracy		*F1-Score*		*Accuracy*		*F1-Score*	
	Train	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*
*Córdoba	0.82	0.82	0.82	0.82	0.60	0.60	0.59	0.59
Cereté	0.78	0.77	0.78	0.77	0.57	0.52	0.57	0.52
Lorica	0.88	0.88	0.88	0.88	0.66	0.58	0.66	0.58
Ciénaga de Oro	FoO	FoO	FoO	FoO	0.71	0.69	0.71	0.69
Montería	NH	NH	NH	NH	NH	NH	NH	NH

*Córdoba (included Cereté, Lorica, and Ciénaga de Oro). Abbreviation: NH = No humidity. FoO = Fail on oversample.

The experiment was carried out after considering the results of previous experiments, and the models with the highest accuracy, Montería for black boll weevils and Lorica for red boll weevils, were used in this test. The purpose of the experiment was to assess whether the best model for one city could result in better classification results for other cities. The models were tested across all other cities to estimate their accuracy levels, and unfortunately, the results showed a decrease in accuracy levels. Specifically, Cereté's accuracy levels dropped from 52% to 29% for black boll weevils and from 77% to 48% for red boll weevils. In other words, the models that worked best for Lorica and Montería did not perform as well in Cereté.

6.3 Results of Task 3 - Diagnosis/prediction of crop yield

This section describes the experiments and scenarios carried out to evaluate the FS. After the FS generated outputs, the results were compared to the crop yield information provided by experts.

Determination of the optimal membership functions for each scenario

Experts were asked to provide specific values for low, medium, and high scales of certain variables through a survey. Each value corresponds to a number on the scale, and the mean and standard deviation were calculated for each value (Table 12).

Table 12

Survey Results: Experts' Assessments.

Variable	Low		Medium		High	
	Mean	Std	Mean	Std	Mean	Std
Attack level of the red boll weevil	3	1.41	16.66	2.35	25	4.08
Attack level of the black boll weevil	2.66	1.69	15	4.08	25	7.07
Rainfall	2.66	0.47	6	0.81	12.33	1.69
Fertilizer	1.66	0.94	5	2.16	10.33	2.35
Crop yield	1.16	0.23	2.33	0.23	3.83	0.23

Abbreviation: Std= standard deviation

The study used different membership functions for variables such as rainfall, black boll weevil attack level, red boll weevil attack level, fertilizer, and crop yield. These functions included triangular/trapezoidal or Gaussian combinations, while other variables like crop stage, pheromone trap, and boll-weevil elimination tube only had triangular/trapezoidal membership functions. Overall, 32 possibilities were generated for each scenario, leading to a total of 288 combinations (9 scenarios x 32 possibilities). The mean and standard deviation were used to create the Gaussian shape in the membership function. The best combination of membership functions was chosen for each scenario, with Table 13 showing the best performance. In some cases, triangular/trapezoidal trends were observed (e.g., scenarios 1 and 6), while in others, Gaussian trends were observed (e.g., scenario 9). The FS results were generally consistent with expert opinion, as shown in the last two columns.

Table 13
Evaluation of the best combination of membership functions.

Scenario	Membership Function								FS	Mean Expert
	Input							Output		
1	T	T	T	G	T	T	T	T	1.236	1.366
2	T	T	T	G	T	T	T	T	1.236	1.633
3	G	G	T	T	G	T	T	T	2.820	2.666
4	G	G	T	T	G	T	T	G	3.831	4
5	G	G	T	T	G	T	T	G	3.831	4
6	T	T	T	T	T	T	T	T	2.880	2.766
7	G	G	T	T	G	T	T	G	1.655	1.5
8	G	G	T	T	G	T	T	G	3.831	4
9	G	G	T	G	G	T	T	G	1.917	1.833

The input variables are fertilizer, crop stage, rainfall, pheromone trap, black attack level, red attack level, and boll-weevil killing tube. The output variable is crop yield. T = triangular / trapezoidal membership function; G = Gaussian membership function

Evaluation of the estimation capabilities of our FS

To further elaborate, the purpose of the test was to evaluate the accuracy and effectiveness of the fuzzy system in predicting crop yield values across various scenarios. The best models, which included formats of the membership functions, were chosen for each scenario, and were used in the test. The test involved considering different values of the input variables that described each scenario, which amounted to more than 50,000 entries. The fuzzy system generated results (FS outputs) for each input value, which were then compared to the crop yield established by experts. In order to compare the results with the crop yield established by the experts, the responses from each scenario were averaged to obtain a single crop yield value per scenario. This average value was then compared to the crop yield for each scenario defined by the experts. By comparing the crop yield values predicted by the FS with those established by the experts, the difference between the two was evaluated. Overall, the test was carried out to determine if the FS was consistent in predicting crop yield values that were comparable to those established by experts. This information could then be used to improve the accuracy of crop yield predictions and ultimately assist in decision-making related to crop production.

To assess the effectiveness of our FS, we employed a duo of measures for evaluating its performance. First, we used R^2 (0.9374), and second, the MSE (0.0661). We can see that the results are very good.

6.4 Results of the Task 4 – Prescribing with strategies for crop management

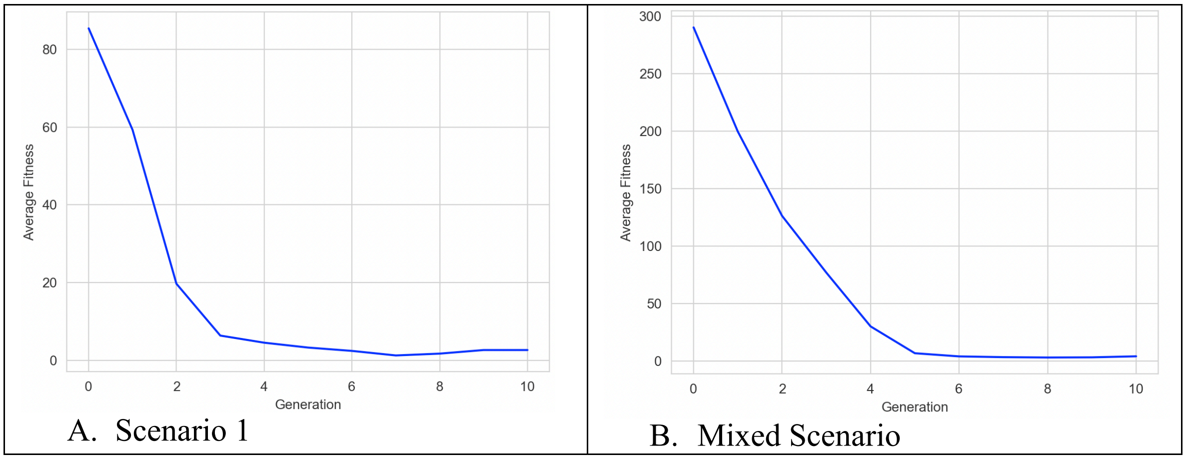
This section shows the results of ACODAT for integrated cotton crop management. For this, real data from cities in the region of Córdoba-Colombia were used. We used different scenarios to validate the experiments. Some scenarios with specific characteristics and others mixed scenarios from the former. In this paper, we present both scenarios to show the application of the autonomous cycle until reaching prescription. Table 14 summarizes the scenarios described in this section. Scenario 1 had a medium level crop yield diagnosis/prediction and Scenario 2 had a low level. According to these levels, a prescription is needed to improve crop yield.

Table 14
Summary of the scenarios.

Scenario	A	B	Crop stage	Rainfall	Fertilizer	C	D	Crop yield
1	Low	Low	Vegetative	High	Medium	Adequate	Adequate	Medium
2	Medium	Medium	Fruiting	Low	NA	NA	NA	Low

Abbreviations: A = Attack level of red boll weevils, B = Attack level of black boll weevils, C = Pheromone trap, D = Boll-weevil killing tube, NA = The farmer did not use this item.

Fig. 8 shows the results using the GA for the scenarios in Table 14. In some scenarios, convergence to optimal prescribing is faster than in others). For example, Fig. 8a. shows a convergence in seven generations, compared to Fig. 8b that shows a convergence in eight generations. The scenarios were tested several times, Fig. 8 shows the average of the generation in which the fitness function reaches the optimal strategy. Fig. 8a begins with values up to 80 and found the best prescription in generation number 7. Fig. 8b begins with values up to 250 and found the best prescription in the generation number 8. The value in the y-axis indicates the values average of the fitness function. The values higher indicated that the individual was penalized. The values closer to cero are appropriated because is an optimization problem of minimizing the costs.



A. Scenario 1 B. Mixed Scenario

Fig. 8. Graph of minimization of the fitness function (with 10 generations).

6.5 General Discussion of Prescriptive Analysis

In the diagnostic/prescribing task, only cases where the crop yield is low, or medium are invoked. Therefore, Table 15 shows high-performance scenarios. All prescription results were 100% correct with all activities included in the strategy and in this sense, the error rate was 0. The generation number needed to reach the prescription was different from scenario to scenario.

Table 15
Example scenarios and their results.

Scenario	The best prescription	N generations	Error	Crop yield	Type
1	100%	7	0	Medium	Mixing
2	100%	7	0	Low	Isolated
3	100%	7	0	Medium	Isolated
4	100%	8	0	Medium	Isolated
5	100%	8	0	Low	Isolated
6	100%	8	0	Low	Isolated
7	100%	7	0	Low	Isolated
8	100%	8	0	Low	Mixing
9	100%	7	0	Low	Mixing

Now, we took two examples to show the results of the prescription in real conditions. The analysis of scenario 1 indicates a medium level of cotton crop yield and scenario 2 a low level.

Scenario 1:

The characteristics of this scenario are: first, it begins with the classification task of the boll-weevil population: The classification task received as input values of temperature, humidity and rainfall of the cultivated area and classified the attack level of the boll weevil as: *low attack level of red boll weevils, low attack level of black boll weevils*.

Second, the diagnosis/prediction task of crop yield received as input values the results of the previous task: a low attack level of red boll weevils and a low attack level of black boll weevils. Additionally, the crop was in the vegetative stage, the rainfall was high (17 mm). Also, at this stage, the farmer used 5 packages of fertilizer (medium), used pheromone traps, and a boll-weevil killing tube. As a result of this task, the diagnosis/prediction of the crop yield was medium (2.88 ton/ha), see Fig. 9.

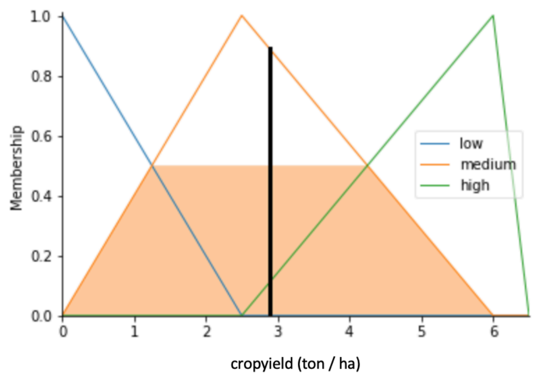


Fig. 9. Defuzzification of the output variable (crop yield with 2.88 tons/ha).

Third, the prescription task for management crop received as input values the results of the previous task: a) a low attack level of red boll weevils, b) a low attack level of black boll weevils, c) a stage of the crop in vegetative, d) a high rainfall (17 mm). Also, at this stage, the farmer e) used five packages of fertilizer (medium), f) used pheromone traps, g) used a boll-weevil killing tube, and mainly, and h) the crop yield was diagnosed as medium. Therefore, according to the medium crop yield, ACODAT should generate a prescription with the best strategy. ACODAT then generates the best strategy as a recommendation to increase the cotton yield to achieve a high level. In this sense, the final prescription is the following chromosome:

1	1	1	1	1	1	0	1	1	0	0	1	1

Fig. 10 Best individual for the first scenario.

Each gene corresponds to an activity. If there is a 0 the activity is not recommended and if there is a 1 the activity is recommended. Table 16 shows the detail of each gene on the previous chromosome.

Table 16
Activity configurations of the best recommendation.

Position on chromosome	Gene	Activity
1	1	The cotton crop should be monitored more frequently.
2	1	The area where the boll weevils were found should be marked, according to the last inspection.

3	1	The cotton buds (squares) of the cotton plants that have fallen to the ground must be collected daily.
4	1	The bolls of the cotton plants that have been affected by the boll weevil must be collected to prevent the boll weevil from feeding and spreading.
5	1	The previously demarcated area should be fumigated.
6	1	Excessive rain must be evacuated using adequate drainage channels.
7	0	The irrigation system should NOT be implemented.
8	1	Soil analysis should be performed.
9	1	The necessary amounts of fertilizer should be applied according to soil analysis and agronomist recommendations.
10	0	Pheromone traps must NOT be placed.
11	0	DO NOT move the pheromone traps frequently.
12	1	Boll-weevil killing tubes should be installed.
13	1	Boll-weevil killing tubes should be moved frequently.

This result is correct because the prescription found an optimal strategy, minimizing costs and using activities that improve crop yield. The prescription points out that the farmer a) should monitor the cotton crop more, b) should mark the area where the boll weevils were found, according to the last inspection, c) should collect daily the cotton buds (squares) of the cotton plants that have fallen to the ground, d) should collect the bolls from cotton plants that have been affected by the boll weevil and thus prevent further feeding and spread of the boll weevils, e) should fumigate the previously demarcated area, f) should evacuate the excessive rain with draining channels, g) should perform a soil analysis, h) should apply the right amount of fertilizer according to soil analysis and agronomist recommendations, i) should install boll-weevil killing tubes, and j) should move frequently the boll-weevil killing tubes. Activities a), b), c), d), and e) should be performed because monitoring and control activities are needed to quickly eradicate the boll weevil. Activity j) included boll-weevil killing tubes and exclude pheromone traps (i.e., the farmer should not use these activities simultaneously because it increases the cost and it is not necessary). In brief, the prescriptive model gives an accurate suggestion regarding the expert opinion on cotton cultivation.

Scenario 2:

The characteristics of this scenario are: first, it begins with the classification task of the boll-weevil population: The classification task received as input values of temperature, humidity and rainfall of the cultivated area and classified the attack level of the boll weevil as*: medium attack level of red boll weevils, medium attack level of black boll weevils*.

Second, the diagnosis/prediction task of crop yield received as input values the results of the previous task: a medium attack level of red boll weevils and a medium attack level of black boll weevils. Additionally, the crop was in the fruiting stage, the rainfall was low (2 mm). Also, at this stage, the farmer did not use fertilizer, pheromone traps, and a boll-weevil killing tube. As a result of this task, the diagnosis/prediction of the crop yield was low (1.23 ton/ha), see Fig. 11.

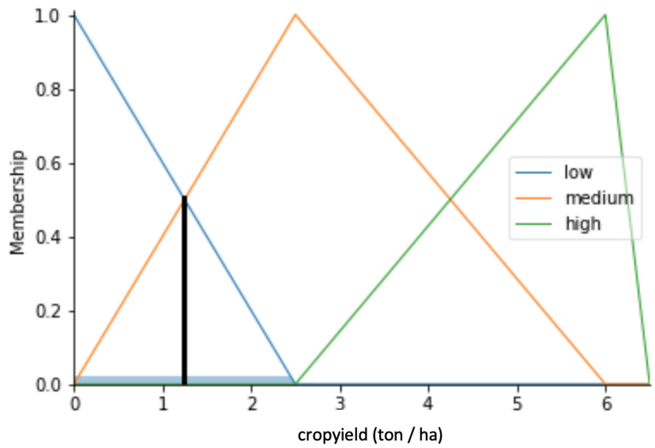


Fig. 11. Defuzzification of the output variable (crop yield with 1.23 tons/ha).

Third, the prescription task for management crop received as input values the results of previous task: a) a medium attack level of red boll weevils, b) a medium attack level of black boll weevils, c) a stage of the crop in fruiting, d) a low rainfall (2 mm), e) at this stage the farmer did not use fertilizer, f) nor pheromone traps, g) no tube kills weevils, and mainly, h) the crop yield was diagnosed as low. Therefore, and according to the low crop yield, ACODAT then generates the best strategy as a recommendation to increase the cotton yield to achieve a high level. In this sense, the final prescription is the following:

0	0	1	1	1	0	1	0	0	0	0	0	0

Fig. 12 Best individual for the second scenario.

Table 17 shows the detail of each gene on the previous chromosome.

Table 17
Activity configurations of the best recommendation.

Position on chromosome	Gene	Activity
1	0	The cotton crop should NOT be monitored more frequently.
2	0	The area where the boll weevils were found should NOT be marked, according to the last inspection.
3	1	The cotton buds (squares) of the cotton plants that have fallen to the ground must be collected daily.
4	1	The bolls affected by the boll weevil should be collected to prevent further feeding and propagation of the boll weevils.
5	1	The previously demarcated area should be fumigated.
6	0	Excessive rain must NOT be evacuated using adequate drainage channels.
7	1	An irrigation system should be implemented.
8	0	Soil analysis should NOT be performed.
9	0	Fertilizer should NOT be applied.
10	0	Pheromone traps should NOT be placed.
11	0	Pheromone traps should NOT be moved frequently.
12	0	Boll-weevil killing tubes should NOT be placed.
13	0	Boll-weevil killing tubes should NOT be moved frequently.

This result is correct because the prescription found an optimal strategy, minimizing costs and using activities that improve crop yield. The prescription points out that the farmer should a) pick up daily the cotton buds (squares) of the cotton plants that have fallen to the ground, b) collect the bolls affected by the boll weevil to prevent further feeding and propagation of the boll weevils, c) fumigate the previously demarcated area, and d) increase water irrigation with an irrigation system. It should be noted that fumigation is recommended considering the previous demarcation, i.e., as the crop is in the fruiting stage, actions in previous stages should have included demarcation. Since the crop is in the fruiting stage, the prescription did not include crop analysis activities, fertilizer application, use of pheromone traps, or use of boll-kill weevil tubes, because they are economically unviable at this stage of cultivation. In brief, the prescriptive model gives an accurate suggestion regarding the expert opinion on cotton cultivation.

6.6 General discussion

Our proposal monitored the data and processed it to generate statistical analyses on the behavior of insect pests on cotton crops. A set of variables and expert opinions were considered to diagnose/predict cotton yield. Finally, we use the data processed above to prescribe the best strategy for integrated cotton crop management.

The classification task of the boll-weevil population was performed using XGBoost with 88% of accuracy using climate data. The results of the diagnosis/prediction of cotton yield showed that can a) manage the uncertainty from the variables of the context or the model, b) manage the knowledge of the experts to adapt the model, and c) use concurrently variables of the climate, of the pests, crops, and fertilizers. The results of the prescription task showed that using GA is possibly found the optimal strategy according to the context. Overall, these results show that the integrated use of data collection, data processing and decision-making technologies are useful in PF for cotton crop management.

6.7 Comparison with previous works

This study defines an ACODAT for integrated cotton management. The tasks have been validated by experts with good results in classification, diagnosis/prediction, and prescription tasks. We introduce a set of qualitative criteria in this section to compare our work with other related works. These criteria are:

- Criterion 1 - Uncertainty model: whether they proposed uncertainty models for diagnosis/prediction.
- Criterion 2 - Integrate management: whether they consider the integrated management of the crop.
- Criterion 3 - Production: whether they considered improve the production of the crops.
- Criterion 4 - Autonomous systems (AS) that include among other tasks, classification, diagnosis/prediction, and prescription tasks to improve the production.
- Criterion 5 - Simultaneous use of Climatic, pests, Fertilizer, and Crop variables (CLFCT).

According to the above criteria, Table 18 shows the comparison with the related works. The existing papers did not meet all the requirements. All the criteria we consider in our work are important because working together allow the operation of a robust system with autonomous tasks for integrated cotton crop management.

Table 18
Comparison with other works.

Work	Uncertainty model	Integrate management	Production	AS	CLFCT
Tribouillois et al. (2022)		✓	✓		
Aggarwal et al. (2022)		✓	✓		
Wu et al. (2020)		✓	✓		
Hajimirzajan et al., (2021)		✓	✓		
This work	✓	✓	✓	✓	✓

Abbreviation: CLFCT= Simultaneous use of Climatic, of pests, of Fertilizer, and of Crop variables. Production = Whether the study considered improving crop production. AS= Autonomous systems that include classification, diagnosis/prediction, and prescription tasks.

Some studies related used integrated management. For example, Tribouillois et al. (2022) build an integrated modeling of crop and water management to optimize irrigation. Hajimirzajan et al., (2021) defined a large-scale crop planning, which involves a comprehensive strategic framework that employs a decision support system to determine the sustainable use of water, as well as optimal crop selection, timing, and cultivation practices. Aggarwal and colleagues (2022) developed a system of geospatial analysis to preserve land fertility, optimize agricultural revenue, and minimize agricultural pollution and water consumption. Wu et al. (2020) developed a model for integrated nutrient management. It should be noted that the previous authors used integrated crop management because they considered different variables to have a broad management of the analyzed context. But no one of them uses different data analysis tasks, with different variables, and an autonomous cycle to integrate them, which our work does. They also do not consider knowledge obtained from expert recommendations to fit the model.

As previously discussed, our approach is the initial one to combine these criteria and propose an integrated cotton management approach using an ACODAT, which can be developed further with multi-agent systems (Aguilar et al., 2007; Terán et al., 2017). The purpose of integrating the multi-agent systems paradigm is to make the system more adaptable, extendable, and autonomous, as described by Vizcarrondo et al. (2017).

7 Conclusions

The objective of this work was to develop a system of PF using an ACODAT for the integrated management of cotton. The cycle used tasks of data processing, classification/prediction of cotton yield, and prescribing strategies for integrated cotton management. In the autonomous cycle, each task communicates with the next and passes processed information. Also, each task has its own AI techniques and the integration of all of them produces strategies according to the context of the crop. The combined use of data analysis tasks in one cycle provided notable advantages compared to isolated techniques. To our knowledge, this is the first work to use an autonomous architecture to support integrated cotton management.

We consider some limitations in this work. First, for the diagnosis/prediction of cotton yield, the fertilizer variable only included the amount used. Secondly, for the diagnosis/prediction of cotton yield we used only the behavior of the boll weevil. Future work should be aimed at improving the diagnosis/prediction model including more variables (e.g., specific fertilizers), and including the behavior of other insect pests and diseases. Third, this proposal did not include pheromone traps with real-time data updating in its case study. This would be an improvement that can be incorporated into the system to have more immediate feedback. In addition, we have planned to integrate this work with an autonomous cognitive architecture for agriculture. Our approach involves defining a meta-learning task, which will enable us to create models of weevil behavior specific to different regions. To achieve this, we will utilize the transfer learning paradigm, which involves transferring knowledge gained from one task to another related task. By doing so, we hope to improve the accuracy and efficiency of the system's predictions and provide valuable insights to farmers and other stakeholders in the agricultural sector. As a final point, the models we develop for weevil behavior will be integrated with our cognitive architecture, which is based on the multi-agent systems paradigm. Our decision to use this approach is rooted in the fact that agent theory has already established many effective modeling capabilities and implementations, which can be leveraged to improve the accuracy and efficiency of our models.

Acknowledgments

We thank the Colombian Agricultural Institute (ICA), Colombian Cotton Confederation (CONALGODON) and Colombian Institute of Hydrology, Meteorology and Environmental Studies (IDEAM). We thank Universidad EAFIT and Universidad de Córdoba for their support in this study.

Funding

This work was supported by the Colombian Science, Technology, and Innovation Fund (FCTeI) of the General Royalty System (SGR); Universidad EAFIT; and Universidad de Córdoba.

Conflicts of Interest

The authors declare there are no conflicts of interest.

Ethical Approval

Not applicable

References

Abbas, G., Fatima, Z., Tariq, M., Ahmed, M., Nasim, W., Rasul, G., Ahmad, S., & others. (2020). Applications of Crop Modeling in Cotton Production. In *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies* (pp. 429–445). Springer. https://doi.org/10.1007/978-981-15-1472-2_11

Aggarwal, S., Srinivas, R., Puppala, H., & Magner, J. (2022). Integrated decision support for promoting crop rotation based sustainable agricultural management using geoinformatics and stochastic optimization. *Computers and Electronics in Agriculture*, *200*. https://doi.org/10.1016/j.compag.2022.107213

Agostini, A., Alenyà, G., Fischbach, A., Scharr, H., Wörgötter, F., & Torras, C. (2017). A cognitive architecture for automatic gardening. *Computers and Electronics in Agriculture*, *138*, 69–79. https://doi.org/10.1016/j.compag.2017.04.015

Aguilar, J., Ardila, D., Avendaño, A., Macias, F., White, C., Gomez-Pulido, J., de Mesa, J. G., & Garces-Jimenez, A. (2020). An autonomic cycle of data analysis tasks for the supervision of HVAC systems of smart building. *Energies*, *13*(12). https://doi.org/10.3390/en13123103

Aguilar, J., Buendia, O., Pinto, A., & Gutiérrez, J. (2022). Social learning analytics for determining learning styles in a smart classroom. *Interactive Learning Environments*, *30*(2), 245–261. https://doi.org/10.1080/10494820.2019.1651745

Aguilar, J., Garcés-Jiménez, A., Gallego-Salvador, N., Gutiérrez de Mesa, J., Gomez-Pulido, J., & Garcìa-Tejedor, A. (2019). Autonomic management architecture for multi-HVAC systems in smart buildings. *IEEE Access*, *7*, 123402–123415. https://doi.org/10.1109/ACCESS.2019.2937639

Aguilar, J., Garces-Jimenez, A., R-Moreno, M. D., & García, R. (2021). A systematic literature review on the use of artificial intelligence in energy self-management in smart buildings. In *Renewable and Sustainable Energy Reviews* (Vol. 151). Elsevier Ltd. https://doi.org/10.1016/j.rser.2021.111530

Aguilar, J., Cerrada, M., & Hidrobo, F. (2007). A Methodology to Specify Multiagent Systems. *Lecture Notes in Computer Science*, *4496*, 92–101.

Ahmad, I., Ghaffar, A., Haider, G., Ahmad, A., Ahmad, B., Tariq, M., Nasim, W., Rasul, G., Fahad, S., Ahmad, S., & others. (2020). Climate Resilient Cotton Production System: A Case Study in Pakistan. In *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies* (pp. 447–484). Springer. https://doi.org/10.1007/978-981-15-1472-2_11

Ahmed, N., Ali, M. A., Danish, S., Chaudhry, U. K., Hussain, S., Hassan, W., Ahmad, F., & Ali, N. (2020). Role of Macronutrients in Cotton Production. In *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies* (pp. 81–104). Springer. https://doi.org/10.1007/978-981-15-1472-2_11

Ahmed, N., Ali, M. A., Hussain, S., Hassan, W., Ahmad, F., & Danish, S. (2020). Essential Micronutrients for Cotton Production. In *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies* (pp. 105–117). Springer. https://doi.org/10.1007/978-981-15-1472-2_11

Ali, M., Deo, R. C., Downs, N. J., & Maraseni, T. (2018). Cotton yield prediction with Markov Chain Monte Carlo-based simulation model integrated with genetic programing algorithm: A new hybrid copula-driven approach. *Agricultural and Forest Meteorology*, *263*, 428–448. https://doi.org/10.1016/j.agrformet.2018.09.002

Alves, A. N., Souza, W. S. R., & Borges, D. L. (2020). Cotton pests classification in field-based images using deep residual networks. *Computers and Electronics in Agriculture*, *174*(April), 105488. https://doi.org/10.1016/j.compag.2020.105488

Anees, M., & Shad, S. A. (2020). Insect pests of cotton and their management. In *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies* (pp. 177–212). Springer. https://doi.org/10.1007/978-981-15-1472-2_11

Astudillo, B., Santorum, M., & Aguilar, J. (2014). *A methodology for Data Analytic Based on Organizational Characterization through a User-centered Design: A Position Paper*. *2008*, 1–5.

Caldeira, R. F., Santiago, W. E., & Teruel, B. (2021). Identification of cotton leaf lesions using deep learning techniques. *Sensors*, *21*(9), 3169. https://doi.org/10.3390/s21093169

Cerrada, M., Aguilar, J., Colina, E., & Titli, A. (2005). Dynamical membership functions: An approach for adaptive fuzzy modelling. *Fuzzy Sets and Systems*, *152*(3), 513–533. https://doi.org/10.1016/j.fss.2004.10.004

Coulibaly, S., Kamsu-Foguem, B., Kamissoko, D., & Traore, D. (2022). Deep learning for precision agriculture: A bibliometric analysis. In *Intelligent Systems with Applications* (Vol. 16). Elsevier B.V. https://doi.org/10.1016/j.iswa.2022.200102

Cui, M., Qian, J., & Cui, L. (2022). Developing precision agriculture through creating information processing capability in rural China. *Journal of Rural Studies*, *92*, 237–252. https://doi.org/10.1016/j.jrurstud.2022.04.002

Eiben, Á. E., Hinterding, R., & Michalewicz, Z. (1999). Parameter control in evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, *3*(2), 124–141. https://doi.org/10.1109/4235.771166

Gandonou, J.-M. A. (2005). *Essays on precision agriculture technology adoption and risk management*. University of Kentucky.

Ghaffar, A., ur Rahman, M., Ali, H. R., Haider, G., Ahmad, S., Fahad, S., & Ahmad, S. (2020). Modern concepts and techniques for better cotton production. In *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies* (pp. 589–628). Springer. https://doi.org/10.1007/978-981-15-1472-2_11

Gosain, A., & Sardana, S. (2017). Handling class imbalance problem using oversampling techniques: A review. *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, *2017-Janua*, 79–85. https://doi.org/10.1109/ICACCI.2017.8125820

Hajimirzajan, A., Vahdat, M., Sadegheih, A., Shadkam, E., & Bilali, H. el. (2021). An integrated strategic framework for large-scale crop planning: sustainable climate-smart crop planning and agri-food supply chain management. *Sustainable Production and Consumption*, *26*, 709–732. https://doi.org/10.1016/j.spc.2020.12.016

Hassanat, A., Almohammadi, K., Alkafaween, E., Abunawas, E., Hammouri, A., & Prasath, V. B. S. (2019). Choosing mutation and crossover ratios for genetic algorithms-a review with a new dynamic approach. *Information*, *10*(12). https://doi.org/10.3390/info10120390

Hoyos, W., Aguilar, J., & Toro, M. (2022). An autonomous cycle of data analysis tasks for the clinical management of dengue. *Heliyon*, *8*(10). https://doi.org/10.1016/j.heliyon.2022.e10846

Hudgins, E. J., Liebhold, A. M., & Leung, B. (2017). Predicting the spread of all invasive forest pests in the United States. *Ecology Letters*, *20*(4), 426–435. https://doi.org/10.1111/ele.12741

Kandalkar, G., Deorankar, A. v, & Chatur, P. N. (2014). Classification of Agricultural Pests Using DWT and Back Propagation Neural Networks. *International Journal of Computer Science and Information Technologies (IJCSIT)*, *5*(3), 4034–4037. www.ijcsit.com

Khattab, A., Habib, S. E. D., Ismail, H., Zayan, S., Fahmy, Y., & Khairy, M. M. (2019). An IoT-based cognitive monitoring system for early plant disease forecast. *Computers and Electronics in Agriculture*, *166*(September), 105028. https://doi.org/10.1016/j.compag.2019.105028

Kong, Q., Kuriyan, K., Shah, N., & Guo, M. (2019). Development of a responsive optimisation framework for decision-making in precision agriculture. *Computers and Chemical Engineering*, *131*. https://doi.org/10.1016/j.compchemeng.2019.106585

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2016). *Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning*. http://arxiv.org/abs/1609.06570

Lobell, D. B., Hammer, G. L., McLean, G., Messina, C., Roberts, M. J., & Schlenker, W. (2013). The critical role of extreme heat for maize production in the United States. *Nature Climate Change*, *3*(5), 497–501. https://doi.org/10.1038/nclimate1832

Mammarella, M., Comba, L., Biglia, A., Dabbene, F., & Gay, P. (2021). Cooperation of unmanned systems for agricultural applications: A theoretical framework. *Biosystems Engineering*. https://doi.org/10.1016/j.biosystemseng.2021.11.008

Mansour, A. I., & Abu-Naser, S. (2019). Expert System for the Diagnosis of Wheat Diseases. *International Journal of Academic Information Systems Research (IJAISR)*, *3*(4), 19–26. www.ijeais.org/ijaisr

Maskey, M. L., Pathak, T. B., & Dara, S. K. (2019). Weather based strawberry yield forecasts at field scale using statistical and machine learning models. *Atmosphere*, *10*(7). https://doi.org/10.3390/atmos10070378

Masood, M. H., Saim, H., Taj, M., & Awais, M. M. (2020). *Early Disease Diagnosis for Rice Crop*. 1–5. http://arxiv.org/abs/2004.04775

Mckinney, W. (2010). *Data Structures for Statistical Computing in Python*.

Morales, L., Ouedraogo, C. A., Aguilar, J., Chassot, C., Medjiah, S., & Drira, K. (2019). Experimental comparison of the diagnostic capabilities of classification and clustering algorithms for the QoS management in an autonomic IoT platform. *Service Oriented*

Computing and Applications, *13*(3), 199–219. https://doi.org/10.1007/s11761-019-00266-w

Pacheco, F., Rangel, C., Aguilar, J., Cerrada, M., & Altamiranda, J. (2014). Methodological Framework for Data Processing based on the Data Science Paradigm. *XL Latin American Computing Conference*. https://doi.org/10.1109/CLEI.2014.6965184

Saha, S., Morita, T., Ospina, R., & Noguchi, N. (2022). A Vision-based Navigation System for an Agricultural Autonomous Tractor. *IFAC-PapersOnLine*, *55*(32), 48–53. https://doi.org/10.1016/j.ifacol.2022.11.113

Sanchez, M., Aguilar, J., Cordero, J., Valdiviezo-Diaz, P., Barba-Guaman, L., & Chamba-Eras, L. (2016). Cloud computing in Smart Educational environments: application in Learning Analytics as service. *Advances in Intelligent Systems and Computing*, *445*. https://doi.org/10.1007/978-3-319-31232-3

Sánchez, M., Exposito, E., & Aguilar, J. (2020). Implementing self-* autonomic properties in self-coordinated manufacturing processes for the Industry 4.0 context. *Computers in Industry*, *121*. https://doi.org/10.1016/j.compind.2020.103247

Say, S. M., Keskin, M., Sehri, M., & Sekerli, Y. E. (2018). Adoption of precision agriculture technologies in developed and developing countries. *The Online Journal of Science and Technology*, *8*(1), 7–15.

Singh, P. K., & Sharma, A. (2022). An intelligent WSN-UAV-based IoT framework for precision agriculture application. *Computers and Electrical Engineering*, *100*. https://doi.org/10.1016/j.compeleceng.2022.107912

Song, X., Yang, C., Wu, M., Zhao, C., Yang, G., Hoffmann, W. C., & Huang, W. (2017). Evaluation of Sentinel-2A satellite imagery for mapping cotton root rot. *Remote Sensing*, *9*(9), 1–17. https://doi.org/10.3390/rs9090906

Sterritt, R., Parashar, M., Tianfield, H., & Unland, R. (2005). A concise introduction to autonomic computing. *Advanced Engineering Informatics*, *19*(3), 181–187. https://doi.org/10.1016/j.aei.2005.05.012

Suleiman, F. M. (2019). Development of A Rule based System for Safflower Disease Diagnosis. *International Journal of Academic Engineering Research (IJAER)*, *3*(8), 1–9.

Terán, J., Aguilar, J., & Cerrada, M. (2017). Integration in industrial automation based on multi-agent systems using cultural algorithms for optimizing the coordination mechanisms. *Computers in Industry*, *91*, 11–23. https://doi.org/10.1016/j.compind.2017.05.002

Thenmozhi, K., & Srinivasulu Reddy, U. (2019). Crop pest classification based on deep convolutional neural network and transfer learning. *Computers and Electronics in Agriculture*, *164*(June), 104906. https://doi.org/10.1016/j.compag.2019.104906

Toscano-Miranda, R., Aguilar, J., Caro, M., Trebilcok, A., & Toro, M. (2023). *A fuzzy classification system to analyze the yield of cotton production*. Submitted to publication

Toscano-Miranda, R., Hoyos, W., Caro, M., Aguilar, J., Trebilcok, A., & Toro, M. (2022). A Classification Model of Cotton Boll-Weevil Population. *2022 XVLIII Latin AmericanvComputer Conference (CLEI)*, 1–5.

Toscano-Miranda, R., Toro, M., Aguilar, J., Caro, M., Marulanda, A., & Trebilcok, A. (2022a). Artificial-intelligence and sensing techniques for the management of insect pests and diseases in cotton: A systematic literature review. *Journal of Agricultural Science*, *160*(1). https://doi.org/10.1017/S002185962200017X

Toseef, M., & Khan, M. J. (2018). An intelligent mobile application for diagnosis of crop diseases in Pakistan using a fuzzy inference system. *Computers and Electronics in Agriculture*, *153*(February), 1–11. https://doi.org/10.1016/j.compag.2018.07.034

Trebilcok, A. (2020). *El cultivo del algodonero en Córdoba*.

Tribouillois, H., Constantin, J., Murgue, C., Villerd, J., & Therond, O. (2022). Integrated modeling of crop and water management at the watershed scale: Optimizing irrigation and modifying crop succession. *European Journal of Agronomy*, *140*. https://doi.org/10.1016/j.eja.2022.126592

Vizcarrondo, J., Aguilar, J., Exposito, E., & Subia, A. (2017). MAPE-K as a service-oriented Architecture. *IEEE Latin America Transactions*, *15*(6), 1163–1175.

Vulpi, F., Marani, R., Petitti, A., Reina, G., & Milella, A. (2022). An RGB-D multi-view perspective for autonomous agricultural robots. *Computers and Electronics in Agriculture*, *202*. https://doi.org/10.1016/j.compag.2022.107419

Wu, H., Yang, T., Liu, X., Li, H., Gao, L., Yang, J., Li, X., Zhang, L., & Jiang, S. (2020). Towards an integrated nutrient management in crop species to improve nitrogen and phosphorus use efficiencies of Chaohu Watershed. *Journal of Cleaner Production*, *272*. https://doi.org/10.1016/j.jclepro.2020.122765

Appendix F

Different transfer learning approaches for insect pest classification in cotton

Applied Soft Computing Journal 153 (2024) 111283

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Different transfer learning approaches for insect pest classification in cotton

Raul Toscano-Miranda [a,b], Jose Aguilar [b,c,d,*], William Hoyos [a,b,e], Manuel Caro [a], Anibal Trebilcok [f], Mauricio Toro [b]

[a] Department of Educational Informatics, Universidad de Córdoba, Colombia
[b] GIDITIC, Universidad Eafit, Medellín, Colombia
[c] CEMISID, Universidad de Los Andes, Mérida, Venezuela
[d] IMDEA Networks Institute, Leganés, Madrid, Spain
[e] ISI Research Group, Universidad Cooperativa de Colombia, Colombia
[f] Department of Agronomic Engineering, Universidad de Córdoba, Colombia

HIGHLIGHTS

- The classification of the boll-weevil population for developing effective pest management strategies.
- Transfer Learning (TL) approaches to improve the Insect Pest Classification in Cotton.
- Instance-based, feature-based, and parameter-based techniques for the pest management

ARTICLE INFO

Keywords:
Transfer learning
Machine learning
Artificial intelligence
Insect pests
Cotton management

ABSTRACT

Boll weevil is an important pest that affects cotton crops worldwide, causing significant economic losses. The classification of the boll-weevil population is crucial for developing effective pest management strategies. However, the low availability of data and features makes classification a challenging task. This study aimed to investigate the use of Transfer Learning (TL) techniques to improve the classification of boll weevil populations. Three types of TL techniques, instance-based, feature-based, and parameter-based, were studied to improve the classification performance of the machine learning algorithms. This work used data from two domains, one with few instances and the other with few features, to test the proposed approaches. Also, climate variables (temperature, humidity, and rainfall) were incorporated as features to predict the level of the boll-weevil attack. The most relevant results of this work are that define (1) How to measure and quantify the similarity or relationship between tasks of different domains; (2) How to select, align, or adapt the relevant features, instances, or models from the source task/domain to the target task/domain; (3) How to reuse parameter settings from the source domain; and (4) How to evaluate and validate the performance and robustness of the TL model on the target task/domain. The proposed approach achieved significant improvements in classification over previous results in the metrics of accuracy and F-measure. For example, in the case with few instances reached an accuracy of 90.79%, while in the case with few features reached an accuracy of 96.28%. Thus, the results demonstrate the effectiveness of TL techniques in improving the classification of boll-weevil populations in cotton crops when few data and/or features are available.

1. Introduction

Machine learning technologies have showcased the amazing capabilities of artificial intelligence in various technological applications. For example, with its multi-layered neural networks, deep learning excels in tasks such as image recognition and natural language processing [15,32]. Reinforcement learning and deep reinforcement learning are also types of machine learning where an agent learns to make decisions in an environment based on reward feedback [10]. They have demonstrated state-of-the-art performance in diverse domains, including game playing, robotics, and natural language processing technologies [19,22,29]. Recently, quantum machine learning has emerged, combining quantum

* Corresponding author at: GIDITIC, Universidad Eafit, Medellín, Colombia.
E-mail address: aguilarjos@gmail.com (J. Aguilar).

https://doi.org/10.1016/j.asoc.2024.111283
Received 16 June 2023; Received in revised form 6 December 2023; Accepted 10 January 2024
Available online 18 January 2024
1568-4946/

computing and machine learning algorithms [45,46], with the potential to solve complex problems beyond traditional techniques. These advancements continue to push the boundaries of AI. An area of machine learning of special interest in recent years is Transfer learning (TL), which utilizes pre-trained models to enhance learning on smaller datasets [8]. TL techniques have shown remarkable success in improving the performance of machine learning algorithms by transferring knowledge from one domain to another [39].

On the other hand, the boll weevil (Anthonomus grandis) is an important pest that affects cotton crops worldwide, causing significant economic losses [5,12]. The classification of the boll-weevil population is crucial for developing effective pest management strategies [35]. However, the classification of boll-weevil populations is a challenging task due to the limited availability of data and features. Thus, traditional classification methods have been used to classify boll weevil populations, but they have limitations in terms of accuracy. For instance, a previous work developed models to classify the population of boll weevils [34]. The results achieved good precision (88%), however, there were limitations related to the number of instances and characteristics in some contexts studied. In this work, we attack these limitations with TL.

1.1. Related works

TL techniques have been used to improve the classification of various pests in different contexts [3,9,14,15,17,32]. For example, TL techniques have been used to improve the classification of pests in crops such as citrus fruit, and tomato [14,15]. Specifically, [3] used TL to develop accurate models for agricultural classification tasks with few data. The study applied TL on ImageNet pre-trained models, where ImageNet was the generic dataset and AgriNet was the target dataset. The pre-trained models were then fine-tuned on the AgriNet dataset to improve their performance. The study found that VGG19 surpassed all other models with an accuracy of 94% and an F1 score of 92%. VGG16 was ranked second, followed by InceptionResNet-v2. The study evaluated the superiority of the proposed models using TL on two agricultural datasets. The AgriNet models achieved higher accuracies than the ImageNet models, and VGG19 was the best-performing model. Thenmozhi, Srinivasulu Reddy ($year$) [32] used TL to retrain deep-learning models and improve the efficiency and accuracy of insect classification tasks. The study used a wide range of insect pests from different field crops such as rice, maize, soybean, sugarcane, and cotton crops. The pre-trained models such as AlexNet, ResNet, GoogLeNet, and VGG were used as fixed feature extractors. By fine-tuning the pre-trained models with TL, the proposed convolutional neural networks (CNN) model achieved higher accuracy in insect classification compared to the pre-trained models alone. The proposed model was evaluated on three different insect datasets, and it achieved high accuracy for each dataset (between 92.25% and 95.97%). The study also analyzed the effects of different hyperparameters on the performance of the proposed model.

Similarly, [20] used TL with pre-trained CNNs to be adapted by retraining them with smaller datasets, with a different distribution than the larger datasets used to train the network. In this study, multiple types of CNN architecture (Densenet 201, Mobilenet, VGG 16, Hyper-parameter Search, and Inception V3) were used on agricultural image data for plant leaf disease detection, pest detection, and weed detection. The fine-tuned Inception V3 model achieved 87.85% accuracy, while the Mobilenet and VGG 16 models achieved accuracies of 91.85% and 78.71%, respectively. The Densenet model performed well with 99.62% accuracy, and the Hyper-parameter Search had 71.07% accuracy. Hadipour-Rokni et al., ($year$) [14] used TL with a deep learning model to leverage pre-existing knowledge from a previously trained model for a different task. The researchers used a pre-trained model on a large image dataset (ImageNet) to extract general features from the citrus fruit images and then fine-tuned the model using the dataset of citrus fruit images to classify the pests. The study found that TL was an effective technique for the early detection of pests in agricultural products using machine vision systems and deep learning. The AlexNet and GoogleNeT models had the highest accuracy (99.33% and 99.27%, respectively) in diagnosing citrus fruit disease, with the AlexNet model having the lowest calculation time. The study suggested that pre-trained models could be used in similar applications to save time and computational resources.

Additionally, [9] used TL to improve the accuracy of the classification model in insect pests. TL used the Inceptionv3 model, which achieved an accuracy of 67.88% in the test set. By leveraging the pre-trained layers of Inceptionv3, the authors were able to reduce the number of network parameters by 41% without affecting the accuracy and loss classification. Also, TL made it possible to use visualization methods to understand what the model has learned, identify biases in the data that affect the training process, and debug the model to visualize these biases. Finally, TL contributed to improving the overall performance of the deep learning model in this study. Huang et al., ($year$) [15] achieved that the knowledge learned from one problem was transferred to another problem in a different but related field. In this case, TL-based CNN models were used to identify tomato pests by transferring knowledge learned from other image recognition tasks. The authors improved the accuracy of tomato pest identification with CNN models (AlexNet, InceptionV3, VGG16, and ResNet50) and image augmentation technology. This approach improved learning efficiency and reduced training time. In summary, these studies used TL techniques with images to improve learning tasks in insect pest classification. However, to the best of our knowledge, the use of TL techniques to classify boll weevil populations has not been explored. Specifically, the use of TL techniques with structured data to classify boll weevil populations has not been analyzed. Therefore, there is a need to investigate the use of TL techniques to improve the classification of boll weevil populations.

In previous works, and to the best of our knowledge, the emphasis was usually on reusing previous models either to save training time or due to lack of data. There is no study that does an exhaustive analysis of how to work on cases where there is a lack of data and features in a problem at the same time. This work seeks to respond to this by proposing various TL schemes.

1.2. Contributions

This study aims to investigate the use of TL techniques to improve the classification of boll weevil populations by incorporating climate variables as features. Three types of TL techniques, instance-based, feature-based, and parameter-based, were studied to improve the classification performance of the XGBoost algorithm, which is the best machine learning algorithm for this type of task according to [34]. The study also aims to test the proposed approach using data from two domains, one with few instances and the other with few features. In summary, our study works with structured data about climate data, insect pest data, and three types of TL. The contributions of this study are the following:

- The definition of a procedure to measure/quantify the similarity or relatedness between two tasks or domains (called source and target). The experiments show how can affect the transferability and effectiveness of the knowledge transfer.
- The design of a method to select, align, or adapt the relevant features, instances, or models from the source task/domain to the target task/domain, which may require different strategies depending on the type and level of TL.
- The specification of a strategy to reuse parameter settings from the source domain and how to measure and determine their validity.
- The definition of a procedure to evaluate and validate the performance and robustness of the TL model on the target task or domain, with appropriate metrics and benchmarks.

The rest of the paper is organized as follows: Section 2 introduces the dataset of the boll weevil in cotton crops and the TL approaches existent

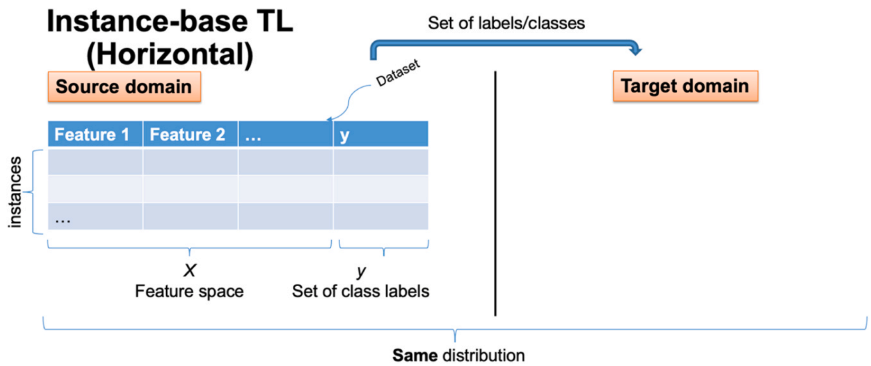


Fig. 1. Example of Instance-base TL.

in the literature. Section 3 presents the design of our approach of TL for the classification of boll-weevil populations. Section 4 shows the instantiation of our TL approach in different case studies in cotton crops. Section 5 presents the results of the case studies, and Section 6 concludes the paper by highlighting some of the future directions of this work.

2. Materials and method

2.1. Mathematical formulation of the TL problem

A domain $D = D(X, P_x)$ consists of a feature space X and a probability distribution P_x for each feature $x \in X$. Given a domain of interest, a task T can be defined by a label space Y and a predictive function f: X -> Y.

In TL, we have two domains, a source domain D_s and a target domain D_t. On the other hand, we have that T_s corresponds to the task executed in D_s and T_t corresponds to the task performed in D_t. Thus, if we have D_s, then T_s is represented as (X_s, Y_s), and in D_t, T_t is represented as (X_t, Y_t). In our case, the objective is to find the parameters, instances or features W_t for the task T_t to determine Y_t. In this case, the idea is to minimize the equation:

$$argmin_{W_t}(C(X_t, Y_t))$$

where $C(X_t, Y_t)$ is the cost function defined on task T_t of domain D_t to determine Y_t (for example, the error); and W_t is the result of a TL process, which can be of parameters, instances or features.

2.2. TL approaches in the literature

According to [26], TL approaches can be divided into four categories: instance-based transfer, feature-based transfer, parameter-based transfer, and relational-based transfer. These categories provide a general framework for understanding the different approaches to transfer learning and are the basis for the development of new TL methods.

2.2.1. Instance-based transfer learning

According to [26], Instance-based TL is an approach that assumes that certain parts of the data in the source domain can be reused for learning in the target domain. Thus, the instance-based transfer involves transferring instances from the source domain to the target domain (see Fig. 1). This approach involves measuring the similarity between a source and a target domain and selecting a similar source domain that has much more training data than the target domain. The approach can choose a pre-trained model that was learned from the source domain and fine-tunes it on the target domain using the re-weighted data [26, 40]. The rationale behind this approach lies in the premise that there are similarities between the source and target domains that can be exploited to improve performance in the target domain. By transferring specific instances from the source domain to the target domain, one seeks to leverage existing knowledge and adapt it to solve similar tasks in the new context. This technique becomes an effective strategy when the source domain has a large amount of training data and significant similarities to the target domain can be identified. However, this approach has its limitations and challenges. One of the potential problems is the assumption that instances from the source domain are applicable and useful in the target domain. If the similarities between the domains are not properly understood, or if there are subtle but significant differences between the data distributions of the two domains, instance-based transfer can lead to poor performance in the target domain. In addition, the quality of the transfer is highly dependent on the correct identification of the relevant instances and the similarity measure used to select them.

2.2.2. Feature-based transfer learning

Feature-based TL is an approach that involves transferring the feature representations learned from the source domain to the target domain [1,25,26]. This approach assumes that the feature spaces between the source and target domains are similar or can be aligned. The learned features from the pre-trained model are then fed as input to a

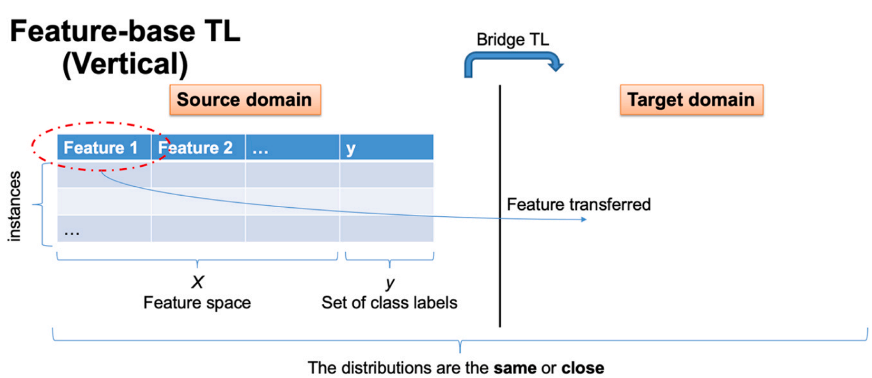


Fig. 2. Example of Feature-based TL.

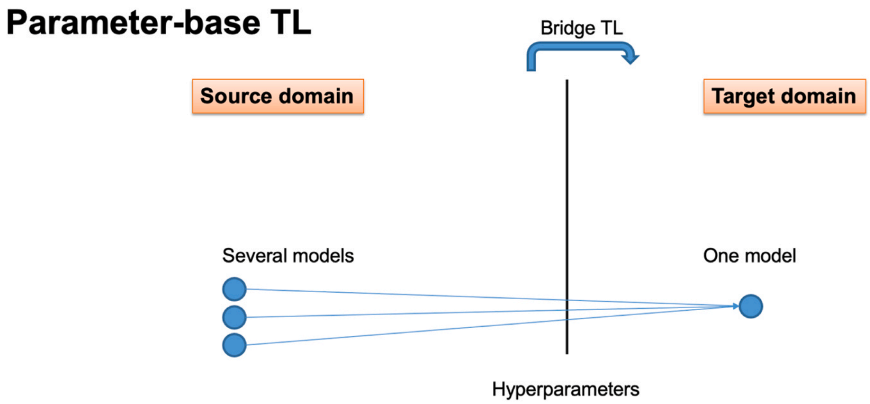


Fig. 3. Parameter-base TL.

new model, which is trained on a different dataset or task. The advantage of feature-based TL is that it can be used when there is not enough data to fine-tune the entire pre-trained model, but still, the learned features can be useful in the new task [23,25,26,41]. Fig. 2 shows the process of transferring features. The rationale behind the assumption that the feature spaces between source and target domains are similar or can be aligned is based on the idea that certain features relevant to a specific task can be generalized and reused in related tasks. Despite the logic behind the assumption, there are several limitations and factors that can affect the results of feature-based TL. For example, the introduction of learned features can lead to overfitting if the target dataset is small. On the other hand, while the learned features may be useful for generic tasks, certain tasks may require more specific knowledge that is not captured in the transferred features. In such cases, feature-based TL may not be sufficient to improve performance. Finally, in some cases, fine-tuning of the pre-trained model is necessary to better adapt it to the new task. Fine-tuning involves training some model settings on the new dataset to refine the learned features.

2.2.3. *Parameter-based transfer learning*

Parameter-based TL is an approach that involves transferring the parameters, or prior distribution of hyperparameters, from the source domain to the target domain (see Fig. 3). This approach assumes that the models for related tasks share some parameters or prior distribution of hyperparameters. This involves learning the source task first and then transferring the learned parameters to the target task. The pre-trained model is adapted to a new task by reusing some or all its pre-trained parameters, which are then fine-tuned on the new task using additional data. The advantage of parameter-based TL is that it can lead to higher performance on the new task, especially when the new task is like the pre-training task [4,6,26]. The rationale behind this approach lies in the observation that certain features and patterns learned during the pre-training task may be applicable and relevant to the target task. By transferring the pre-trained parameters, the target model can benefit from this prior knowledge, which can speed up the training process and, in some cases, significantly improve performance on the new task. This is especially true when the target task is like the pre-training task. However, it is important to note that the success of parameter-based TL is highly dependent on the similarity between the source task and the target task. If the tasks are too different in terms of structure, nature of data or requirements, then direct parameter transfer may not be beneficial or even detrimental to performance on the target task. In addition, another critical factor that can affect the results of parameter-based TL is the quantity and quality of data available for the target task. If the target task data is sparse or of low quality, then pretrained parameter transfer may be more prone to overfit the model to the limited training data, leading to poor performance on unseen data. In such situations, it is important to consider strategies such as regularization and careful fine tuning to avoid overfitting.

2.2.4. *Relational-based transfer learning*

Relational-based TL is an approach that focuses on learning the relations between the source and target domains [21,26]. Particularly, relational-based transfer involves transferring relational knowledge from the source domain to the target domain (see Fig. 4). This approach finds past knowledge in the source domain to be used in the current context by the target domain. This assumes that there is a relationship between the source and target domains that can be leveraged to improve the performance of the target task. Relational-based TL can be used in scenarios where the domains of the source and target tasks are not the same but interrelated [11,26,30].

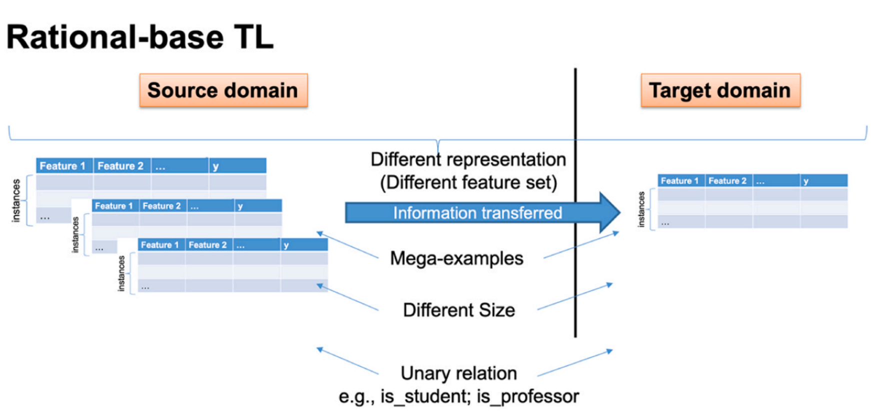


Fig. 4. Relational-base TL.

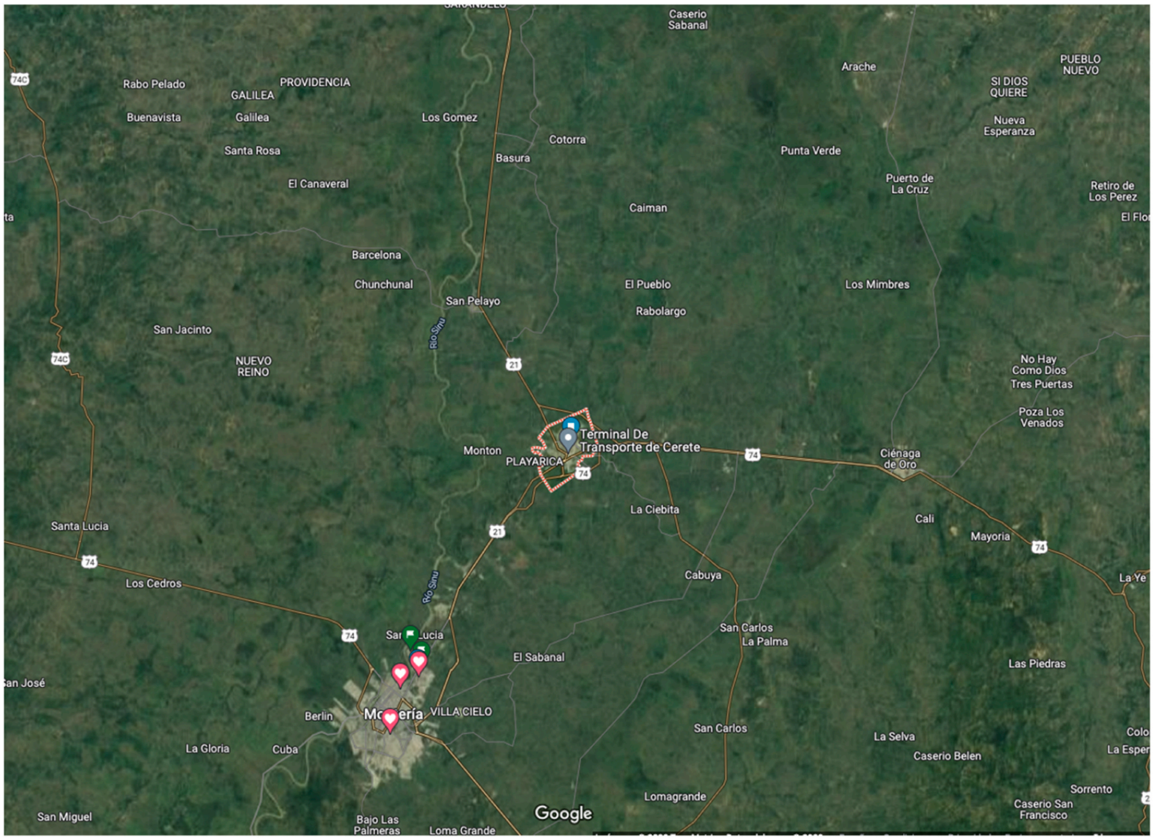


Fig. 5. Geographical regions of cotton crops in Córdoba.

2.3. Study area and datasets

To validate our proposal, we utilized cotton crops as a case study, which are in various regions of Córdoba, Colombia. These regions include cities that make up the Sinú Valley, namely High, Middle, and Low Sinú, as mentioned in [36], and are situated at approximately ~8°55'33.6"N, 75°48'16.5"W (see Fig. 5). We collected data for our study from the net of monitoring of the boll weevil established by the Colombian Agricultural Institute (ICA) and climate data recorded by the Colombian Institute of Hydrology, Meteorology and Environmental Studies (IDEAM). Specifically, we analyzed data from the cities of Montería, Cereté , Lorica, and Ciénaga de Oro, covering the period of 2018 to 2021. The reason for selecting these areas was due to their cultivation of cotton and the availability of pheromone trap records.

Datasets from regions of the Sinú Valley were used. These datasets included three climatic variables (rainfall, humidity, and temperature) and the number of red and black boll weevils captured in pheromone traps. *Rainfall* is the amount of rain that falls during the day, measured in millimeters. Rainfall can affect the survival and reproduction of boll weevils, as well as the growth and quality of cotton plants. *Humidity* is the hourly relative humidity (average of the day), measured in percentage. Humidity can also influence the development and activity of boll weevils, as well as the susceptibility and resistance of cotton plants to pests and diseases. *Temperature* is the average daily temperature, measured in degrees Celsius. Temperature can determine the life cycle and population dynamics of boll weevils, as well as the phenology and yield of cotton plants. *Red boll* weevils are the youngest stage of the pest, which has not yet reached maturity. This feature represents the number of captures of these boll weevils in a trap. Red boll weevils indicate the presence and intensity of infestation, as well as the potential damage to the cotton buds and flowers. *Black boll* weevils are the ones that can procreate, which have reached maturity and can mate and lay eggs. This feature represents the number of captures of these boll weevils in a trap.

Table 1
Dataset distribution.

City	Samples	Temperature	Humidity	Rainfall
Lorica	1800~	✓	✓	✓
Cereté	4000~	✓	✓	✓
Ciénaga de Oro	900~	✓	| ✓ Black	| ✓ Black
Montería	1000~	✓		| ✓ Black

Black boll weevils indicate the reproductive capacity and the future generation of the pest, as well as the potential damage to the cotton bolls and seeds. Table 1 shows the distribution after processing and cleaning the dataset [34]. Lorica, Cereté included the climate variables and the red and black boll weevils. However, Ciénaga de Oro and Montería did not have enough samples for the red boll weevils and climatic variables. This particularity made it suitable to apply our proposal of TL.

We used five techniques, eXtreme Gradient Boosting (XGBoost), Random Forest, Decision Tree, Support Vector Machine, and Artificial Neural Network with multilayer perceptron, which were chosen because they have been techniques that have given very good results in previous works in various disciplines [35]. In that work, these techniques were subjected to a process of optimizing their parameters before comparing them, using the grid search hyperparameter optimization method [2]. As a result of the comparison of these five techniques, XGBoost gave the best accuracy (88%). These analyses and results were presented in previous works [34]. Therefore, XGBoost is used in the present study.

In this work, several domains with few instances or few features are presented, which reach very poor quality metrics values compared to the domains with more instances and features. This is an opportunity to design and apply TL strategies, in order to improve the results of domains with few instances or features.

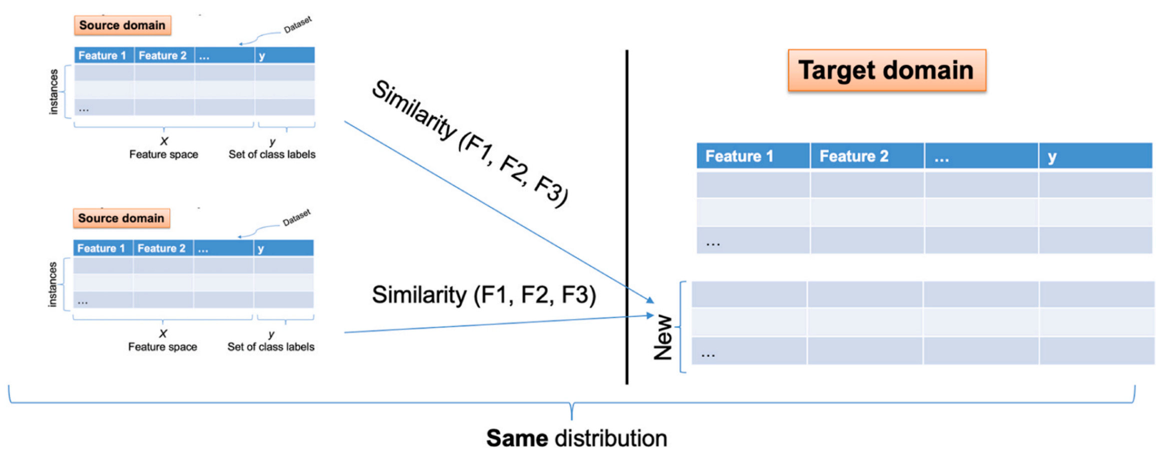


Fig. 6. Instance-based TL in our case study.

3. Design of our TL approaches

The TL was applied from the source domain (X^s) to the target domain (X^t) using a metric to determine the similarity between the domains. For the determination of this similarity metric, a statistical analysis with mean, standard deviation and variance was used. This study uses the instance-, feature-, and parameter-based TL techniques. In the next subsections, we will describe each technique in our work.

Algorithm 1, the instances of the source domain that are like the target domain are determined, and then, added to the target domain dataset as new instances. Then, in step 5 of Algorithm 1, the model for the target domain is trained with both old instances and new instances. Finally, in step 6, the set of tests is used in the model of the target domain to evaluate its quality.

Algorithm 1. : instance-based TL algorithm.

Input:
X^s: Source domain of boll weevil
X^t: Target domain of boll weevil

Output: TL in the target domain

1. Train several $X^s{}_i$
2. Test $X^s{}_i$
3. Analysis of statistical similarity of all instances of the best $X^s{}_i$'s vs X^t, using the features F1, F2, F3
4. The best instances similar to the selected $X^s{}_i$ pass as new instances to the X^t
5. Train with the entire X^t with the new instances
6. Test X^t

3.1. *Instance-based transfer learning*

This technique uses a horizontal treatment of the data set (that is, samples/instances). From now on, we will call them instances. The source domain transfers data because the datasets of the target domain are few. Fig. 6 shows the use of the instance-based TL approach. We use the similarity of the instances with all the features (F1, F2, F3) between the source domain and target domain to compare the different sources. The source domain that has major similarity with the target domain, and also, good accuracy, is selected. Thus, this process generates new instances in the target domain according to the similarity of the instances.

Algorithm 1 shows the steps to reach the instance-based TL. A similarity threshold is defined to establish the instances to select. Instances with a similarity greater than 75% were selected. Thus, during step 4 of

3.2. *Feature-based transfer learning*

This technique uses a vertical treatment of the dataset. We used as the source domain the datasets of the cities with the best accuracy and similarity to the target domain. Suppose we have three features (e.g., F1, F2 and F3). The selected source domain transfers features to the target domain because the target domain datasets do not have all the features (missing features, e.g., F2 and F3). We applied statistical similarity of the common features (i.e., F1) between possible source domains (with good accuracy) and the target domain to select one of them. Like the previous technique, samples were selected whose features (columns) obtained a similarity greater than 75%. Then, F2 and F3 from the selected source domain are transferred to the target domain (see Fig. 7). Note that the similarity analysis was made based on the statistical metrics of F1.

Algorithm 2 shows the steps to reach the TL. A similarity threshold was defined to establish the instances to select. In this case, samples were selected whose features (columns) obtained a similarity greater than 75%. In step 3.2 of Algorithm 2, the source dataset whose F1 is most similar to the F1 of the target dataset is selected. Then in step 4, the most similar instances according to feature F1 of the selected source dataset are selected to take their other features. In step 5, the new features (F2, F3) are added to the target domain dataset. Finally, the model of the target domain is trained and tested with both the old and new features (see steps 6 and 7).

Algorithm 2. : feature-based TL algorithm.

Input:
X^s: Source domain of boll weevil
X^t: Target domain of boll weevil

Output: TL in the target domain

1. Train several $X^s{}_i$
2. Test $X^s{}_i$
3. Select the best $X^s{}_i$'s
   3.1 Analysis of similarity of F1 (temperature) between each selected $X^s{}_i$ and X^t
   3.2 Select the $X^s{}_i$ more similar to X^t
4. Select the instances more similar from the selected $X^s{}_i$ (according to F1)
5. The instances more similar pass to the target domain with their new features (F2, F3) to the X^t
6. Train with the entire X^t with the new features
7. Test X^t

3.3. Parameter-based transfer learning

With the parameter-based TL technique, we improved the target domain (X^t) using the parameters of the best model applied to the source domain (X^s). For this purpose, firstly, we selected the best model trained on the source domain. Second, we transferred the parameters of this model to the model of the target domain to improve it. Source domains (X^s) are those machine learning models with the highest precision and whose data sets it trained on have a most statistically similar to the target domain (X^t). The most similar and best model is the one used to transfer all its parameters (see Fig. 8).

Algorithm 3 shows the steps to reach the TL. The best source domain

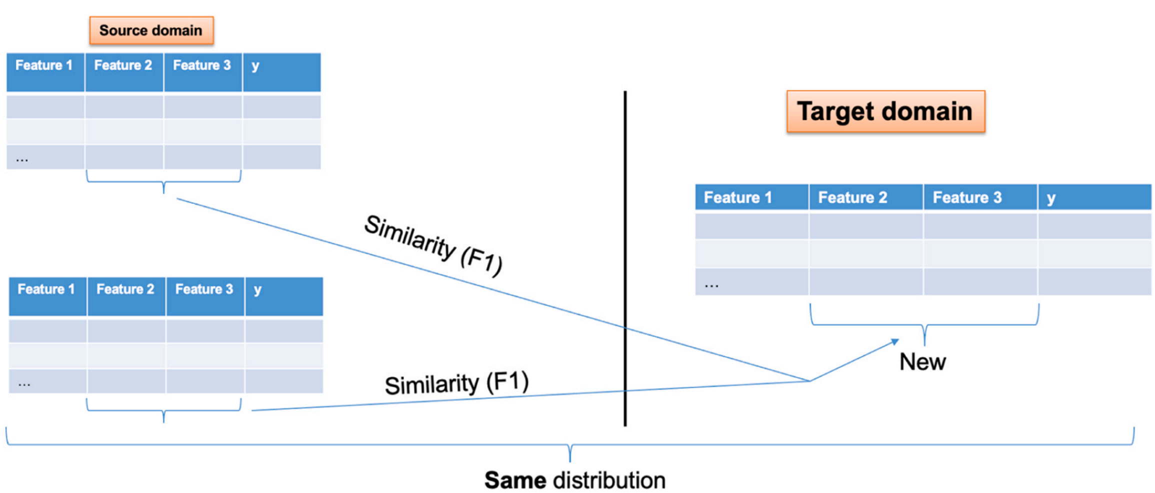


Fig. 7. Feature-based TL in our case study.

is selected for transferring its parameters to the target domain (see steps 3.1 and 3.2). In step 4, the model of the target domain is trained with the new parameters. Finally, in step 5, testing is performed with the target domain model.

Algorithm 3. parameter-based TL algorithm.

Input:
X_1^s: Source domain of boll weevil
X_2^s: Source domain of boll weevil
X^t: Target domain of boll weevil

Output: TL in the target domain

1. Train several X^s_i's
2. Test X^s_i's
3. If the accuracy of X^t is not good
   3.1 Select the best X^s_i
   3.2 Transfer the parameters of the selected X^s to X^t
4. Train X^t with the transferred parameters
5. Test X^t

4. Instantiation of our TL approaches in our case study

TL techniques were applied using the XGBoost algorithm. XGBoost is the technique with the best results in previous works [13,18,31,34]. The experiments included the datasets with information about climate data (temperature, rainfall, and humidity) and the level attack of the red boll weevil. The black boll weevil obtained low accuracy in all the cases (lower than 70%) in a previous work [34], and therefore, it was not used in this study.

Table 2 shows the distribution of the dataset in each city and the TL technique that was used in each case study. In previous work, Lorica city had the best results of accuracy in the classification model (see Table 2 and [34]). The city of Cereté had more samples but less accuracy than the city of Lorica. Therefore, we used parameter-based TL from Lorica to Cereté to improve its accuracy. Ciénaga de Oro city has less samples than Lorica city, therefore, we used instance-based TL from Lorica to Ciénaga de Oro to improve its accuracy using the similarities between these domains. Monteria city did not have all the climatic data. Montería only had the temperature. Therefore, we used with Montería city a feature-based TL approach to improve the accuracy using the statistical similarity between the common features.

Eq. (1) and Eq. (2) were used to determine the similarity between source and target domains. The similarity of each feature is given by:

$$S(i) = \left[1 - \frac{\left|X^i_{source}\right| - X^i_{target}}{\max\left(X_{target}\right)}\right] \quad (1)$$

where i indicates the current feature, S is the percentage of similarity, X_{source} is the source domain, X_{target} is the target domain.

The similarity of all features per instance is given by,

$$S(h) = \frac{1}{n}\sum_{i=1}^{n} S(i) \quad (2)$$

where h is the current instance and n is the number of features.

4.1. Instance-based transfer learning

The purpose of this technique was to improve the target domain (Ciénaga de Oro) using as the source domain Cereté and Lorica. In a

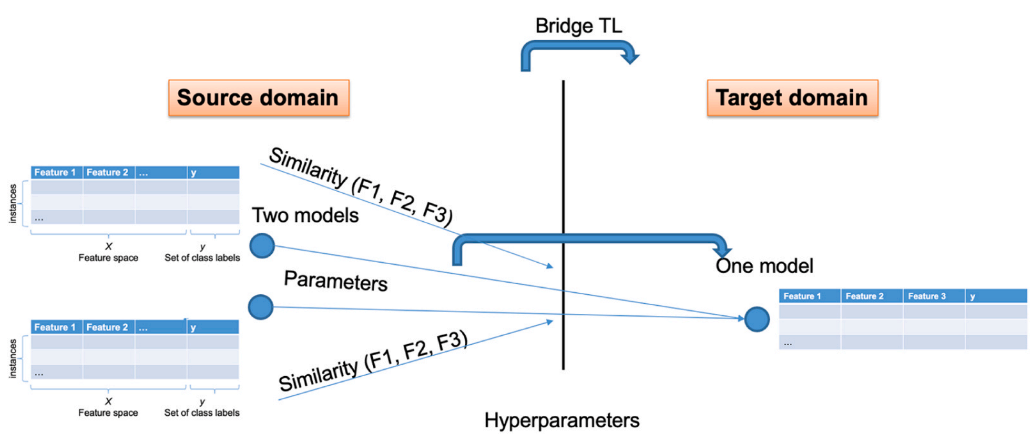


Fig. 8. Parameter-based TL in our case study.

Table 2
Dataset distribution vs TL technique.

City	Remarks	Samples	TL	Previous Accuracy
Lorica	The best accuracy. Used as a source domain	1800~	NA	88%
Cereté	With more instances. Used as source and target domains	4000~	C	76,68%
Ciénaga de Oro	Used as a target domain	900~	A, C	NA
Montería	Used as a target domain	1000~	B	NA

Abbreviations: A = Instance-based TL, B = Feature-based TL, C = Parameter-based TL, NA = Not apply

Table 3
Target domain: Distribution of the quantity of instances per Low, Medium, and High class.

Class of boll weevil	Instances
Low (0 to 4)	946
Medium (5 to 20)	36
High (> 20)	3
Total	985

previous work [34], the Ciénaga de Oro domain failed because it did not have enough instances in each class. The dataset of the target domain had 985 instances. The quantity of red boll weevil captured in pheromone traps was recorded and classified as Low, Medium, and High. The Low class means the number of red boll weevils between 0 to 4. The Medium class means the number of red boll weevils between 5 to 20. High class means the number of red boll weevils is greater than 20. For the Low class, there were 946 instances with information about of the number of boll weevils, 36 for Medium, and 3 for High (see Table 3).

In this case, the oversampling technique failed in the target domain due to there were few instances in the High class.

4.2. Feature-based transfer learning

In a previous work ([34], the results of the Montería domain only included the feature of temperature. For this reason, it was selected as the target domain in this study. Thus, the Montería dataset is the target domain because it has one feature of climate. In our case study, we used as the source domain the datasets of the cities with the best accuracy (i.e., Lorica and Cereté). The source domains have three features (temperature, humidity, and rainfall).

Table 4
Target domain and source domains with their instances and features.

City	Domain	Instances	Features		
			Temperature	Humidity	Rainfall
Montería	Target	1052	✓		
Lorica	Source	1775	✓	✓	✓
Cereté	Source	4083	✓	✓	✓

Table 5
Increase of new instances in the target domain.

Class	S-L	T-C-O	T-C-TL		
	Instances	*Instances*	*A*	*B*	*C*
0	1668	946	2614	2591	2544
1	95	36	129	127	113
2	12	3	11	8	8
Total	1775	985	2754	2726	2665
Increase of new instances:			1769	1741	1680
Percentage increase:			179.59%	176.75%	**170.56%**

Similarity between source and target domains: A: 75%, B: 90%, C: 95%.
Abbreviations: S-L= Source Lorica, T-C-O: Target - Ciénaga de Oro - Original, T-C-TL: Target - Ciénaga de Oro - Processed with TL.

We applied statistical similarity of the common features (in this case, the temperature) between the possible source domains and the target domain. Feature 2 (humidity) and feature 3 (rainfall) from the selected source domain, are then transferred to the target domain according to the similarity of feature 1 (temperature) between the selected source domain and the target domain. The similarity analysis was made based on Eq. (1).

The source domain included climate data (temperature, rainfall, and humidity) and the level attack of the red boll weevil (i.e., categorized as Low, Medium, or High class). The two domains were compared focusing on the common feature (temperature). If the instance of the target domain was in the same class (level attack Low, Medium, or High) of the source domain, then Eq. (1) was applied to determine the percentage of similarity between the feature of the temperature of the target domain and source domain. If the feature was in the threshold of similarity, then the other two features (humidity and rainfall) were transferred as new features to the target domain (for the same instance).

For the target domain, 1052 instances with only one feature were analyzed. In the source domains, 1775 and 4083 instances with three features were analyzed (see Table 4). In summary, only the temperature features that reach the similarity threshold are considered to pass their features to the target domain as new features.

4.3. Parameter-based transfer learning

In our case study, first, we selected the dataset of the city of Cereté as the target domain. Second, we use a parameter-based TL to improve its model results. We used the machine learning models developed by the dataset from the city of Lorica as source domains to improve the machine learning models from the city of Cereté . Thus, the parameters of the best models in the selected source domain are transferred to the model in the target domain. The Cereté dataset was used as the target domain because the precision was lower than that of Lorica. Thus, this technique used the configuration of the parameters from the source domain to the destination domain. In this way, this experiment aims to reduce the time to configure hyperparameters in the target domain.

5. Results

In this section, the proposed approaches to improve the prediction using TL paradigm are presented. The results of ([34] were improved with our three techniques of TL. Based on the confusion matrix, three metrics (Accuracy, Recall, and F1 score) were used to evaluate the performances of our models. These metrics are given by [24],

$$Accuracy = \frac{Number \quad of \quad boll \quad weevil \quad correctly \quad predicted}{Total \quad number \quad of \quad input \quad boll \quad weevil \quad samples}$$

$$Recall = \frac{Number \quad of \quad boll \quad weevil \quad correctly \quad predicted}{Total \quad number \quad of \quad true \quad cases}$$

$$F1score = 2 * \frac{Recall * Accuracy}{Recall + Accuracy}$$

Additionally, to test the results of our parameter-based transfer learning approach, we defined the following hypotheses [28,37]:

$$H_0 : \overline{\gamma}_{TL} = \gamma_{no-TL}$$

$$H_1 : \overline{\gamma}_{TL} > \gamma_{no-TL}$$

where $\overline{\gamma}_{TL}$ is the accuracy mean of 1000 runs on the testing set using the parameters transferred from Lorica city and γ_{no-TL} corresponds to the accuracy of the model without the application of transfer learning. To test the hypotheses, the Student's t test with superior tail alternative and 95% confidence was used. Finally, for each of the case studies, the

Table 6
Increase of new instances in the target domain using as source domains the combination of Lorica and Cereté .

Class	S-LC	T-C-O	T-C-TL		
	Instances	*Instances*	*A*	*B*	*C*
0	5510	946	7402	7379	7332
1	268	36	338	334	276
2	80	3	58	23	11
Total	5858	985	7798	7736	7619
Increase of new instances:			6813	6751	6634
Percentage increase:			**691,68%**	685,38%	673,50%

Similarity between source and target domains: A: 75%, B: 90%, C: 95%.
Abbreviations: S-LC= Source Lorica+Cereté , T-C-O: Target - Ciénaga de Oro - Original, T-C-TL: Target - Ciénaga de Oro - Processed with TL.

XGBoost parameters were optimized using the grid search hyper-parameter optimization method [2].

5.1. Instance-based transfer learning

In this case study, we used as source domain the datasets of the cities with the best accuracy (i.e., Lorica and Cereté cities). The target domain was Ciénaga de Oro. After applying Eq. (1) and Eq. (2), the experiments were conducted with 75%, 90%, and 95% of similarity. Also, three experiments were conducted: The first experiment included the source domain of Cereté and the data of the red boll weevil. The second experiment included the source domain of Lorica and the data of the red boll weevil. The third experiment included the source domain using the combination of instances of Cereté and Lorica with the data of the red boll weevil.

The two domains (source and target) were compared using the Algorithm (1): if the instance of the target domain was in the same class (Low, Medium, or High-level attack) as the source domain, then the Eq. (2) was applied to determine the percentage of similarity. If the instance was in the threshold of similarity, then it was transferred as a new instance to the target domain. Finally, new instances were added, with a minimum of 170.56% (see Table 5) and a maximum of 691.68% (see Table 6) of increase. Table 5 shows the increase of new instances in the target domain using the best results from the best source domain (Lorica) and different similarity thresholds (A, B, C).

Table 6 shows that the combination of the two source domains (Lorica and Cereté) added more instances than just Lorica (shown in Table 5).

The target domain was then balanced with SMOTE and normalized with StandardScaler.

Table 7 shows the results of the experiments with the different combinations of the source domains and the similarity threshold. The results showed that the model was improved and gave an accuracy of 90.79%.

In general terms, the results showed that the accuracy increased with the similarity. It means that using 95% of similarity as the threshold gave the best results. Also, of the source domains used, Lorica showed the best accuracy. It is worth mentioning that experiments with 98% of similarity failed in oversample. Also, the results show that the instance-based TL gave better results in Ciénaga de Oro city (90.79%), compared with the best result of Lorica city (88%) found in [34]. In general, with least similarity threshold, the experiments gave less precision, although more instances were added (see Tables 5 and 6). On the other hand, with the source domain of Lorica is obtained the best results than with other combinations (e.g., Cereté, or Lorica + Cereté). Finally, instance-based TL helped a target domain that was having trouble finding predictions because it didn't have enough instances can now achieve higher accuracy. The results showed that instance-based TL can achieve high accuracy rates by selecting the most similar instances from the source domain to the target domain, based on a similarity measure. This reduces the negative transfer and increases the relevance of the transferred data.

Table 7
Results for the set of testing using the target domain to Ciénaga de Oro and three source domains.

Source domains	A		B		C	
	Accuracy	*F1-Score*	*Accuracy*	*F1-Score*	*Accuracy*	*F1-Score*
Cereté	0.8329	0.8329	0.8821	0.8821	FoO	
Lorica	0.9018	0.9018	0.9074	0.9074	**0.9079**	**0.9079**
Lorica + Cereté	0.8982	0.8982	0.8862	0.8862	0.8875	0.8875

Similarity between source and target domains: A: 75%, B: 90%, C: 95%

Table 8
Description of the four experiments with feature-based TL.

Experiment	Description
First (SMOTE)	Two features were added, and the dataset was balanced with SMOTE.
Second (Hybrid: Manual + SMOTE)	Two features were added. Additionally, a set of instances of the High class of boll-weevil attack level were selected of the source domain and added to the target domain. This set of instances had three features. Then, the dataset was balanced with SMOTE.
Third (Pure)	Two features were added.
Fourth (Automatic hybrid)	Two features were added. Then, new instances were automatically added using the instance-based TL approach Finally, the dataset was balanced with SMOTE.

5.2. Feature-based transfer learning

In this case study, we use as source domains the datasets of the cities with the best precision (i.e., the cities of Lorica and Cereté) and as the target domain to Montería. The datasets included information related to the red boll weevil and climate data. Four experiments were conducted. In each experiment, two new features were added to the target domain using Eq. (1), and then, further actions were applied as follows: The first experiment used class balanced with SMOTE. The second experiment included new instances belonging to the High class of boll-weevil attack level. These instances were manually selected. Then, the entire dataset was balanced with SMOTE. In the third experiment, SMOTE was not used. The fourth experiment applied an automatic hybrid technique (feature-based plus instance-based). After adding the two features using feature-based TL, new instances were automatically added using instance-based TL by Eq. (2). Then, the dataset was balanced with SMOTE. Table 8 shows a summary of these experiments.

In summary, of the four experiments to test the feature-based TL approach, two experiments had instances added (manually or using the instance-based TL approach) and the other two did not. The experiments were conducted with 75%, 90%, and 95% of similarity.

Table 9 shows the results of the four experiments with the three similarity thresholds. The first experiment had better quality by including the source domain of the city of Lorica and the oversampling technique with SMOTE. This was because similarity was better between the temperature characteristic of the source and target domains, compared to the other three experiments. The second experiment gave a lower precision than the first, although the difference was small. Manually selecting certain instances helped but was not the best strategy. The third experiment with the pure technique could not oversample because there were not enough instances in the high class. Finally, the results showed that the predictive model improved with the feature-based TL technique and gave an accuracy of up to 96.28%, surpassing the results of the city of Lorica (88%) found in [34]. The feature-based TL technique demonstrates that it is possible to transfer learning from one domain to another. The results showed that feature-based TL can

Table 9
Results for the set of testing using the target domain to Montería and four experiments source domains.

Experiment	A		B		C	
	Accuracy	*F1-Score*	*Accuracy*	*F1-Score*	*Accuracy*	*F1-Score*
First (SMOTE)	0.9442	0.9442	0.93	0.93	**0.9628**	**0.9628**
Second (Hybrid: Manual + SMOTE)	0.9256	0.9256	0.9344	0.9344	0.9584	0.9584
Third (Pure)	FoO	FoO	FoO	FoO	FoO	FoO
Fourth (Automatic hybrid)	0.887	0.887	0.8928	0.8928	0.8836	0.8836

Similarity between source and target domains: A: 75%, B: 90%, C: 95%. Abbreviation: FoO = Fail on oversample.

achieve high accuracy rates by using cross-domain feature similarity analysis, allowing those features to be reused in the target domain to enrich its information and build more robust models.

5.3. Parameter-based transfer learning

The purpose of the experiment was to assess whether the best model for one city could result in better classification results for Cereté . As a hyperparameter optimization process was carried out in XGBoost, we transferred the values of the parameters with which the best performance of the model was obtained, which significantly affected the results. Table 10 shows the hyperparameters transferred from the source domain to the target domain, based on the previous work [33], where a significant number of hyperparameters were tested and it was found that those defined in Table 10 are the ones that are really relevant to obtain the best possible model. To mention one example, the parameter called *min _weight_fraction_leaf*, which defines the minimum weighted fraction of the total sum of weights, was not carried over because it does not affect the model performance. Other parameters such as *min-_impurity_decrease, random_state, verbose*, among others, also do not contribute to the superior performance of the model, therefore, they were not transferred either.

Table 11 shows the mean and standard deviation of the metrics that evaluate the performance of the models. Additionally, it shows the results of the statistical test to check if the use of TL allows the improvement of the performance of the model without TL. Based on the results, the use of the parameter-based TL approach significantly improves the performance (accuracy and F1-score) of the model without TL (p-value $= < 0.001$).

In summary, the results showed an increase in accuracy levels from 67% to 79.2%. In other words, the parameters that worked best for Lorica (88%) also performed better in Cereté .

Table 10
Brief description of parameters transferred from the source domain to the target domain.

Parameter	Brief description	Value
subsample	The portion of samples allocated for fitting the individual base learners.	0.8
n_estimators	The quantity of boosting rounds to execute.	1000
min_child_weight	denotes the smallest total weight a node must have to split into a child.	1
max_depth	Defines the depth of the tree based on the quantity of splits performed.	4
learning_rate	Reduces the tree weights in each round of boosting	0.4

Table 11
Descriptive statistics and Student's t-test (p-value) results to determine performance improvement using TL.

Metric	$\overline{\gamma}_{TL}$	σ_{TL}	γ_{no-TL}	p-value
Accuracy	0.792	0.006	0.670	< 0.001
F1-Score	0.771	0.006	0.710	< 0.001

6. Comparison with previous works

This study aimed to investigate the use of TL techniques to improve the classification of boll weevil populations. Three types of TL techniques, instance-based, feature-based, and parameter-based, were studied to improve the classification performance of the machine learning algorithms. We introduce a set of qualitative criteria in this section to compare our work with other related works. These criteria include:

Criterion 1 - Integration of three techniques of TL: whether they proposed the use of instance-based, feature-based, and parameter-base TL to improve the knowledge models.
Criterion 2 - TL for insect pests: whether they consider knowledge models related to insect pests such as boll weevil.
Criterion 3 - Climate data: whether they considered the climate data using structured data.
Criterion 4 - TL for prediction model: whether they propose a TL approach for prediction models.
Criterion 5 - Adaptability: whether the proposal can be applied to other knowledge models in other fields.

According to the above criteria, Table 12 shows the comparison with the related works. The existing works did not meet all the requirements. All the criteria were considered in our work because they allow improving the accuracy of the predictive models implemented in previous works.

Some works have used TL techniques to improve the performance of machine learning algorithms by transferring knowledge from one domain to another. For example, TL techniques have been used to

Table 12
Qualitative comparison with other works.

Criteria	[1]	[2]	[3]	[4]	Our work
Integration of three techniques of TL					✓
TL for insect pests					✓
Climate data					✓
Prediction model integration with TL	✓	✓	✓	✓	✓
Adaptability	✓	✓	✓	✓	✓

Abbreviations: [1] Thenmozhi & Srinivasulu Reddy [32], [2] Meena et al. [20], [3] [14]), [4] Coulibaly et al. [9].

Table 13
Quantitative comparison with other works.

Work	Instance	Feature	Parameter
	Accuracy	Accuracy	Accuracy
[42]	90.68%	NU	NU
[38]	80%	NU	NU
[7]	95.62%	NU	NU
[43]	NU	84%	NU
[27]	NU	81.8%	NU
[16]	NU	97.9%	NU
[44]	NU	81.13%	NU
Our	90.79%	96.28%	79.2%

Abbreviation: NU= Not used.

improve the classification of insect pests in several crops [9,14,20,32]. These works used a feature-based TL approach to improve the quality of the models in the target domains. In general, the benefits reflect increased accuracy, efficiency, and the use of pre-trained models that could be used in similar applications to save time and computational resources. However, none of these works includes the study of three TL techniques to improve the machine-learning models. This integration facilitates the identification of the TL technique that leads to enhanced performance in the target domain. In addition, our proposal includes climatic data to determine the behavior of the boll weevil, which is not usually considered. Also, the prediction model was integrated with TL and can be adapted to other scenarios or application domains.

Finally, some works applied TL techniques on structured data to improve results when there is little data. Thus, three works used instance-based TL [7,38,42]. Four works used feature-based TL with structured data [16,27,43,44]. Unlike previous works, our approach used three techniques of TL (instance-, feature-, and parameter-based TL), which represents an improvement with respect to previous works. Our approach is more flexible and adaptable to different scenarios and domains, by allowing us to select, transform and adjust the most relevant instances, features and/or parameters from the source domain to the destination domain, as required by the context of the problem. This improves the accuracy and robustness of the models, by reducing the risk of negative transfers, overfitting, or underfitting, by enhancing the generalization and representation capabilities of the models, as can be seen in Table 13, where very good results were obtained with our approach for the different types of TL.

7. Conclusions

This study aimed to investigate the use of TL techniques to improve the classification of boll-weevil populations by incorporating climate variables as features. Three types of TL techniques, instance-based, feature-based, and parameter-based, were employed to improve the classification performance of the XGBoost algorithm. The study used data from two domains, one with few instances and the other with few features, to test the proposed approach.

Particularly, the study demonstrates the potential of the different TL types to overcome the limitations of the availability of data or features, which are common challenges in data analysis in many domains. TL can leverage the existing knowledge from related domains to enhance the learning performance in new domains, thus reducing the need for costly and time-consuming data collection. On the other hand, the TL approaches studied in this work can be applied to other classification problems similar to the one studied in this work, where there are cases with enough structured data from which good models can be built, and other cases with little data or characteristics. The application domains can be very different, from agricultural to medical and industrial fields.

Specifically, the results of this study have important implications for the development and implementation of climate-smart pest management strategies for cotton crops and can be adapted to other crops. Climate-smart pest management reduces pest-induced crop losses, enhances ecosystem services, reduces the greenhouse gas emissions intensity per unit of food produced, and strengthens the resilience of agricultural systems in the face of climate change. The utilization of TL techniques to improve the classification of boll-weevil populations based on climate variables is useful for monitoring and predicting pest dynamics and risks in different regions and seasons.

The proposed approach achieved significant improvements in classification accuracy for both a few-instance domain and a few-feature domain. The target domain with few instances reached an accuracy of 90.79%, while the target domain with few features reached an accuracy of 96.28%. The highest accuracies were found with the 95% similarity threshold. In addition, parameter-based TL experiments were performed. The tests showed that the target domain improved accuracy. The results demonstrate the effectiveness of TL techniques in improving the classification of boll-weevil populations in cotton crops when few data and characteristics exist.

Our models were tested with metrics such as Accuracy, F1 score, and Student's t-test for the estimation of the quality of the prediction; however, other metrics could be added. In this sense, in future works, it would be interesting to test with other metrics, in order to test the sensibility of our approach. Also, we plan to add other cities with more instances, and test the cases where it failed by oversample. In addition, we plan to merge the work done here with previous research on autonomous cycles in integrated cotton crop management. Finally, we would like to use TL techniques to enhance learning in the application of metacognitive functions in a metacognitive architecture for agriculture.

Ethical Approval

Not applicable.

Funding

This work was supported by the Colombian Science, Technology, and Innovation Fund (FCTeI) of the General Royalty System (SGR); Universidad EAFIT; and Universidad de Córdoba.

CRediT authorship contribution statement

Trebilcok Anibal: Data curation, Formal analysis, Investigation, Methodology, Writing – review & editing. **Toro Mauricio:** Funding acquisition, Investigation, Methodology, Writing – review & editing. **Hoyos William:** Data curation, Formal analysis, Investigation, Methodology, Writing – original draft. **Toscano-Miranda Raul:** Data curation, Formal analysis, Software, Writing – original draft, Investigation. **Caro Manuel:** Funding acquisition, Investigation, Writing – review & editing. **Aguilar Jose:** Conceptualization, Funding acquisition, Investigation, Project administration, Writing – original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We thank the Colombian Agricultural Institute (ICA), Colombian Cotton Confederation (CONALGODON) and Colombian Institute of Hydrology, Meteorology and Environmental Studies (IDEAM). We thank Universidad EAFIT and Universidad de Córdoba for their support in this study.

References

[1] J. Aguilar, A. Garcès-Jimènez, N. Gallego-Salvador, J. De Mesa, J. Gomez-Pulido, A. Garcìa-Tejedor, Autonomic management architecture for multi-HVAC systems in smart buildings, IEEE Access 7 (2019) 123402–123415, https://doi.org/10.1109/ACCESS.2019.2937639.

[2] J. Aguilar, C. Salazar, H. Velasco, J. Monsalve-Pulido, E. Montoya, Comparison and evaluation of different methods for the feature extraction from educational contents, Computation 8 (2) (2020) 1–20, https://doi.org/10.3390/COMPUTATION8020030.

[3] Al Sahili, Z., Awad, M., 2023. The power of transfer learning in agricultural applications: AgriNet. Convolutional neural networks and deep learning for crop improvement and production.

[4] S. Bashath, N. Perera, S. Tripathi, K. Manjang, M. Dehmer, F.E. Streib, A data-centric review of deep transfer learning with applications to text data, Inf. Sci. 585 (2022) 498–528, https://doi.org/10.1016/j.ins.2021.11.061.

[5] E. Ben Guerrero, R. Salvador, P. Talia, Evaluation of hydrolytic enzyme activities from digestive fluids of Anthonomus grandis (Coleoptera: Curculionidae), Arch. Insect Biochem. Physiol. (2020), https://doi.org/10.1002/arch.21730.
[6] A. Chakraborty, C. Anitescu, X. Zhuang, T. Rabczuk, Domain adaptation based transfer learning approach for solving PDEs on complex geometries, Eng. Comput. 38 (5) (2022) 4569–4588, https://doi.org/10.1007/s00366-022-01661-2.
[7] Chowdhury, S.B.R., Annervaz, K.M., & Dukkipati, A. , 2021. Instance-based inductive deep transfer learning by cross-dataset querying with locality sensitive hashing. In: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource Natural Language Processing - Proceedings, 183–191. https://doi.org/10.18653/v1/d19–6120.
[8] T. Cody, P.A. Beling, A systems theory of transfer learning, IEEE Syst. J. 12 (1) (2023), https://doi.org/10.1109/jsyst.2022.3224650.
[9] S. Coulibaly, B. Kamsu-Foguem, D. Kamissoko, D. Traore, Explainable deep convolutional neural networks for insect pest recognition, in: Journal of Cleaner Production, 371, Elsevier Ltd,, 2022, https://doi.org/10.1016/j.jclepro.2022.133638.
[10] C. Cuartas, J. Aguilar, J. Hybrid algorithm based on reinforcement learning for smart inventory management, J. Intell. Manuf. 34 (2023) 123–149, https://doi.org/10.1007/s10845-022-01982-5.
[11] O. Day, T.M. Khoshgoftaar, A survey on heterogeneous transfer learning, J. Big Data 4 (1) (2017), https://doi.org/10.1186/s40537-017-0089-0.
[12] J.F.J. Grigolli, L.A. Souza, M.G. Fernandes, A.C. Busoli, Spatial distribution of adult anthonomus grandis Boheman (Coleoptera: Curculionidae) and damage to cotton flower buds due to feeding and oviposition, Neotrop. Entomol. 46 (4) (2017) 442–451, https://doi.org/10.1007/s13744-016-0471-1.
[13] F. Guo, Z. Liu, W. Hu, J. Tan, Gain prediction and compensation for subarray antenna with assembling errors based on improved XGBoost and transfer learning, IET Microw., Antennas Propag. 14 (6) (2020) 551–558, https://doi.org/10.1049/iet-map.2019.0182.
[14] R. Hadipour-Rokni, E. Askari Asli-Ardeh, A. Jahanbakhshi, I. Esmaili paeen-Afrakoti, S. Sabzi, Intelligent detection of citrus fruit pests using machine vision system and convolutional neural network through transfer learning technique, Comput. Biol. Med. 155 (2023), https://doi.org/10.1016/j.compbiomed.2023.106611.
[15] M.L. Huang, T.C. Chuang, Y.C. Liao, Application of transfer learning and image augmentation technology for tomato pest identification, Sustain. Comput.: Inform. Syst. 33 (2022), https://doi.org/10.1016/j.suscom.2021.100646.
[16] I. Jung, J. Lim, H. Kim, Pf-tl: payload feature-based transfer learning for dealing with the lack of training data, Electronics 10 (10) (2021), https://doi.org/10.3390/electronics10101148.
[17] W. Li, T. Zheng, Z. Yang, M. Li, C. Sun, X. Yang, Classification and detection of insects from field images using deep learning for smart pest management: a systematic review, Ecol. Inform. 66 (2021), https://doi.org/10.1016/j.ecoinf.2021.101460.
[18] W. Liu, W.D. Liu, J. Gu, Predictive model for water absorption in sublayers using a joint distribution adaption based XGBoost transfer learning method, J. Pet. Sci. Eng. 188 (2020), https://doi.org/10.1016/j.petrol.2020.106937.
[19] W. Liu, B. Wang, J. Fan, Y. Ge, M. Zidan, A quantum system control method based on enhanced reinforcement learning, Soft Comput. 26 (14) (2022) 6567–6575, https://doi.org/10.1007/s00500-022-07179-5.
[20] S.D. Meena, M. Susank, T. Guttula, S.H. Chandana, J. Sheela, Crop yield improvement with weeds, pest and disease detection, Procedia Comput. Sci. 218 (2023) 2369–2382, https://doi.org/10.1016/j.procs.2023.01.212.
[21] L. Mihalkova, R.J. Mooney, Transfer learning from minimal target data by mapping across relational domains, IJCAI Int. Jt. Conf. Artif. Intell. (2009) 1163–1168 (July).
[22] L. Morales, C.A. Ouedraogo, J. Aguilar, C. Chassot, S. Medjiah, K. Drira, Experimental comparison of the diagnostic capabilities of classification and clustering algorithms for the QoS management in an autonomic IoT platform, Serv. Oriented Comput. Appl. 13 (3) (2019) 199–219, https://doi.org/10.1007/s11761-019-00266-w.
[23] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2014).
[24] F. Pacheco, C. Rangel, J. Aguilar, M. Cerrada, J. Altamiranda, Methodological framework for data processing based on the data science paradigm, XL Lat. Am. Comput. Conf. (2014), https://doi.org/10.1109/CLEI.2014.6965184.
[25] Pan, J. , 2010. Feature-based transfer learning with real-world applications. Hong Kong.
[26] J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359, https://doi.org/10.1109/TKDE.2009.191.
[27] H. Qing, L. Biwen, S. Beijun, Y. Xia, Cross-project software defect prediction using feature-based transfer learning, ACM Int. Conf. Proc. Ser. 6 (2015) 74–82, https://doi.org/10.1145/2875913.2875944.
[28] M. Sánchez, J. Aguilar, J. Cordero, P. Valdiviezo-Díaz, L. Barba-Guamán, L. Chamba-Eras, Cloud computing in smart educational environments: application in learning analytics as service, in: Á. Rocha, A. Correia, H. Adeli, L. Reis, M. Mendonça Teixeira (Eds.), New Advances in Information Systems and Technologies. Advances in Intelligent Systems and Computing vol. 444, Springer, 2016, https://doi.org/10.1007/978-3-319-31232-3_94.
[29] M. Sánchez, E. Exposito, J. Aguilar, Implementing self-* autonomic properties in self-coordinated manufacturing processes for the Industry 4.0 context, Comput. Ind. (2020) *121*, https://doi.org/10.1016/j.compind.2020.103247.
[30] Tan, B., Zhang, Y., Pan, S.J., & Yang, Q. , 2017. Distant domain transfer learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. ⟨http://qzone.qq.com⟩.
[31] Tawalbeh, S., Hammad, M., & Al-Smadi, M. , 2020. SAJA at TRAC 2020 Shared task: Transfer learning for aggressive identification with XGBoost.
[32] K. Thenmozhi, U. Srinivasulu Reddy, Crop pest classification based on deep convolutional neural network and transfer learning, Comput. Electron. Agric. 164 (June) (2019) 104906, https://doi.org/10.1016/j.compag.2019.104906.
[33] Toscano-Miranda, R., Aguilar, J., Caro, M., Trebilcok, A., & Toro, M., 2023. A fuzzy classification system to analyze the yield of cotton production. Submitted to Publication.
[34] Toscano-Miranda, R., Hoyos, W., Caro, M., Aguilar, J., Trebilcok, A., & Toro, M., 2022. A classification model of cotton boll-weevil population. In: Proceedings of the 2022 XVLIII Latin American Computer Conference (CLEI), 1–5. https://doi.org/10.1109/CLEI56649.2022.9959893.
[35] R. Toscano-Miranda, M. Toro, J. Aguilar, M. Caro, A. Marulanda, A. Trebilcok, Artificial-intelligence and sensing techniques for the management of insect pests and diseases in cotton: a systematic literature review, J. Agric. Sci. 160 (2022) 16–31, https://doi.org/10.1017/S002185962200017X.
[36] Trebilcok, A., 2020. El cultivo del algodonero en Córdoba.
[37] Vizcarrondo, J., Aguilar, J., Exposito, E., Subias, A., 2012. ARMISCOM: Autonomic reflective middleware for management service composition. In: Proceedings of the Global Information Infrastructure and Networking Symposium (GIIS), 1–8. https://doi.org/10.1109/GIIS.2012.6466760.
[38] M. Wang, H. Yang, Research on personal credit risk assessment model based on instance-based transfer learning, Int. J. Intell. Sci. 11 (01) (2021) 44–55, https://doi.org/10.4236/ijis.2021.111004.
[39] X. Xu, H. Qiao, X. Ma, G. Yin, Y. Wang, J. Zhao, H. Li, An automatic wheat ear counting model based on the minimum area intersection ratio algorithm and transfer learning, Measurement (2023) 112849, https://doi.org/10.1016/j.measurement.2023.112849.
[40] Q. Yang, Y. Zhang, W. Dai, S.J. Pan, Instance-based transfer learning. Transfer Learning, Cambridge University Press,, 2020, pp. 23–33, https://doi.org/10.1017/9781139061773.004.
[41] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks? Adv. Neural Inf. Process. Syst. 27 (2014).
[42] L. Zhang, L. Guo, H. Gao, D. Dong, G. Fu, X. Hong, Instance-based ensemble deep transfer learning network: a new intelligent degradation recognition method and its application on ball screw, Mech. Syst. Signal Process. 140 (2020) 106681, https://doi.org/10.1016/j.ymssp.2020.106681.
[43] Zhao, J., Shetty, S., Pan, J.W., 2017. Feature-based transfer learning for network security. In: Proceedings - IEEE Military Communications Conference MILCOM, 17–22. https://doi.org/10.1109/MILCOM.2017.8170749.
[44] X. Zhong, S. Guo, H. Shan, L. Gao, D. Xue, N. Zhao, Feature-based transfer learning based on distribution similarity, IEEE Access 6 (2018) 35551–35557, https://doi.org/10.1109/ACCESS.2018.2843773.
[45] M. Zidan, A.H. Abdel-Aty, A. Khalil, M. Abdel-Aty, H. Eleuch, A novel efficient quantum random access memory, IEEE Access 9 (2021) 151775–151780, https://doi.org/10.1109/ACCESS.2021.3119588.
[46] M. Zidan, S.F. Hegazy, M. Abdel-Aty, S.S.A. Obayya, Rapid solution of logical equivalence problems by quantum computation algorithm, Appl. Soft Comput. 132 (2023) 109844, https://doi.org/10.1016/j.asoc.2022.109844.