The Detection of Paddy Brown Leaf Spot and Bacterial Leaf Blight Using Interpretable Image Processing with Human-Centred AI

A dissertation submitted for the Degree of Master of Computer Science

P. H. Senanayake

University of Colombo School of Computing

2024

DECLARATION

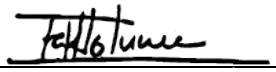
Name of the student: P. H. Senanayake
Registration number: 2018/MCS/082
Name of the Degree Programme: Master of Computer Science
Project/Thesis title: The Detection of Paddy Brown Leaf Spot and Bacterial Leaf Blight Using Interpretable Image Processing with Human-Centred AI

1. The project/thesis is my original work and has not been submitted previously for a degree at this or any other University/Institute. To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.
2. I understand what plagiarism is, the various types of plagiarism, how to avoid it, what my resources are, who can help me if I am unsure about a research or plagiarism issue, as well as what the consequences are at University of Colombo School of Computing (UCSC) for plagiarism.
3. I understand that ignorance is not an excuse for plagiarism and that I am responsible for clarifying, asking questions and utilizing all available resources in order to educate myself and prevent myself from plagiarizing.
4. I am also aware of the dangers of using online plagiarism checkers and sites that offer essays for sale. I understand that if I use these resources, I am solely responsible for the consequences of my actions.
5. I assure that any work I submit with my name on it will reflect my own ideas and effort. I will properly cite all material that is not my own.
6. I understand that there is no acceptable excuse for committing plagiarism and that doing so is a violation of the Student Code of Conduct

Signature of the Student	**Date (DD/MM/YYYY)**
	14/09/2024

Certified by Supervisor(s)

This is to certify that this project/thesis is based on the work of the above-mentioned student under my/our supervision. The thesis has been prepared according to the format stipulated and is of an acceptable standard.

	Supervisor 1	**Supervisor 2**	**Supervisor 3**
Name	Dr. Thilina Halloluwa		
Signature			
Date	14/09/2024		

I want to dedicate this thesis to my family and many friends. Special gratitude to my beloved parents, whose encouragement and push for persistence resound in my ears.

I also dedicate this dissertation to my friends who have supported me. I will always appreciate all he has done, especially W. M. C. Wickramathunga, for helping me develop my technology skills and for the many hours of proofreading.

ACKNOWLEDGEMENTS

Without the assistance of the supervisors, this effort would not have been possible. I am especially indebted to Dr Thilina Halloluwa, Lecture at the University of Colombo School of Computing, and Mr D. T. Bamunuarachchi, Lecture at the University of Colombo School of Computing, for their invaluable advice, continuous support, and patience during my study. Their vast knowledge and wealth of experience have inspired me in my academic research and daily life.

I appreciate everyone I had the privilege of working with on this research. Without my classmates, especially W. M. C. Wickramathunga, whose editing assistance, late-night feedback sessions, and moral support made this endeavour feasible. I should also extend my gratitude to the university's librarians, research assistants, and study participants, who impacted and encouraged me. I want to thank the management at Pearson and my co-workers for giving me the time to do the research.

The people in my family have been more significant to me in pursuing this mission than anyone else. I want to thank my parents for their support and affection in whatever I do. They are the ideal examples to follow.

ABSTRACT

Detecting agricultural diseases, such as paddy brown leaf spot (BLS) and bacterial leaf blight (BLB), poses significant global challenges to crop management and food security. Leveraging the advancements in image processing and artificial intelligence, this research investigates the application of human-centred artificial intelligence (AI) techniques for interpretable disease detection in paddy fields. This study addresses the critical need for transparent and understandable AI models by integrating human-centred design principles with state-of-the-art explainable AI (XAI) techniques. Through a comprehensive literature review, we explore the landscape of image processing algorithms, disease characteristics, and XAI methodologies, laying the groundwork for our research.

The methodology section outlines an evaluation of the two trained models through XAI models tailored for object detection. Emphasis is placed on the ethical considerations and human-centric design choices guiding the implementation process. Theoretical frameworks elucidate the foundations of image processing algorithms, machine learning models, and human-centred design principles, providing a holistic understanding of the research context.

Implementation details delve into dataset descriptions, XAI model configurations, and training procedures. The results and analysis section evaluates the performance and interpretability of the XAI models, incorporating user feedback and perception analysis to assess the system's usability. Case studies showcase the real-world application of the XAI system, including agricultural settings, and highlight its impact on disease detection and farming practices. Future directions outline potential enhancements and ethical considerations for further research and implementation.

The Human Centred explainable AI (HCXAI) approach involves iterative analysis, including "why not" and "what if" questions, to refine the model. Insights from the first iteration highlight key findings, challenges, and opportunities, leading to actionable recommendations for improving model performance, data quality, interpretability, fairness, and robustness. These enhancements, prioritised based on potential impact and feasibility, are aligned with stakeholder objectives and resource constraints. Clear goals and performance metrics for subsequent iterations are established to measure success. This iterative, human-centred approach ensures responsible use of technology, promoting ethical, safe, and mindful engagement, ultimately leading to improved performance, transparency, and trustworthiness in the AI system's deployment and operation.

In conclusion, this research contributes to advancing interpretable AI in agriculture, bridging the gap between technological innovation and human-centred design principles. By empowering stakeholders with transparent and understandable AI models, we aim to revolutionise disease management practices and foster sustainable agricultural development.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1. INTRODUCTION

In the realm of modern agriculture, ensuring the health and vitality of crops is paramount to safeguarding food security and sustaining agricultural economies (Hu and Hillary, 2023). Two significant threats to paddy crops, namely brown leaf spot and bacterial leaf blight, have posed formidable challenges to farmers and researchers alike. These diseases can rapidly spread and devastate fields, resulting in significant yield losses (Deng et al., 2021). As a result, for efficient disease management and crop protection, proper and timely detection of these diseases is essential.

Conventional approaches to disease identification in paddy fields frequently depend on labour-intensive, time-consuming, and susceptible to human error manual examination. The convergence of image processing and AI has recently opened up new possibilities for revolutionising how we detect and manage crop diseases (Rawat et al., 2023). This intersection of technologies allows us to harness the power of machine learning and computer vision to provide rapid, accurate, and scalable solutions for identifying paddy brown leaf spot and bacterial leaf blight.

This introductory chapter serves as a gateway to understanding the transformative potential of image processing through human-centred artificial intelligence in addressing these agricultural challenges (Holzinger et al., 2022). It lays the foundation for comprehensively exploring the methods, tools, and applications discussed throughout this work.

The problem is defined as the interpretability of detecting Paddy Brown Leaf Spot and Bacterial Leaf Blight diseases.

Paddy brown leaf spot (BLS) and bacterial leaf blight (BLB) are two distinct yet equally devastating diseases that afflict paddy crops (Oryza sativa).

Paddy Brown Leaf Spot (BLS**)**: BLS primarily occurs in the fungus Bipolaris oryzae and is distinguished by the formation of minute, brown lesions on the rice plant's leaves. These lesions can coalesce and cover large portions of the leaf surface, resulting in decreased crop output, stunted growth, and reduced photosynthesis. Early detection of BLS is essential to implement timely disease management strategies, such as fungicide application or crop rotation, to mitigate its impact.

Bacterial Leaf Blight (BLB**)**: BLB, on the other hand, occurs by the bacterium Xanthomonas oryzae pv. oryzae is known for its rapid spread and destruction of paddy leaves. Infected leaves exhibit water-soaked lesions that turn yellow and eventually wither, severely affecting the plant's ability to photosynthesise and produce grains. Effective management of BLB relies on early detection and resistant crop varieties.

The dual challenge of timely detection and differentiation of these diseases in the vast expanses of paddy fields calls for innovative solutions that combine the strengths of image processing and human-centred artificial intelligence. By harnessing the interpretability of these technologies, we can empower farmers with accessible tools for early diagnosis and informed decision-making, ultimately safeguarding crop health and ensuring food security.

This study embarks on a journey to explore the integration of image processing techniques and AI-driven interpretability in the context of paddy disease detection. In the following chapters, we will delve into the intricacies of image acquisition, feature extraction, machine learning algorithms, and user-centric interfaces, all working harmoniously to provide farmers and agricultural stakeholders with practical solutions to combat paddy brown leaf spot and bacterial leaf blight. Through this exploration, we aim to not only define the problem but also present a pathway toward its solution, one that is both innovative and human-centred.

1.1.Motivation

The motivation for conducting this research on "The Detection of Paddy Brown Leaf Spot and Bacterial Leaf Blight Through the Interpretability of Image Processing via Human-centred Artificial Intelligence" is rooted in the pressing need to address critical challenges in modern agriculture, enhance food security, and empower farming communities.

Agriculture is the lifeblood of societies worldwide, providing sustenance, livelihoods, and economic stability. However, this vital sector faces increasingly complex challenges, including climate change, diminishing resources, and the relentless onslaught of crop diseases. Among these challenges, the threat posed by paddy brown leaf spot (BLS) and bacterial leaf blight (BLB) is a substantial impediment to agricultural sustainability and food production.

Agricultural Sustainability: Sustainable agriculture maximises yield, minimises inputs such as pesticides and fertilisers, reduces environmental impact, and promotes biodiversity. The overreliance on traditional disease detection methods in paddy fields, often involving

indiscriminate pesticide use, threatens agriculture's ecological balance and long-term viability. Our study promotes sustainable practices by enabling precise, targeted disease management.

Food Security: With the global population on the rise, ensuring food security is paramount. Crop diseases like BLS and BLB can lead to significant yield losses, exacerbating food shortages and affecting the livelihoods of millions of farmers. Our research aims to contribute to food security by providing a tool for early disease detection, which is fundamental to crop protection and yield optimisation.

Empowering Farmers: Smallholder farmers, who make up a considerable proportion of the global farming community, frequently encounter barriers to obtaining sophisticated agricultural technologies. We intend to bridge this technology gap by developing human-centred artificial intelligence solutions. Empowering farmers with accessible tools for disease detection allows them to make informed decisions, reduce losses, and enhance their economic well-being.

Interdisciplinary Innovation: Integrating image processing and artificial intelligence in agriculture represents a promising intersection of disciplines. Our motivation is driven by a passion for innovation that leverages cutting-edge technologies to address real-world challenges, bringing together computer science, agriculture, and human-centred design.

Interpretability and Trust: As AI systems become increasingly integral to decision-making, it is essential to ensure they are interpretable and transparent. By focusing on the interpretability of AI models in the context of disease detection, we aim to build trust among users and stakeholders, making the technology more accessible and user-friendly.

In summary, this study is motivated by the urgent need to address the challenges posed by paddy brown leaf spot and bacterial leaf blight in agriculture. We aim to contribute to sustainable agriculture, enhance food security, empower farmers, foster interdisciplinary innovation, and promote transparent, user-centric technology by harnessing the potential of image processing and human-centred artificial intelligence. This research endeavours to catalyse positive change in the agricultural landscape, ultimately benefiting communities worldwide.

1.2.Statement of the problem

The paddy farming community faces a critical challenge in the timely and accurate detection of paddy brown leaf spot (BLS) and bacterial leaf blight (BLB), two devastating diseases that threaten crop health, food security, and the livelihoods of millions of farmers, necessitating

the development of a robust, accessible, reliable, safe, trustworthy and human-centred artificial intelligence-driven solution for disease identification and management.

In addition to the chemical fertiliser challenge and the pandemic, water scarcity, unstable ground conditions, and diseases caused by fungi, bacteria, viruses, and nematodes all contributed to sluggish paddy crop growth. These diseases are recorded to be likely to occur in paddy fields, leading to epidemics and severe crop losses, causing financial and economic difficulties. Biological and non-biological factors, such as fungi, bacteria, viruses, and nematodes, have become the leading causes of paddy diseases.

Manual detection of plant diseases is costly and time-consuming, often involving experts on large farms. This method has led to a decline in rice production due to poor management. Continuous improvement is crucial to reduce pesticide use, save costs, and improve quality (Sethy et al., 2020). Accurate early conclusions can help to reduce pesticide usage.

Sometimes, the disease's appearance on the leaves is similar, so farmers cannot detect the correct disease. Usually, the symptoms are manually identified. However, manual detection is time-consuming, and the disease will likely not be seen promptly, resulting in delays in correct treatments.

An efficient and effective Rice Leaf Disease Identification System is required to address this issue. Regarding crop disease management, the Rice plant disease recognition system primarily focuses on precise and timely illness prognoses. Various machine learning and image processing techniques (Daniya and Vigneshwari, 2022) are employed to identify and categorise rice diseases.

One may compare the performance and computing requirements to those of conventional algorithms. Adopting AI-based methodologies across research, technology, and industry has resulted in more substantial evidence-based decision-making in agriculture and other sectors. As interest in the ethical dimensions of AI and machine learning has increased, emphasis has shifted to assuring the reliability of current and future activities. This emphasis reflects the awareness that sustaining trust (Toreini et al., 2020) in AI may be crucial for guaranteeing the acceptance and effective adoption of AI-powered services and products.

The increasing integration of AI and advanced analytics into company operations and the subsequent automation of decision-making raises the need for transparency in the decision-generating processes of these models. Diverse methods that generate explanations and seek to

increase user confidence in Deep Neural Network models have been proposed (Pugliese et al., 2021). Nevertheless, the effect of machine learning-generated illustrations on trusting humans for complicated decision-making tasks in industries is not fully known.

To attain this level of transparency, how might we leverage the efficiencies of AI?

This is where Explainable AI (XAI) can be of use. In this study, we are attempting to model human-centred explainable AI. By placing humans at the core of AI efforts, human-centred AI attempts to alleviate fears of existential dangers and increase benefits for users and society. Human-centred Artificial Intelligence (HCAI) is an emerging field that aims to develop AI systems that enhance and improve human abilities rather than replace them. HCAI attempts to preserve human control to ensure that artificial intelligence fulfils our needs while operating transparently, giving equitable results, and protecting privacy. Human-centred artificial intelligence learns through human input and collaboration, concentrating on algorithms inside a broader system focused on humans. Human-centred artificial intelligence is characterised by systems that continually improve due to human input and facilitate a successful interaction between humans and robots. Human-centred AI surpasses the capabilities of earlier artificial intelligence solutions by striving to comprehend human language, emotion, and behaviour through the development of machine intelligence. This approach seeks to establish a connection between machines and humans.

Well-designed technologies that provide substantial degrees of human control and computer automation will augment human performance rather than supplant them. Human-centred AI technologies are more likely to result in Reliable, Safe, and Trustworthy designs. Achieving these goals will dramatically improve human performance (Shneiderman, 2020b) while supporting human self-efficacy, mastery, creativity, and responsibility.

In HCAI, model correctness is merely one criterion by which the AI model is evaluated. Conversely, other human-centric attributes warrant our attention, including, but not limited to, interpretability, engagement, and justice. This research intends to pursue the reliability, Safety, and trustworthiness of previous research studies conducted to detect paddy leaves and address the gaps by human-centred AI.

1.3.Research Aims and Objectives

Human-centred artificial intelligence (HCAI) aims to centre AI development on humans rather than technology. Nevertheless, the extent to which current HCAI principles and

procedures achieve this objective remains uncertain. An objective is to determine whether HCAI places adequate emphasis on people.

1.1.1 Aim

The primary aim of this research is to validate the interpretability of a comprehensive and effective solution for the early detection and management of paddy brown leaf spot (BLS) and bacterial leaf blight (BLB) in rice crops through the integration of image processing techniques and human-centred artificial intelligence.

1.1.2 Objectives

1. To validate and evaluate the current image acquisition system capable of capturing high-quality images of paddy leaves for disease detection.
2. To validate and evaluate image processing algorithms that can accurately identify and differentiate symptoms of BLS and BLB in paddy leaves.
3. To evaluate trained machine learning models for automated disease classification and severity assessment, focusing on interpretability.
4. To establish confidence in AI systems, transparent rationales for their determinations are necessary. Understanding the rationale behind a particular decision increases users' propensity to trust AI.
5. To help identify and mitigate bias in AI models, ensuring that decisions are fair and unbiased, especially in critical domains.
6. Assessing the environmental and economic ramifications of disease detection system implementation concerning less pesticide application, increased agricultural output, and improved sustainability.
7. To provide recommendations and guidelines for the widespread adoption of this technology in paddy farming communities, considering factors such as scalability, affordability, and ease of maintenance.

By achieving these objectives, this research aims to offer a holistic solution that not only addresses the technical challenges of validating the interpretability of disease detection but also places a strong emphasis on usability and real-world applicability, ultimately benefiting farmers and contributing to the sustainability of paddy agriculture.

1.2 Scope

Analysing this research's core functions and goals determines the key factors and non-focused areas of primary concern.

- The proposed model will be based on existing classification algorithms. Therefore, existing HCAI models and XAI will be reviewed.
- Review existing HCAI and XAI techniques.
- To evaluate the model, object detection by image processing will be used. For that, the following boundaries will be taken.
    - The classification task is mainly based on the images of rice paddy leaf diseases because the same data set will be used during the evaluation.
    - Only selected rice paddy leaf diseases will be brought into this work. They are Bacterial blight and Brown spot.

2 LITERATURE REVIEW

This literature review exhaustively investigates the current corpus of knowledge concerning the detection of paddy BLS and BLB via the interpretability of image processing utilising AI with a human-centred focus. This approach promises to expedite the identification process and empower farmers and agricultural stakeholders, who may have limited technical expertise, with accessible disease detection and management tools.

Within the landscape of human-centred artificial intelligence research, several key themes emerge in the context of agriculture, machine learning and interpretability.

2.1 Image Processing for Disease Detection

Image processing methodologies, including feature extraction and image segmentation, are indispensable for the automated detection of disease symptoms on plant leaves.

T. Islam et al. (Islam et al., 2018) have presented a research article on a faster rice disease detection technique. According to his study, he has used green pixel masking with Naïve Bayes' classifier. He has identified that this technique can detect bacterial blight and rice brown spots with an accuracy of 89% and 90%, respectively. The present study employed the Naïve Bayes Classifier, marked by a notable drawback in its reliance on independent predictors. In its implicit assumption, Naïve Bayes considers each attribute mutually independent. However, it is unlikely that a definite set of independent predictors could be determined.

Mohd Adzhar Abdul Kahar et al. (Kahar et al., 2015) presented a study about an integrated method for recognising disease in paddy plant leaves. The research has focused on three primary paddy diseases: Bacterial Leaf Blight. He has opted for the neuro-fuzzy expert system as the recognition method. The accuracy result for the recognition is 74.21%. However, the study elaborates on issues when applying the neuro-fuzzy expert system, such as noises and other lighting problems due to external forces. This study's accuracy rate is too low to identify the Bacterial Leaf Blight on paddy leaves.

Recent developments in automation have immensely improved the identification of paddy diseases. Khaing War Htun and Chit Su Htwe (Htun and Htwe, 2018) have presented an automated system to classify the four types of paddy diseases, including bacterial leaf blight and brown leaf spots. In this study, the researchers proposed Principal Component Analysis

(PCA), Colour Grid-centred Moment, GLCM for feature extortion, and SVM for classification technologies. The system accuracy was 90% for modified grayscale conversion. The author used the Support vector machine (SVM) for classification in this study. Several critical SVM algorithm parameters must be appropriately configured to produce optimal classification outcomes for a given problem. Therefore, more reviews must be conducted on the parameters to achieve the best results.

Chowdhury Rafeed Rahman et al. (Rahman et al., 2020) illustrate using deep learning methodology to identify rice diseases and pests and propose a Convolutional Neural Network. This study's results show that the proposed architecture can achieve the desired accuracy of 93.3% with a significantly reduced model size. However, deep learning methodology contains several layers for classification. Hence, recognising the diseases will likely take longer than other diseases.

Suman and Dhruvakuma (T and T, 2015) have presented a method to classify paddy leaf diseases using shape and colour. The research applied appropriate preprocessing techniques and a histogram to classify normal and diseased leaves. In the proposed method, the shape and colour features of the diseased leaves were extracted, and the combined features of colour and shape were used to classify bacterial leaf blight, brown spot, narrow brown spot and rice blast. The researcher has used a support vector machine classifier. According to the study's findings, a 70% accuracy is achieved for four diseases. This study's accuracy rate is too low to identify the four diseases.

Shampa Sengupta et al. (Sengupta and Das, 2017) presented an article on particle Swarm optimisation-based incremental classifier design for rice disease prediction. The researchers used statistical measures and tests to establish their significance and effectiveness in this study. The results show that 84.02% accuracy is achieved for four diseases. The paper uses the Particle Swarm optimisation technique and Association Rule Mining concepts to design an incremental rule-based classification system that can be modified for better performance and accuracy.

Various methods have been explored, from traditional computer vision approaches to advanced deep learning algorithms. In conclusion, image processing for disease detection in crops has demonstrated significant potential in improving agricultural practices. Through a comprehensive approach that combines advanced image processing techniques and machine learning algorithms, studies have provided valuable insights and practical tools for early and

accurate disease diagnosis. The implementation of convolutional neural networks (CNNs) for feature extraction and classification yielded high accuracy rates, leveraging the network's ability to learn complex patterns directly from the images. Machine learning models, particularly CNNs, showed exceptional performance in classifying diseases with high precision, recall, and overall accuracy.

The above studies highlight the transformative impact of combining image processing with artificial intelligence in agriculture. By improving disease detection and providing actionable insights, this technology holds promise for enhancing crop health management and ensuring food security in the face of growing agricultural challenges.

2.2 Artificial Intelligence and Machine Learning

Support vector machines (SVMs), convolutional neural networks (CNNs), and decision trees are examples of machine learning models that have demonstrated potential in classifying and quantifying disease symptoms. The interpretability of these models is crucial for understanding their decisions and ensuring trust among users.

2.3 Human-centred Design

The adoption of a human-centred design approach ensures that the tools developed are accessible, user-friendly, and aligned with the needs of farmers and local communities. This theme explores how technology can be tailored to be intuitive and effective for non-technical users.

The study "Vision, challenges, roles and research issues of Artificial Intelligence in Education" by Yanqing Duan et al. states that artificial intelligence (AI) tries to enable computers to execute tasks by emulating intelligent human behaviours, such as inference, analysis, and decision making. (Duan et al., 2019).

A detailed examination of the various components of AI Systems reveals that each architecture and algorithm has distinct qualities, strengths, and weaknesses: Some architects function better with more data, while others do better with less. Some configurations may support unlabelled data, while others might not. In addition, various Architectures require certain Input Data. At the same time, some approaches can be integrated as ensembles, while others cannot.

Alfred Früh et al. have presented that the technical progress of AI and ML is rapid, and the state of the art is constantly evolving. Nonetheless, it is essential to have a comprehensive understanding of the present state of the art. Expanding on fundamental terms and elaborating on the particulars of this technology enable legal scholars to make credible legal statements and progress research in this sector (Früh and Haux, 2022).

A non-profit public policy organisation based in Washington, DC., the Brookings Institution's former presidents, John R. Allen and Darrell M. West, conducted a report in 2018 and presented that Artificial intelligence and data analytics are on the verge of changing numerous industries worldwide. Significant deployments in banking, national security, health care, criminal justice, transportation, and smart cities have already modified decision-making, business models, risk mitigation, and system performance. These advancements have

significant economic and social benefits. Nevertheless, AI systems development has substantial ramifications for civilisation; it affects how policy difficulties are addressed, how ethical dilemmas are resolved, how legal realities are dealt with, and how much transparency is necessary for AI and data analytics solutions. How judgments are made and integrated into organisational procedures is influenced by the decisions made by humans about software development. How precisely these operations are carried out must be better comprehended, as they will significantly impact the public soon and in the foreseeable future (West and Allen, 2018).

Human users must be able to interpret and rely on the results and output generated by machine learning algorithms to bridge the gap between AI results and human understanding. The term "explainable AI" characterises an AI model, its anticipated impact, and any potential biases.

Plamen P. Angelov at el. Illustrates in their study that artificial intelligence (AI) and machine learning (ML) have demonstrated their potential to revolutionise industries, public services, and society by achieving or even exceeding human levels of performance in terms of accuracy for a variety of problems, including image and speech recognition and language translation. However, their most effective deep learning (DL) product in terms of accuracy is frequently described as a "black box" and opaque. Indeed, such models contain a vast number of parameters that are intended to store the information learnt from training data. Not only are there many of these weights, but their connection to the problem's physical context is difficult to distinguish. This makes describing such AI extremely difficult for users. Since the applications of advanced AI and ML, including DL, are expanding rapidly throughout the digital health, legal, transportation, finance, and defence industries, the challenges of transparency and explaining ability are becoming increasingly recognised as crucial.

Figure 1 demonstrates that explaining ability is an ongoing research question for some of the most accurate types of ML, such as SVMs, DL, and many other ANNs (Angelov et al., 2021).

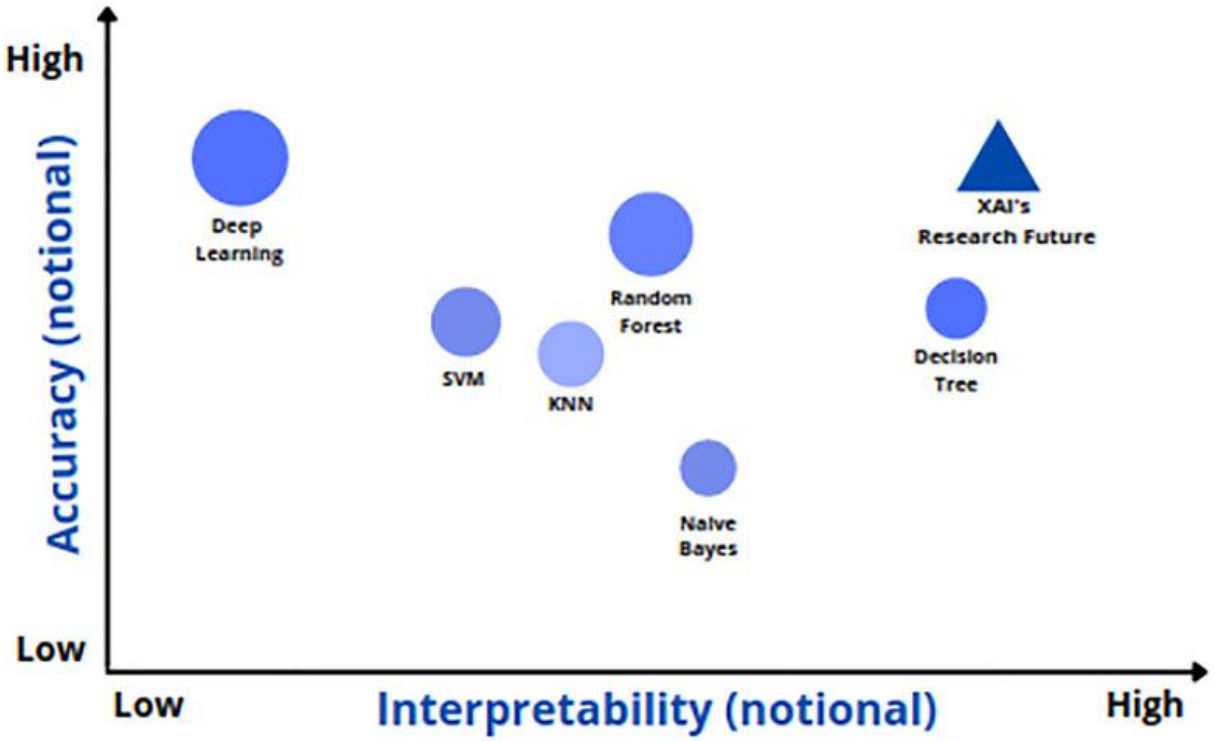


Figure 1 Accuracy versus interpretability for different machine learning models.

source: https://wires.onlinelibrary.wiley.com/

Since the beginning of AI research, scientists have argued that intelligent systems should explain AI outcomes, particularly concerning decisions. As a result of the fact that human experts establish and develop the knowledge and regulations within expert systems, these are straightforward for humans to comprehend and interpret. The decision tree is a standard method with an explicable structure (Xu et al., 2019).

The current output of Deep Neural Networks (DNNs) cannot be explained without fundamentally new explanatory methods. This is true regardless of whether one considers the neural network, an external descriptive component, or the system's developer. CNN, RNN, and LSTM are all DNNs utilising distinct architectures to address problem classes and input data. All of them must be regarded as black boxes whose internal inference processes are neither visible nor interpretable to humans (Xu et al., 2019).

The ability of a machine learning model is typically inversely proportional to its prediction accuracy; as it increases, explainability decreases. As illustrated in Figure 2, the DARPA Explainable AI (XAI) programme presents a visually appealing chart that draws attention to these noteworthy phenomena. Among the learning strategies enumerated, decision trees exhibit the highest capacity for explanation but the lowest accuracy in predictions. However,

regarding predictive ability, deep learning approaches are inferior to all other learning techniques; furthermore, they are the least likely to be explicable. (Xu et al., 2019).

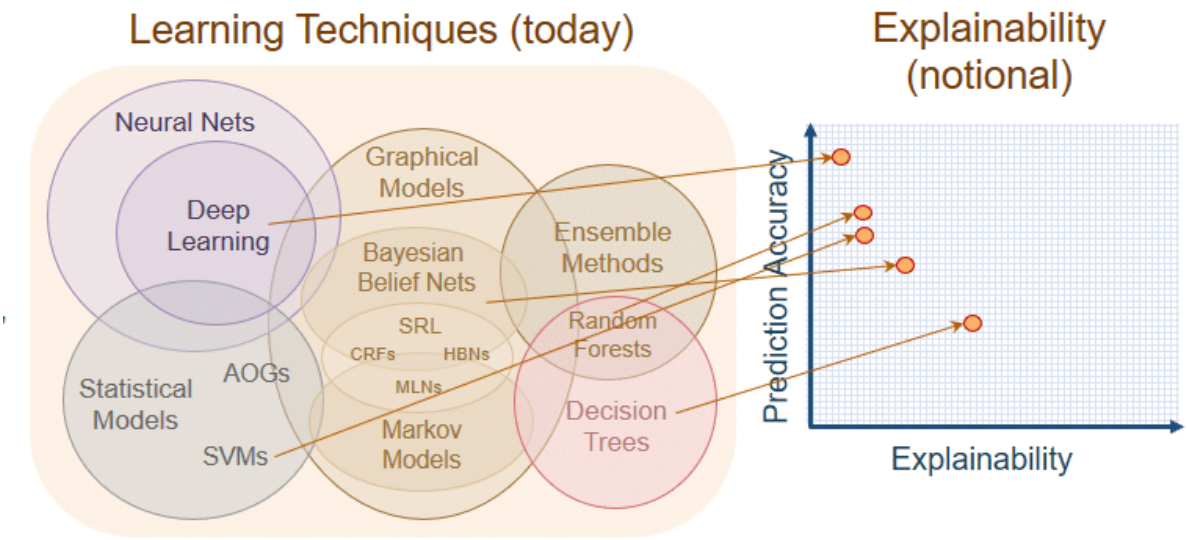


Figure 2 Explainability of machine learning models appear inverse to their prediction accuracy

source: DARPA

The Defence Advanced Research Projects Agency (DARPA) demonstrates that the explosion of Artificial Intelligence (AI) applications has resulted from the phenomenal success of machine learning. Continued advancements are expected to generate autonomous systems that can independently perceive, learn, decide, and act. However, the current inability of machines to explain their decisions and actions to human users hinders the effectiveness of these systems (DARPA, n.d.).

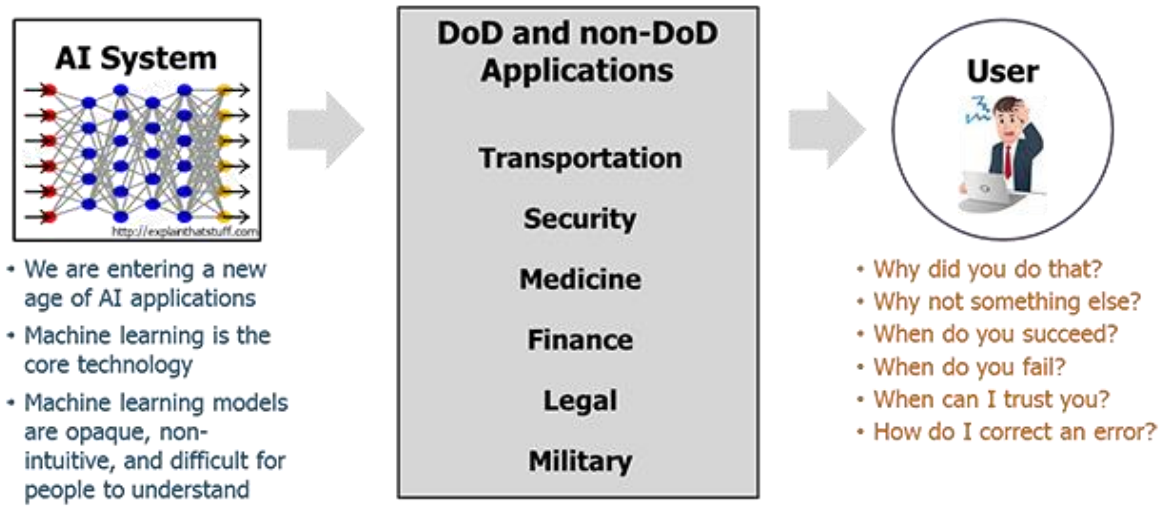


Figure 3 The Need for Explainable AI

source: DARPA

Upol Ehsan explains that AI systems' capabilities are vital for holding them accountable, as they increasingly impact our lives by driving high-stakes decisions in fields such as healthcare and law in the study "Human-Centred Explainable AI " (HCXAI): Beyond Opening the Black Box of AI". Meanwhile the study explains the understanding who interacts with the black box of AI is at least as crucial as "opening" it when it comes to Explainable AI (XAI). Nevertheless, the discourse on XAI has been predominately centred on the black box, resulting in failures to address user needs and a worsening of algorithmic opaqueness. Researchers have called for human-centred methods to XAI to solve these concerns (Ehsan et al., 2022).

Emmanuel Adjei Domfeh (Adjei Domfeh et al., 2022) presented a research article exploring the concepts of human-centred AI research and considered the various theoretical principles, theories, and paradigms. Additionally, the study investigates the multiple advancements and prospects in human-centred AI subsequent to conducting a systematic literature review across multiple online journal databases. The study screened and classified available literature into numerous categories using the PRISMA model. The paper concurs that a balance between increasing computer automation and human involvement is necessary. This is especially pertinent in the age of chatbots and other AI systems when achieving fair, just, and dependable systems. Additionally, the associated literature emphasises many applications that centre on AI, focusing on human needs. Among numerous other suggestions for future research, we propose that Shneiderman's two-dimensional human-computer autonomy be expanded. The study further suggests more commitment and attention to balancing human control over current intelligent systems. Ben Shneiderman (Shneiderman, 2020b) has illustrated the use of well-designed technologies that offer high levels of human power and that high levels of computer automation can increase human performance, leading to broader adoption.

Understanding the circumstances in which complete human or computer control is required and designing for high levels of human power and high levels of computer automation to improve human performance are both objectives of the Human-Centred Artificial Intelligence (HCAI) framework. Prevent the risks associated with overbearing human or computer control.

HCAI methods are more likely to generate Reliable, Safe, and Trustworthy (RST) designs. Accomplishing these objectives will significantly enhance human performance while fostering self-efficacy, mastery, originality, and accountability.

A study by Wei Xu et al. (Xu et al., 2019) remarked that while AI has benefited humanity, it can also be harmful if not developed properly. To uncover these obstacles, they conducted a comprehensive literature research and (Human Centred Interaction) HCI-oriented analysis of existing work in constructing AI systems. Their analysis and review shed light on the recent developments in AI technology and the challenges that HCI specialists face when attempting to implement the human-centred AI (HCAI) methodology for developing AI systems. Furthermore, the research identified seven notable obstacles in human interaction with AI systems that HCI specialists did not encounter during the development of non-AI computing systems. To aid in the execution of the HCAI strategy, the research identified novel HCI opportunities associated with distinct HCAI-driven design objectives, which HCI specialists could utilise to resolve these emerging challenges. Finally, their assessment of current HCI methodologies exposes the constraints related to their use in developing HCAI systems. Alternative strategies for overcoming these restrictions have been given in the study to aid HCI specialists in efficiently applying the HCAI methodology to developing AI systems.

Andreas Holzinger et al. (Holzinger et al., 2022) have stated in the study "Digital Transformation in Smart Farm and Forest Operations Needs Human-Centred AI: Challenges and Future Directions" that ML models frequently respond to even slight perturbations, which can have profound implications on their outcomes. Consequently, the use of AI in critical human life domains (agriculture, forestry, climate, health, etc.) has increased the need for trustworthy AI with two key characteristics: explainability and resilience. Utilising expert knowledge is a means of enhancing the robustness of AI. Consequently, human-centred artificial intelligence (HCAI) is a blend of "artificial intelligence" and "natural intelligence" designed to empower, magnify, and augment human performance as opposed to replacing people. For HCAI to attain practical success in agriculture and forestry, this article identifies three crucial frontier research topics:

1. Intelligent information fusion;
2. Robotics and embodied their evaluation intelligence;
3. Augmentation, explanation, and verification for trusted decision support.

In the paper "Designing Theory-Driven User-Centric Explainable AI," Danding Wang, Qian Yang, Abdul Ashraf, and Brian Lim present a conceptual framework for developing XAI that is human-centric and rational, drawing from a comprehensive review of relevant literature (Wang et al., 2019). By utilising this paradigm, the research establishes the channels through

which prevalent cognitive biases can be mitigated by XAI and the human mental patterns that necessitate their development.

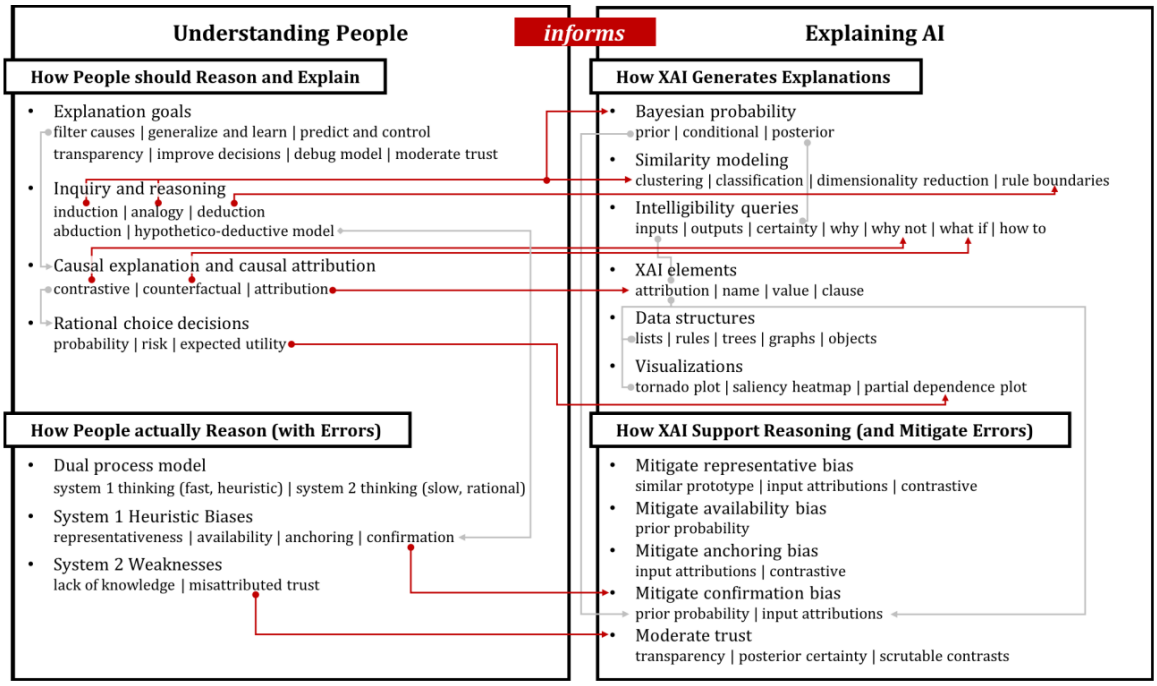


Figure 4 Conceptual framework for Reasoned Explanations that describes how human reasoning processes inform XAI techniques

A conceptual framework for Reasoned Explanations that describes how human reasoning processes (left) inform XAI techniques (right) (Wang et al., 2019).

The study done by Wang et al. leads application developers to select XAI techniques pathways connecting to human reasoning goals.

The study by Q. Vera Liao and Kush R. Varshney examines recent works in both our field and others within human-computer interaction (HCI) that adopt human-centred strategies in crafting, assessing, and offering conceptual and methodological resources for XAI. They pose the inquiry of how human-centred approaches contribute to XAI, pinpointing three key roles: guiding technical decisions based on users' demands for explainability, identifying shortcomings in current XAI techniques to inspire novel approaches, and furnishing conceptual structures for XAI that align with human needs and preferences (Liao and Varshney, 2022).

Q. Vera Liao and Kush R. Varshney created a mapping tool that links categories of user inquiries from the XAI question bank to various XAI methods capable of addressing these questions. Each technique is accompanied by a description of its output, presented in the

"Explanations" column. XAI methods are selected based on their availability in open-source XAI toolkits.

Question	Explanations	Example XAI techniques
Global how (global model-wide)	Describe the general model logic as feature impact, rules or decision trees (sometimes need to explain with a surrogate simple model) If the user is only interested in a high-level view, describe the top features or rules considered.	ProfWeight, Global Feature Importance, PDP, DT Surrogate
Why	Describe how features of the instance, or what key features, determine the model's prediction. Or describe rules that the instance fits to guarantee the prediction. Or show examples with the same predicted outcome to justify the model's prediction.	LIME, SHAP, LOCO, Anchors, ProtoDash
Why not (a different prediction)	Describe what features of the instance determine the current prediction and with what changes the instance would get the alternative prediction. Or show prototypical examples that had the alternative outcome.	CEM, Counterfactuals, ProtoDash (on alternative prediction)
How to be that (a different prediction)	Highlight feature(s) that, if changed (increased, decreased, absent, or present), could alter the prediction to the alternative outcome, often requiring minimum effort. Or show examples with minimum differences but had the alternative outcome.	CEM, Counterfactuals, DiCE
How to still be this (the current prediction)	Describe features/feature ranges or rules that could guarantee the exact prediction. Show examples that differ from the particular instance but still have the same outcome.	CEM, Anchors

What if	Show how the prediction changes corresponding to the inquired change of input.	PDP, ALE
Performance	Provide performance metrics of the model. Show uncertainty information for each prediction. Describe the potential strengths and limitations of the model.	Precision, Recall, Accuracy, F1, AUC, Uncertainty Quantification 360, FactSheets, Model Cards
Data	Document comprehensive information about the training data, including the source, provenance, type, size, coverage of population, potential biases, etc.	FactSheets, DataSheets
Output	Describe the scope of output or system functions. Suggest how the output should be used for downstream tasks or user workflow.	FactSheets, Model Cards

Table 1 Question, Explanations and Example XAI techniques

2.4 Validation and Field Testing

Rigorous field testing assesses the real-world applicability of disease detection systems. These studies consider environmental variability, lighting conditions, and data collection challenges.

Explanation of the model is incorporated as a critical component of the machine-learning pipeline. The option of maintaining a machine learning model as a "black box" has been eliminated. The way these models or applications process to yield the results is still a mystery to both the developer and the user. Validating these kinds of systems is a challenge. This lack of interpretability of AI models and applications makes them hard to trust.

The following researchers have developed several techniques to mitigate the validation issue.

2.4.1 Local Interpretable Model-agnostic Explanations (*LIME*)

The LIME approach is precisely engineered to furnish human-comprehensible and interpretable explanations for the predictions generated by intricate machine learning models, regardless of the underlying model architecture.

LIME was presented in the paper *"Why Should I Trust You?": Explaining the Predictions of Any Classifier* and aims to explain any black box model by creating such a local approximation the complex models are complete black boxes, and the internals are hidden for LIME (Ribeiro et al., 2016). So, it's just based on the inputs and outputs of a model it works on. Almost any input, such as text, tabular data, images or graphs. Usually, the domain experts in a particular field, medicine, have some prior knowledge about the problem. For example, sports have a positive impact on overall health. If the lime explanation tells us that sports increase the probability of a stroke, there's most likely something wrong in our developed model, which helps to build trust, and we can assess whether it makes sense. The paper also states that providing explanations improves the acceptance of a predictive algorithm for LIME. The only requirement is that the explanations are locally faithful, but they might not make sense globally, so we focus on that local area around our prediction.

Critical characteristics of LIME include:

- LIME is model-agnostic, meaning it can be applied to any machine learning model regardless of its type (e.g., decision trees, support vector machines, neural networks). This flexibility allows users to interpret the predictions of even the most complex models.
- LIME focuses on explaining individual predictions at a local level. Instead of explaining the entire global behaviour of the model, it approximates the model's decision-making process near a specific data point of interest.
- LIME generates local interpretations by perturbing the features of a given instance and sampling from the perturbed data. By creating variations of the input data, LIME aims to understand how feature changes impact the model's predictions.
- LIME constructs interpretable surrogate models (often more straightforward and transparent than the original) based on the angry and sampled data. These surrogate models approximate the complex model's local decision boundary.
- LIME assigns weights to the sampled instances based on their proximity to the original instance. Instances closer to the original point receive higher weights, reflecting their importance in approximating the local decision function.
- The output of LIME is a set of weighted features, indicating their contribution to the prediction for the specific instance. This information is often visualised in ways humans can interpret, such as a bar chart or heatmap.
- LIME is commonly used in various applications, including image classification,

natural language processing, and any scenario where the interpretability of individual predictions is crucial. It has been applied to improve understanding and trust in machine learning models' predictions.

LIME has proven to be a valuable tool for enhancing the interpretability of complex models in real-world applications. It provides users, especially non-experts, with insights into why a model made a specific prediction for a given instance, contributing to increased transparency and trust in machine learning systems.

2.4.2 SHapley Additive Explanations (SHAP)

The paper "*A Unified Approach to Interpreting Model Predictions*", authored by Scott Lundberg and Su-In Lee (Lundberg and Lee, 2017), presents a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations).

The SHAP method explains the output of any machine learning model. It derives its name from the Shapley values in cooperative game theory and offers a technique for equitably allocating the impact of each feature on the prediction. SHAP originated from the desire to bring a consistent and fair approach to attributing a model's prediction across its input features.

Critical Characteristics of SHAP:

- SHAP applies to any machine learning model, including support vector machines, decision trees, neural networks, and other algorithms, as it is model-agnostic.
- SHAP values provide individualised explanations for each prediction, breaking down the model's output into the contribution of each feature for a specific instance.
- SHAP values satisfy consistency and fairness properties, ensuring that the sum of the feature contributions equals the difference between the model's prediction for the instance and the average prediction for all the cases.
- SHAP values are based on Shapley values from cooperative game theory. They represent the average contribution of a feature across all possible combinations, considering all possible orders in which features could be added to the model.
- SHAP values adhere to the additivity property, meaning the sum of the SHAP values for all features equals the difference between the model's prediction for a specific instance and the average prediction for all the cases.

By employing SHAP, users gain insights into how each feature influences model predictions, fostering transparency and aiding the understanding and trustworthiness of complex machine learning models.

2.4.3 Counterfactual explanations and adversarial attacks

Counterfactual explanations and adversarial attacks are two concepts often discussed in the context of machine learning and artificial intelligence. They represent two different aspects of model behaviour and security. This approach was presented in the paper “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR”, illustrated by Sandra Wachter, Brent Mittelstadt and Chris Russell (Wachter et al., 2018). In this paper, they have introduced the notion of unconditional counterfactual explanations as an innovative form of elucidating automated decisions, addressing numerous hurdles encountered in existing efforts towards algorithmic interpretability and accountability.

Counterfactual Explanations

Counterfactual explanations offer valuable insights into the hypothetical modifications to a model's prediction, assuming specific input features remained constant while altering others. It answers, "What changes would need to be made to the input for the model's prediction to change?".

Key Points

- **Interpretability**: Counterfactual explanations enhance the interpretability of machine learning models by showing users what features are critical for a particular prediction.
- **Applications**: They are used in various domains, including finance (e.g., loan approval), healthcare (e.g., diagnosis), and recommendation systems, to provide transparent and actionable explanations for model decisions.
- **Use Cases**: For example, in a loan approval scenario, a counterfactual explanation could show which features (e.g., income, credit score) would need to change for a rejected applicant to be approved.

Adversarial Attacks

Definition: Adversarial attacks involve intentionally perturbing input data so that a machine-learning model makes incorrect predictions. These perturbations are often imperceptible to humans but can significantly alter the model's behaviour.

Key Points

- **Vulnerability**: Adversarial attacks highlight vulnerabilities in machine learning models, showing that even state-of-the-art models can be susceptible to manipulation.
- **Security Concerns**: They raise security concerns, especially in critical applications like autonomous vehicles, where a malicious actor could intentionally deceive the model to cause accidents.
- **Defence Mechanisms**: Researchers develop defence mechanisms against adversarial attacks, such as negative training, input preprocessing, and robust optimisation techniques, to enhance model resilience.

While counterfactual explanations focus on understanding model behaviour and providing transparent insights to users, adversarial attacks exploit weaknesses in models to deceive them. However, both concepts contribute to the broader understanding of model behaviour and security in machine learning and artificial intelligence.

2.4.4 Layer-Wise Relevance Propagation (LRP)

LRP is an explanation technique used in eXplainable Artificial Intelligence (XAI) to understand the decision-making process of deep neural networks (DNNs). Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller authored the paper titled "*Layer-Wise Relevance Propagation: An Overview*," wherein they provide a brief overview of LRP. They discuss the ease and efficiency of implementing propagation rules, the theoretical justification of the propagation procedure as a 'deep Taylor decomposition,' strategies for selecting propagation rules to ensure high-quality explanations at each layer, and the adaptability of LRP to various machine learning scenarios beyond deep neural networks (Montavon et al., 2019). LRP aims to attribute the model's output to its input features, providing insights into which features are most relevant for a given prediction.

Critical Components of Layer-Wise Relevance Propagation

- LRP defines propagation rules that specify how relevance scores are propagated backwards through the network layers. These rules determine how much relevance from a neuron in the output layer is attributed to neurons in the preceding layers.
- Given a prediction made by the neural network, LRP calculates relevance scores for each neuron in the network. The relevance scores indicate the individual contributions of each neuron towards the ultimate prediction.

- LRP redistributes relevance scores from the output layer back to the input layer, passing through intermediate layers. This redistribution process provides insights into the importance of different features at each network layer.
- By propagating relevance scores layer by layer, LRP offers layer-wise interpretability, allowing users to understand how features at different layers contribute to the model's decision.

Overall, Layer-wise Relevance Propagation is a powerful XAI technique that offers detailed insights into the decision-making process of deep neural networks, enabling users to understand and trust the predictions made by these complex models.

2.5 Impact on Agriculture

Research often investigates the economic, ecological, and social impacts of disease detection technology, including the potential reduction in pesticide usage, increased crop yields, and enhanced sustainability.

The principal aim of this literature review is to consolidate current understanding, identify deficiencies, and furnish an all-encompassing synopsis of the research landscape concerning paddy BLS and BLB detection. By examining previous work, we aim to lay the foundation for our research and contribute to developing more effective, user-centric, and sustainable solutions for disease management in paddy crops.

3 METHODOLOGY

The study will be conducted as Design Science Research (DSR). DSR is used to develop foundational knowledge about the design of artefacts such as software, methodologies, models, and concepts. Hence, the research will be conducted on DSR principles.

Design Science Research (DSR) is a methodology that focuses on creating and evaluating innovative solutions to real-world problems. The research methodology can be structured as follows when applying DSR to developing eXplainable Artificial Intelligence (XAI) to detect paddy brown leaf spots and bacterial leaf blight.

As mentioned in the objectives of this research, the problem is to provide interpretability on using HCAI in intelligent agriculture and model to get maximum yield and detect diseases in a Human-centric environment. The scope of the agriculture is as follows: research will be conducted on two central paddy leaf diseases: Bacterial Blight and Brown Spot. To ensure that this knowledge is elaborated understandably, scientific methods are used in the DSR.

3.1 Overview

This study defines the problem as the need for accurate and user-friendly disease detection in paddy crops. Understand the specific challenges and requirements of farmers, agricultural stakeholders, and the agricultural context.

The literature helps develop theories and uncover research gaps. In addition, it helps DSR researchers comprehend the problem space by locating existing literature and flagging unexplored areas. Since that research will be conducting a literature evaluation, the literature will be shown on Paddy Leaf disease detection methods and their evaluation as Human-centred Artificial Intelligence.

If the interviewees are part of the problem's stakeholder group, interviews can be used to establish requirements for the solution space with a solid basis in the problem area. So, a few interview sessions will be held with the “National Institute of Plantation Management” (NIPM) members.

3.2 Design and Conceptualization

Develop a conceptual framework for the Human Centred eXplainable Artificial Intelligence (HC-XAI) system, outlining the key components and design principles. Identify the system's essential features, such as interpretability, user-friendliness, and real-time disease detection.

Implementing eXplainable Artificial Intelligence (XAI) in the context of "The Detection of Paddy Brown Leaf Spot and Bacterial Leaf Blight" involves combining techniques and approaches that prioritise human understanding, interpretability, and user-centred design. Below is a methodology for developing XAI systems for this specific application:

3.2.1 Data Collection and Preprocessing

Gather a diverse dataset of images of paddy leaves, including those affected by brown leaf spot and bacterial leaf blight. Annotate the dataset to label the presence and severity of diseases. Preprocess the images, removing noise, enhancing contrast, and standardising image sizes.

3.2.2 Feature Extraction

Use image processing techniques to extract relevant features from the images. These features may include colour, texture, shape, and lesion patterns. Apply strategies such as Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and colour histograms to capture discriminative characteristics.

3.2.3 Model Selection

Choose machine learning or deep learning models suitable for image classification. Convolutional Neural Networks (CNNs) are often effective for image analysis tasks. Train models to classify images: healthy, brown leaf spot, or bacterial leaf blight.

Time and resources can be saved, and the three sections mentioned above can be omitted by getting a previously developed solution documented in the literature. The study aims to validate the interpretability of detecting and managing paddy brown leaf spot (BLS) and bacterial leaf blight (BLB) in rice crops by integrating image processing techniques and human-centred artificial intelligence. Since then, there hasn’t been system development. The study conducted by Junaid Iqbal, Israr Hussain, Ayesha Hakim, and Sami Ullah on "Early Detection and Classification of Rice Brown Spot and Bacterial Blight Diseases Using Digital Image Processing" the research system aims to detect symptoms of diseases on rice leaves. After identifying these disorders, they are initially classified through image processing. The

process involves capturing multiple images of both healthy and diseased leaves. Furthermore, features are extracted following image preprocessing. Subsequently, the rice leaf images are categorised as either beneficial or diseased. When there is an infection, the system accurately recognises and classifies it. In the study, classifiers such as Inception V3 and VGG19 were employed, with VGG19 demonstrating superior performance, achieving an accuracy of 97.94% (Iqbal et al., 2023).

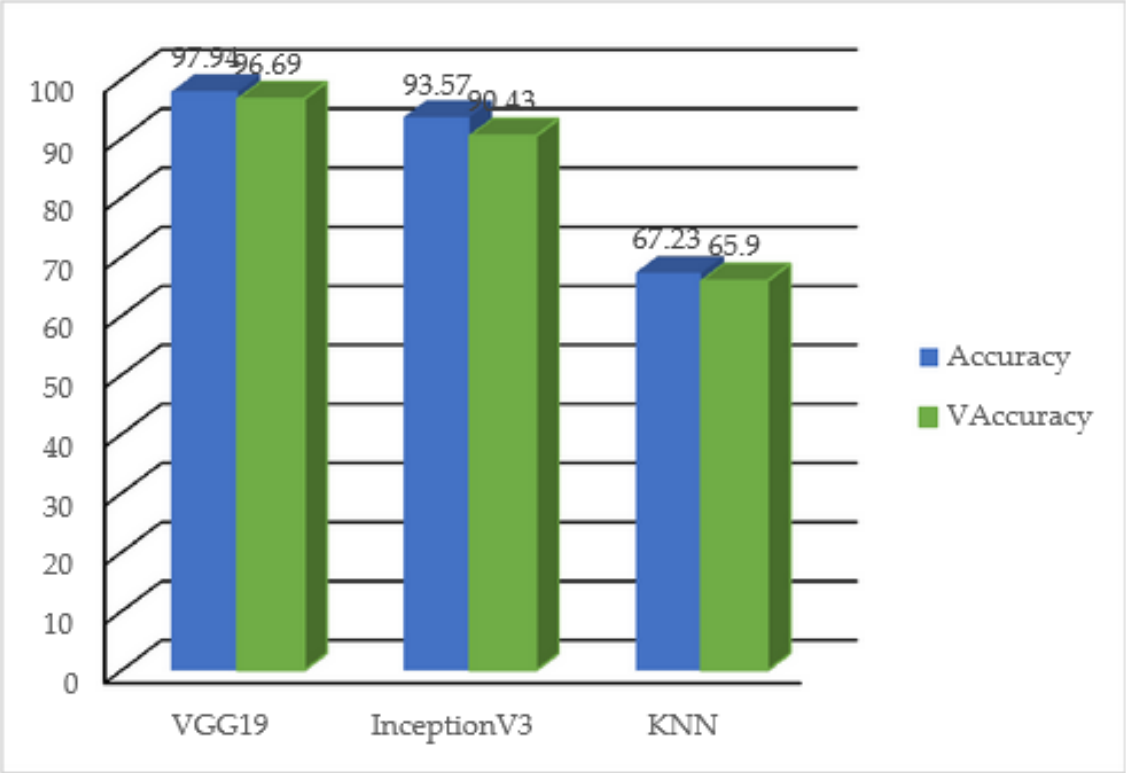


Figure 5 Accuracy and the Validation Accuracy of the study 1

	VGG19	**InceptionV3**	**KNN**
Accuracy	97.94%	93.57%	97.23%
V Accuracy	96.69%	90.43%	65.9%

A YOLOv5 model trained by the study's author will be used for comparison. This study achieved 97.6% accuracy in identifying Bacterial Blight and Brown Spot diseases in PaddyLeaf. The above models will be mentioned in the study.

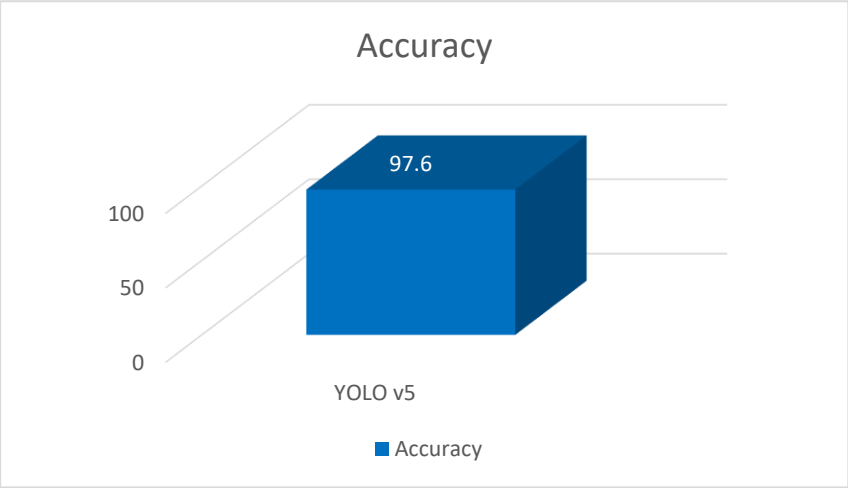


Figure 6 Accuracy of the study 2

	YOLO v5
Accuracy	97.6%

Model 01: Early Detection and Classification of Rice Brown Spot and Bacterial Blight Diseases Using Digital Image Processing (Iqbal et al., 2023).

Model 02: Authors trained the model by YOLO v5.

3.2.4 Interpretability Techniques

Incorporate XAI methodologies, such as Local Interpretable Model-agnostic Explanations (LIME), Shapley Additive Explanations (SHAP), Diverse Counterfactual Explanations (DiCE), and Layer-wise Relevance Propagation, to enhance the interpretability of models (LRP). Produce saliency maps or heat maps that emphasise the visual portions that impact the model's determination most.

Let’s discuss how each XAI tool performs on the two studies selected. In explainable artificial intelligence (XAI), various techniques have been developed to provide interpretable insights into the decision-making processes of machine learning models, particularly in image analysis. Methodologies such as Layer-wise Relevance Propagation (LRP), Local Interpretable Model-agnostic Explanations (LIME), and Counterfactual explanations and adversarial attacks (DiCE) are among those that are designed explicitly for image-based explanations. These methods aim to elucidate the contributions of different image regions or features to the model's predictions, facilitating a deeper understanding of how these models perceive and classify visual information. Through the application of these image-specific XAI technologies, stakeholders can gain transparent and actionable insights into the functioning of complex neural networks, enhancing trust, interpretability, and usability in various image-centric applications ranging from medical diagnosis to autonomous driving systems since that study has conducted the validation on those three technologies.

3.2.5 Local Interpretable Model-agnostic Explanations (LIME)

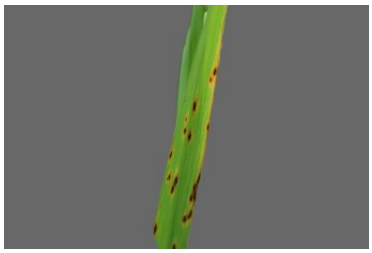
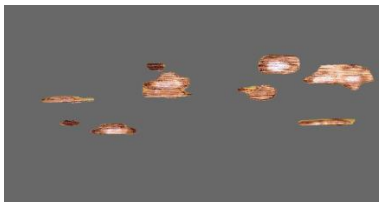
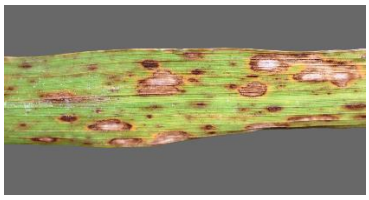
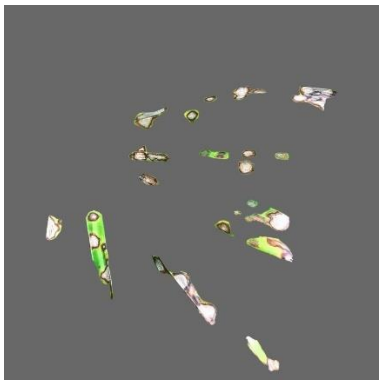
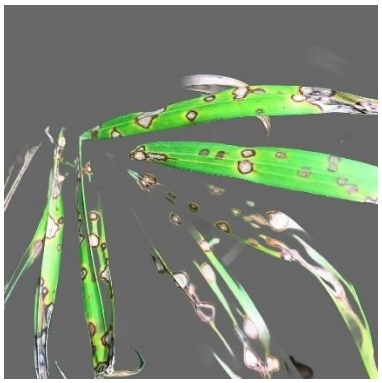
Original image	Model 01	Model 02
		
	Explaining brown_spot = True	Explaining brown_spot = True
		
	Explaining brown_spot = True	Explaining brown_spot = True
		
	Explaining brown_spot = True	Explaining brown_spot = True

Table 2 LIME on Brown spot disease detection

Observation: Both models can identify brown spot disease, even if the leaf is not a paddy leaf (third image). However, both models fail to recognise the leaf category.

	brown_spot		
	Image 1	**Image 2**	**Image 3 (Error)**
Model 1	TRUE	TRUE	TRUE
Model 2	TRUE	TRUE	TRUE

Original image	Model 01	Model 02
		
	Explaining bacterial_leaf_blight = True	Explaining bacterial_leaf_blight = True
		
	Explaining bacterial_leaf_blight = True	Explaining bacterial_leaf_blight = True
		
	No Detection	Explaining bacterial_leaf_blight = True

Table 3 LIME on Bacterial leaf blight disease detection

Observation: Both models are capable of identifying the bacterial leaf blight disease. However, Model 2, even though it is not a paddy leaf (third image), identifies it as a disease.

	bacterial_leaf_blight		
	Image 1	**Image 2**	**Image 3 (Error)**
Model 1	TRUE	TRUE	FALSE
Model 2	TRUE	TRUE	TRUE

3.2.6 Diverse Counterfactual Explanations (DiCE)

Counterfactual explanations and adversarial attacks

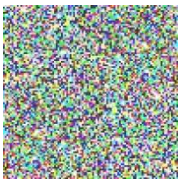
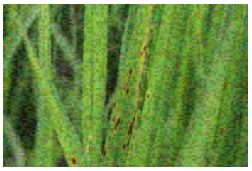
Original Image	Noise	Model 01	Model 02
			
		No Detection	No Detection
			
		Brown spot	No Detection
			
		No Detection	No Detection

Table 4 DiCE on Brown spot disease detection

Observation: Both models cannot identify brown spot disease when noise is added to the image. Data image clarity and visibility are essential to identifying the disease. Both models fail to recognise the disease category. Model 1 has a slight improvement in identifying the disease due to the higher zoom level of the image.

	brown_spot		
	Image 1	**Image 2**	**Image 3 (Error)**
Model 1	FALSE	TRUE	FALSE
Model 2	FALSE	FALSE	FALSE

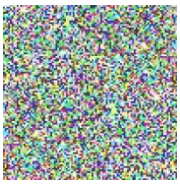
Original Image	Noise	Model 01	Model 02
			
		bacterial_leaf_blight	No Detection
			
		bacterial_leaf_blight	No Detection
			
		No Detection	No Detection

Table 5 DiCE on Bacterial leaf blight disease detection

Observation: In identifying bacterial leaf blight disease, Model 1 can identify the disease in every scenario. Even in the wrong leaf category (third image). Model 2 has poor identification of the XAI method.

	bacterial_leaf_blight		
	Image 1	**Image 2**	**Image 3 (Error)**
Model 1	TRUE	TRUE	FALSE
Model 2	FALSE	FALSE	FALSE

3.2.7 Layer-wise Relevance Propagation (LRP)

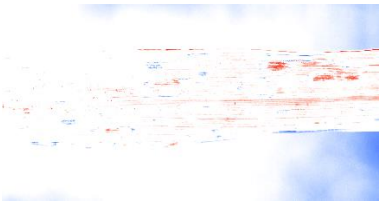
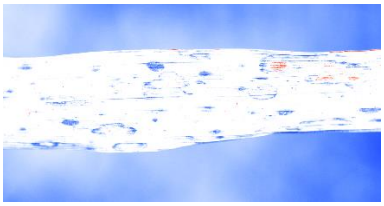
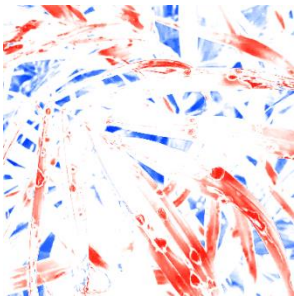
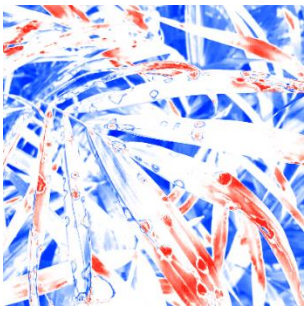
Original Image	Model 1	Model 2
		
	Brown spot	Brown spot
		
	Brown spot	Brown spot
		
	Brown spot	Brown spot

Table 6 LRP on Brown spot disease detection

Observation: As presented in the heat map, Model 1 uses the disease spot to classify the disease. However, Model 2 also considers the background. As mentioned in the LIME scenario, neither model can categorise the leaf type. Because the third scenario does not involve a paddy leaf, both models detected the disease.

	brown_spot		
	Image 1	**Image 2**	**Image 3 (Error)**
Model 1	TRUE	TRUE	TRUE
Model 2	TRUE	TRUE	TRUE

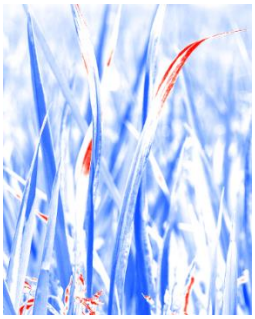
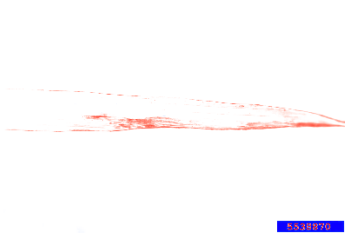
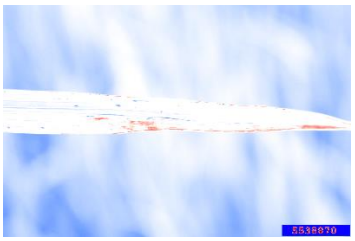
Original Image	Model 1	Model 2
		
	bacterial_leaf_blight	bacterial_leaf_blight
		
	bacterial_leaf_blight	bacterial_leaf_blight
		
	No Detection	bacterial_leaf_blight

Table 7 LRP on Bacterial leaf blight disease detection

Observation: As mentioned above, the same scenario is reflected here, but Model 1 neglected to detect the disease.

	bacterial_leaf_blight		
	Image 1	**Image 2**	**Image 3 (Error)**
Model 1	TRUE	TRUE	FALSE
Model 2	TRUE	TRUE	TRUE

3.3 Human-Centred Design

So far, the study has proved that completed trained models have high accuracy. However, they lack trustworthiness, and these models are not human-centred in the real world. In addition, the work that has been done is all post-hoc explainability. This study aims to design a framework that integrates the XAI system and ensures it is user-friendly and accessible to domain experts or agricultural stakeholders.

The study conducted by Danding Wang, Qian Yang, Abdul Ashraf, and Brian Lim, which was mentioned in the Literature, explains how to connect human reasoning theories to XAI techniques. The study can infer what types of explanations users need by understanding how users reason. According to the above research, it has been described that the contrastive and counterfactual, which come under casual explanation and causal attribution, mapped to XAI intangible queries; why not and what if, respectively (Wang et al., 2019).

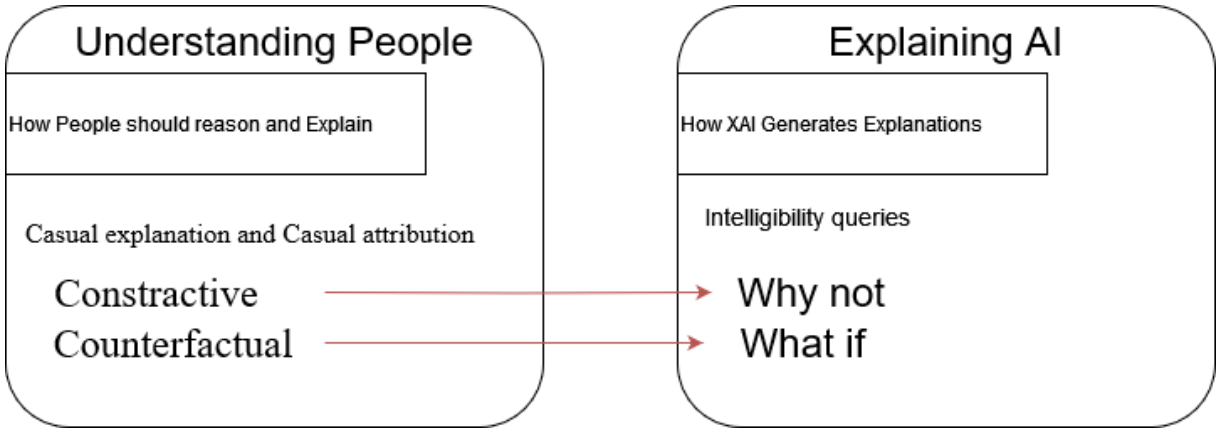


Figure 7 Extracted part of the Framework

Q. Vera Liao and Kush R. Varshney created a mapping tool that links categories of user inquiries from the XAI question bank to various XAI methods capable of addressing these questions. From the mapping tool that links categories of user inquiries from the XAI question created by Q. Vera Liao and Kush R. Varshney, we can identify the necessary XAI methods capable of addressing the above questions (Liao and Varshney, 2022).

3.3.1 Why not

When someone asks "Why not this output has come?" in explaining AI, they typically seek clarification or justification for the result produced by an AI system. This question suggests that the user expected a different outcome or is confused about why the AI generated a specific output. In response to this question, an explanation of the AI's decision-making process and the factors that influenced the output is necessary. This might involve the following attributes.

Model Explanation

Providing insights into how the AI model operates, including its architecture, training data, and algorithms used. Explaining which features the model considered most important in making its decision can explain why a particular output was generated.

Input Analysis

Examining the input data or user input that the AI processed to understand any biases, noise, or missing information that could have influenced the output. This helps identify potential reasons why the expected output did not occur.

Confidence and Uncertainty

Discussing the level of confidence or uncertainty associated with the AI's output. If the AI is unsure about its decision, it may produce unexpected results, which can be explained to the user.

Error Analysis

Explain any errors or limitations in the AI system that may have led to the unexpected output. These could include issues such as data quality, model complexity, or the inherent limitations of AI technology.

Feedback and Improvement

Encouraging feedback from the user to improve the AI system's performance in the future. Understanding user expectations and concerns can help refine the AI model and enhance its accuracy and relevance.

Overall, addressing the question, "Why does this output come?" requires a detailed explanation of the AI's decision-making process, input data analysis, confidence levels, potential errors, and opportunities for improvement. By providing clear and transparent explanations, AI designers can better understand the AI system and its outputs.

3.3.2 What if

When someone asks, "What if this output has come?" in explaining AI, they are typically interested in understanding the potential consequences or implications of a different output generated by the AI system. This question suggests that the user is exploring hypothetical scenarios and seeking insights into how alternative outcomes could impact the situation. In response to this question, an explanation of the possible implications of a different output can be provided. This might involve the following reasons.

Alternative Scenarios

Discuss different possible outputs the AI could have generated and explain how each would have affected the decision-making process or outcome.

Risk Analysis

Evaluating the risks associated with the alternative outputs, including potential benefits and drawbacks. This involves considering accuracy, fairness, ethical considerations, and stakeholder impact.

Sensitivity Analysis

Examining how sensitive the overall outcome is to AI output variations. This helps assess the robustness of the decision-making process and identify potential areas of uncertainty or instability.

Decision-Making Framework

Explain the decision-making framework used by the AI system and how it weighs different factors to generate outputs. This provides insights into why specific outputs are favoured and helps users understand the rationale behind the AI's decisions.

Mitigation Strategies

Discuss strategies for mitigating the potential consequences of alternative outputs, such as adjusting input data, refining the AI model, or incorporating human oversight into the decision-making process.

In summary, the question "What if this output has come?" in Explaining AI involves exploring hypothetical scenarios, evaluating potential risks and benefits, and providing insights into how alternative outputs could impact the situation. AI designers and developers can benefit from this discourse by acquiring a more profound comprehension of the AI system's decision-making mechanism and the ramifications that this has on the given task.

The utilisation of "Why not" and "What if" in AI explanations clarifies and provides a more profound comprehension of the decision-making mechanism and possible consequences of the AI system.

“Why not” suggests that the user expected a different output from the AI system and wants to understand the reasons behind the discrepancy. An explanation of the factors influencing the AI's decision-making process is necessary. This might involve discussing the input data, model architecture, training process, biases, and uncertainties contributing to the generated output. By providing transparency into the AI's decision-making mechanisms, users can understand why the expected output did not occur and how to interpret the results.

“What if” explores hypothetical scenarios and alternative outcomes that the AI system could have generated. Users may be interested in understanding different outputs' potential consequences and implications. Responding to this question involves discussing various possible outputs, evaluating their risks and benefits, and considering how they would impact the decision-making process or outcome. By engaging in this analysis, users can better understand the AI system's capabilities, limitations, and potential implications for the task at hand.

In essence, answering the inquiries "Why not?" and "What if?" necessitates presenting clear and comprehensive elucidations of the decision-making mechanism employed by the AI, encompassing crucial determinants that impact the result and possible alternate situations. By fostering a dialogue around these questions, designers and ML developers can better understand the AI system's behaviour and its implications for their specific use case or application.

However, these tools can still be applied at the end of the AI model's lifecycle. Can we use an AI's explainability at the beginning of a training model? This study, incorporating the above studies, introduces the following framework to answer that.

3.4 Proposed framework

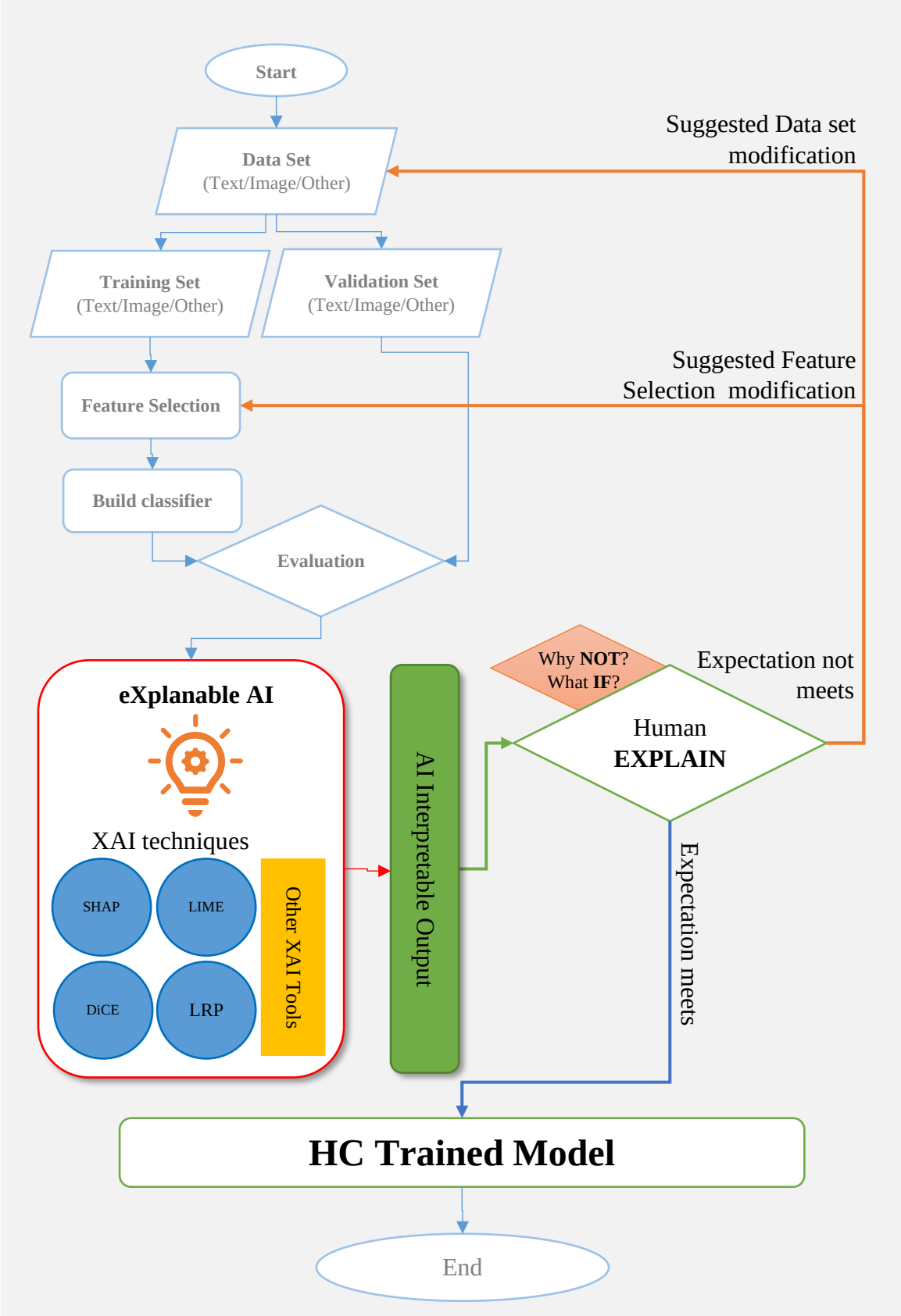


Figure 8 Proposed Framework

Generally, the training model should be an incremental process and not a one-time training. After the evaluation, an interpretable output can be derived through an explainable AI. A person can verify that interpretable output. As a result of this new step, Reliability, Safety, & Trustworthiness can be included in the trained model.

Cloud computing has evolved, and on-demand services have been expanded to training AI models. Leveraging this advancement of technology, training a small model reduces the cost of operation of computing power and time.

Four individuals are needed in four separate stages for this framework. First, we identify the four individuals.

- **Designer**: A person who gathers equipment and expectations for the AI system.
- **Developer**: The person who develops the model (Person can be a data scientist).
- **Domain expert**: A person who knows the Domain that AI is building. (Agriculture, construction)
- **User**: End user who users the system.

Below are the four stages involving the designer, developer, domain expert, and user.

Stage 1 - **Data collection**

Stage 2 - **Performance of the model**

Stage 3 - **eXplainability of the model**

Stage 4 - **End-user experience**

The following is the mapping of the human factor at each stage. The proposed framework is illustrated in the diagram below.

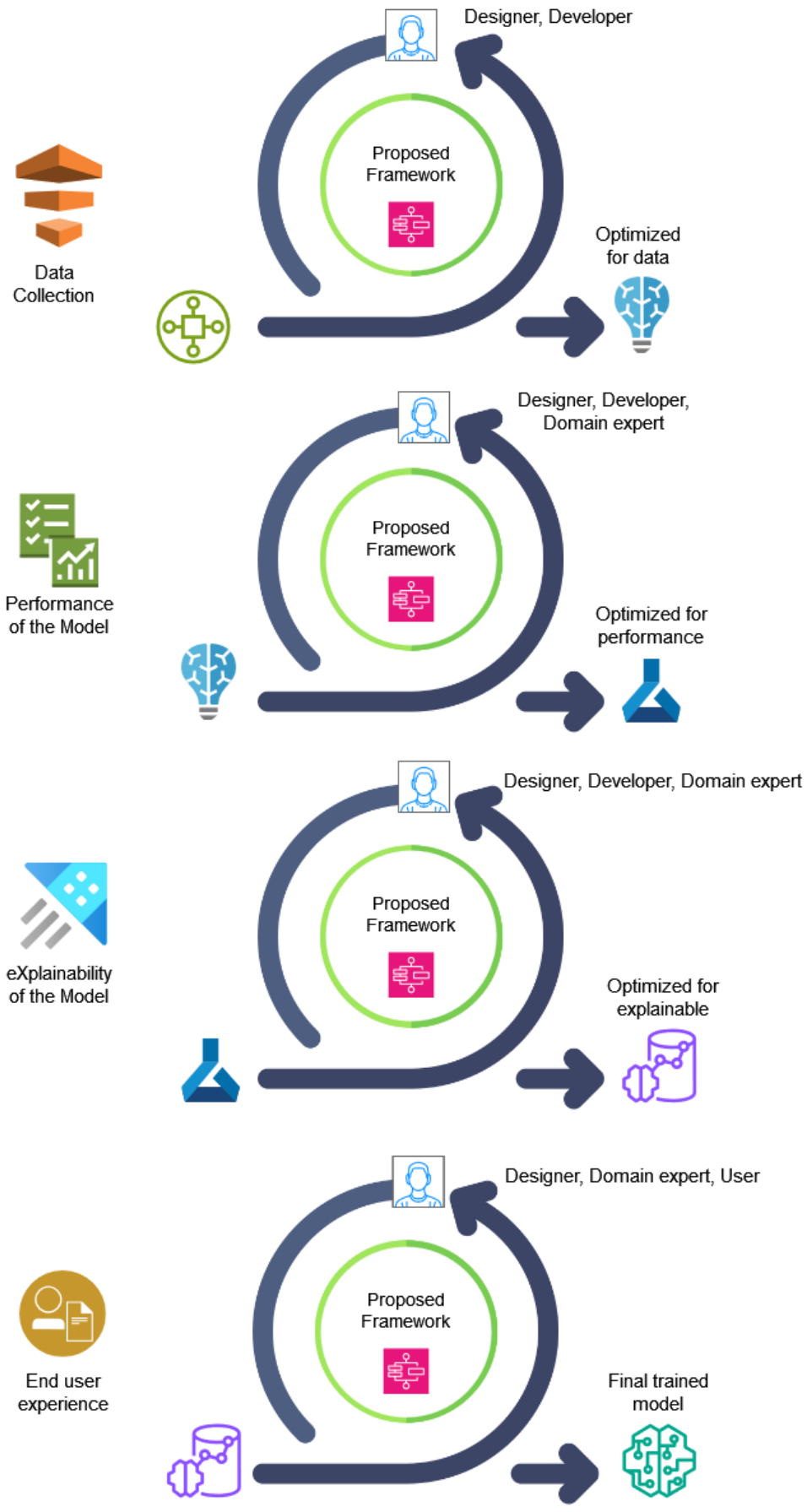


Figure 9 illustration of the proposed framework

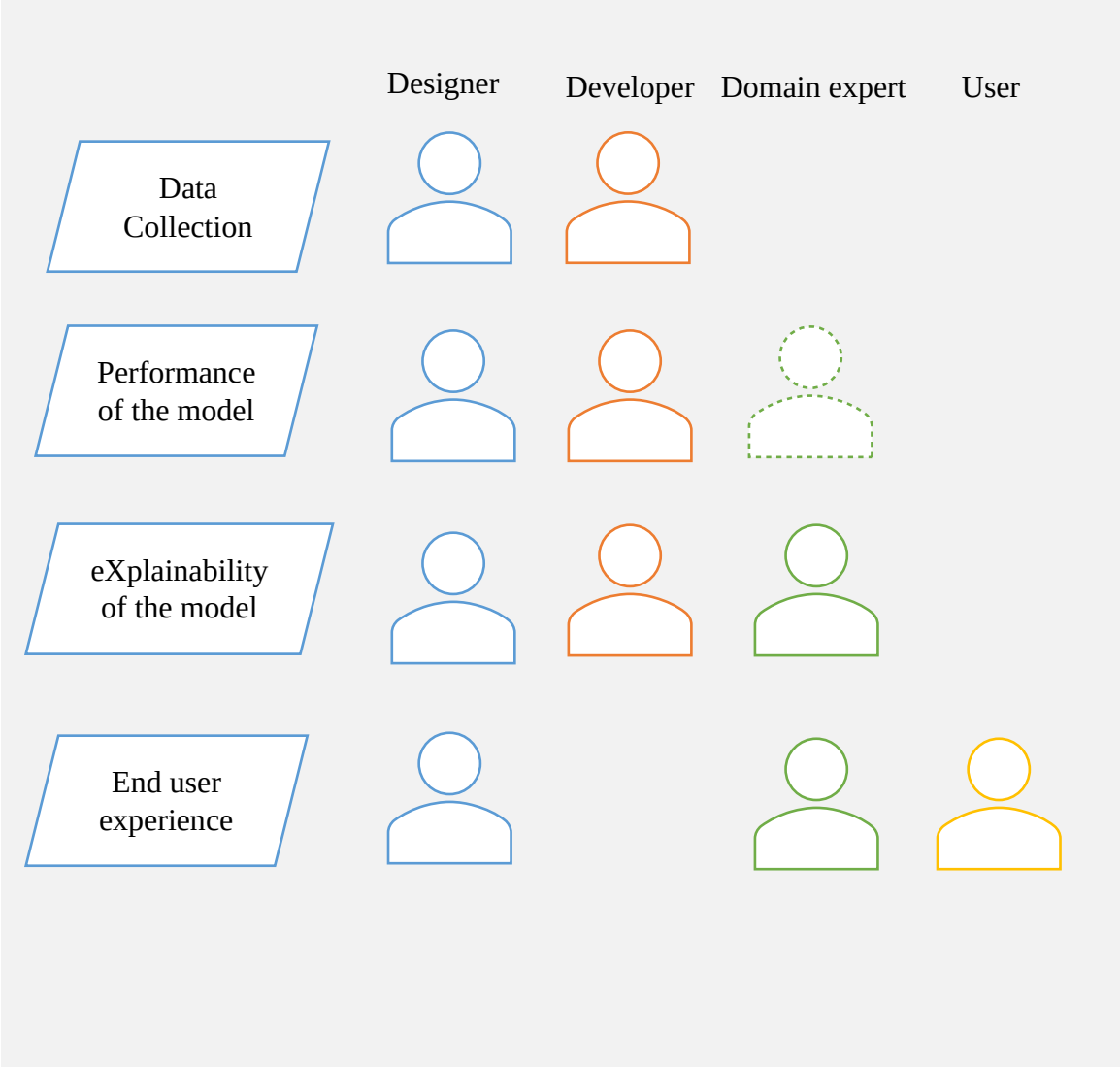


Figure 10 Stages mapped to the human interaction

As per the above illustration, the Designer and Developer engage in data collection as the first step of the proposed framework. In the second stage, which is the model's performance, the Designer and Developer take responsibility for it, and the Domain expert also takes part in the model's performance if needed at this stage. The model's eXplainability is the third stage of the proposed framework. Designer, Developer and Domain expert are utilised at this stage. The final stage of the proposed framework is the End user experience. The designer, Domain expert and User are involved in this last stage. With a stage, many iterations of the training model will be trained.

3.4.1 Validation and Field Testing

Validate the HC XAI framework on pilot training models. Evaluate the accuracy of and the usability of the HC XAI interface in a real-world context.

This methodology emphasises an approach to HC XAI for disease detection in paddy crops. It combines advanced AI techniques with user-centred design principles to provide interpretable and actionable information to the end-users, contributing to the overall goal of agricultural sustainability and food security.

3.5 Reflection and Learning

This DSR methodology for HC XAI in the context of paddy brown leaf spot and bacterial leaf blight detection emphasises the development of an effective HC XAI system and its successful implementation for other disciplinary AI models. The iterative nature of DSR allows for ongoing improvements to the system based on real-world feedback, ultimately contributing not only to agricultural sustainability but also to other disciplinary advancements.

3.6 Ethical Considerations

Throughout the process, consider ethical aspects of data privacy, informed consent, and responsible AI use in agriculture. Ethical considerations are crucial in the development and deployment of image processing and AI technologies for disease detection in agriculture. Addressing these ethical issues helps ensure that the technology is used responsibly and equitably, providing benefits to all stakeholders involved while mitigating potential harm. By focusing on data privacy, fairness, transparency, accessibility, economic impact, environmental sustainability, consent, and inclusion, developers and policymakers can create a more just and effective technological ecosystem in agriculture.

4 EVALUATION AND RESULTS

The Evaluation and Results chapter comprehensively analyses the outcomes obtained through the research methodology outlined in the preceding chapters. This section aims to critically evaluate the data collected, assess the effectiveness of the research methods employed, and elucidate the findings about the research objectives. This chapter seeks to uncover patterns, trends, and insights from the empirical investigation through meticulous analysis and interpretation, thereby contributing to a deeper understanding of the research topic. This chapter begins with an overview of the research design and methodology before delving into the presentation and discussion of the results obtained. Subsequently, the findings are analysed in light of existing literature, allowing for a nuanced evaluation of their significance and implications. Through this process, this chapter addresses the research questions posed at the outset, drawing meaningful conclusions that advance knowledge in the field.

4.1 Results Analysis

The study conducted three different XAI techniques to explain two other models: Early Detection and Classification of Rice Brown Spot and Bacterial Blight Diseases Using Digital Image Processing (Iqbal et al., 2023), and the authors trained the model using YOLO v5.
Both of these models have gone through the following XAI methods and interpreted the explainability of the models.

- Local Interpretable Model-agnostic Explanations (LIME)
- Counterfactual explanations and adversarial attacks (DiCE)
- Layer-wise Relevance Propagation (LRP)

Within the validation, we used a similar but not a paddy leaf with the same disease characteristics. This helps evaluate the disease detection capability and the truthfulness of the outcome.

4.1.1 Local Interpretable Model-agnostic Explanations (LIME)

The given table shows a comparison or evaluation of the two different models (Model 1 and Model 2) in their ability to correctly identify or classify images as either "brown spot" or "bacterial leaf blight" explained through Local Interpretable Model-agnostic Explanations (LIME) [Table 2 LIME on Brown spot disease detection, Table 3 LIME on Bacterial leaf blight disease detection]. Each row represents a model, and each column represents a specific image. The values in the table indicate whether each model correctly identified the presence

of either "brown spot" or "bacterial leaf blight" in the corresponding image.
There are three images labelled "Image 1," "Image 2," and "Image 3." The “Image 3” in both disease classes is a false image. Each image is evaluated for "brown spot" and "bacterial leaf blight." The values in the table are either "TRUE" or "FALSE," indicating whether the model correctly identified the presence of the respective disease in each image.

	brown_spot			bacterial_leaf_blight		
	Image 1	Image 2	Image 3 (Error)	Image 1	Image 2	Image 3 (Error)
Model 1	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
Model 2	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Table 8 LIME XAI result summary

Image 1 and Image 2 in the two disease categories are accurate diseased Paddy leaf images. Image 3 in the two disease categories are False Leafs, which are also not Paddy Leafs.
Model 1 and Model 2 have accurately identified "brown spot" and “bacterial leaf blight” diseases. This answers the question, "**Why is brown spot disease not shown as bacterial leaf blight disease?**” the trained model accurately identifies the two diseases separately.

In the case of "Image 3," Model 1 successfully identified the presence of a "brown spot", while Model 2 also correctly identified it. In the case of "Image 3," Model 1 didn’t recognise the presence of "bacterial leaf blight," while Model 2 correctly identified it. This answers the “**What if the diseased leaf is not similar to a paddy leaf?**” question. Even though the leaf differs, models 1 and 2 identify it as a "brown spot” disease.

This suggests that Model 1 and Model 2 have limitations in identifying the leaf type through the LIME XAI tool kit, even though they accurately identify "brown spot" and "bacterial leaf blight" in Image 3.

4.1.2 Counterfactual explanations and adversarial attacks (DiCE)

The table presents a comparative assessment of two distinct models (Model 1 and Model 2) regarding their accuracy in classifying or identifying images as "bacterial leaf blight" or "brown spot" [Table 4 DiCE on Brown spot disease detection, Table 5 DiCE on Bacterial leaf blight disease detection]. Each column corresponds to a distinct image, while each row represents a model. The value of "bacterial leaf blight" or "brown spot" that each model correctly identified in the corresponding image is denoted in the table.

There are three images labelled "Image 1," "Image 2," and "Image 3." Each image is evaluated for "brown spot" and "bacterial leaf blight." The values in the table are either "TRUE" or "FALSE," indicating whether the model correctly identified the presence of the respective disease in each image.

	brown_spot			bacterial_leaf_blight		
	Image 1	Image 2	Image 3 (Error)	Image 1	Image 2	Image 3 (Error)
Model 1	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE
Model 2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table 9 DiCE XAI result summary

In this scenario, both Model 1 and Model 2 failed to correctly identify the presence of "brown spot" or "bacterial leaf blight" in some images.

Model 1 correctly identified "brown spot" in Image 2 and "bacterial leaf blight" in Image 1 and Image 2. Model 2 did not identify either disease correctly in any of the images. Both Models didn't recognise the disease in Image 3, which is a false image.

Considering the above results, this answers the question, "**Why were brown spot disease and bacterial leaf blight disease not detected while adding noise to the image?" the trained model considers** the clarity of the disease representation in the leaf.

To the question "**What if the image has different representations?**" because the model has been trained for clear image identification, both models fail to identify the disease when the image is unclear.

Regarding counterfactual explanations and adversarial attacks, XAI shows that Model 2 depends on image quality features. Even though the image is slightly disrupted, Model 1 can successfully identify the bacterial leaf blight disease.

4.1.3 Layer-wise Relevance Propagation (LRP)

The given table is a comparative analysis of two models (Model 1 and Model 2) assessing their performance in detecting the presence of "brown spot" and "bacterial leaf blight" in three different images labelled as Image 1, Image 2, and Image 3 on Layer-wise Relevance

Propagation (LRP) [Table 6 LRP on Brown spot disease detection, Table 7 LRP on Bacterial leaf blight disease detection].

Each cell in the table contains a boolean value indicating whether the respective model correctly identified the presence of the specified disease in the corresponding image.

	brown_spot			bacterial_leaf_blight		
	Image 1	Image 2	Image 3 (Error)	Image 1	Image 2	Image 3 (Error)
Model 1	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
Model 2	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Table 10 LRP XAI result summary

Both models (Model 1 and Model 2) correctly identified the presence of a "brown spot" in all three images (Image 1, Image 2, and Image 3). Similarly, both models accurately detected the presence of "bacterial leaf blight" in Image 1 and Image 2. However, there's a discrepancy in Model 2's performance regarding "bacterial leaf blight" in Image 3, where it erroneously identified the disease (denoted by "TRUE").

Similar to LIME, in LRP, we are considering answering the same question from the XAI tool kit. Model 1 and Model 2 have accurately identified "brown spot" and “bacterial leaf blight” diseases. This answers the question, "**Why is brown spot disease not shown as bacterial leaf blight disease?**” the trained model accurately identifies the two diseases separately.

In the case of "Image 3," Model 1 successfully identified the presence of a "brown spot", while Model 2 also correctly identified it. In the case of "Image 3," Model 1 didn’t recognise the presence of "bacterial leaf blight," while Model 2 correctly identified it. This answers the “**What if the diseased leaf is not similar to a paddy leaf?**” question. Even though the leaf differs, models 1 and 2 identify it as a "brown spot” disease.

LRP XAI shows the same result as the LIME XAI outcome. Even this toolkit indicates that the error image is identified as a diseased paddy leaf.
The above XAI toolkits show that the AI model that was trained has an issue. It identifies the brown spot and the bacterial leaf blight disease, even if the leaf is not a paddy leaf.

4.2 Lesson learned

The questions we have answered from XAI tool kits show that the paddy leaf diseased image data that needs to be collected for training should include the leaf's diseased area and the whole leaf. Images of diseased paddy plants must be added to achieve a better outcome. It shows that “Why not” and the “What if” questions can effectively drive the XAI validations.

This knowledge was captured through two domain experts: a data science analyst and a paddy disease expert at “The National Institute of Plantation Management (NIPM)”.

Both parties' outcomes were that the trained data was mainly focused on the disease, not the whole paddy plant.

Figure 11 expected images

So, the training data should be recollected according to the parameters suggested by the domain experts and added to the current data set.

4.3 Next Step

After evaluating the eXplainable AI (XAI) results from the first iteration and deriving insights from the "why not" and "what if" questions, the next step involves refining the AI system or the decision-making process through a second iteration. This iterative approach allows for continuous improvement and optimisation based on the feedback and insights gathered from the initial XAI analysis.

The insights from analysing the "why not" and "what if" questions are integrated into the AI system's development process. This includes incorporating feedback on model performance, data quality, decision factors, and potential risks or benefits associated with alternative outputs.

Based on the insights derived, adjustments may be made to the AI model architecture, algorithms, or parameters to address identified limitations or improve performance. This could involve fine-tuning the model, optimising training data, or exploring alternative techniques to enhance accuracy, fairness, or interpretability.

Efforts are made to improve the quality, diversity, and relevance of the input data used by the AI system. This may involve collecting additional data, preprocessing existing data to remove biases or inconsistencies, or augmenting the dataset with synthetic or external sources to better represent real-world scenarios.

Following this iterative approach, the AI system evolves, becoming more effective, reliable, and trustworthy. Each iteration builds upon the insights gained from the XAI analysis and addresses identified gaps or opportunities for improvement, ultimately enhancing the system's value and impact in its intended application domain.

5 CONCLUSION AND FUTURE WORK

Human-centred explainable artificial intelligence (XAI) transforms technical environments by prioritising human needs, preferences, and experiences. This approach shifts the focus from purely technical solutions to creating systems that are intuitive, accessible, and user-friendly. By engaging directly with users, human-centred AI aims to design spaces that are functional and enhance the overall user experience. This re-framing emphasises the importance of empathy and understanding in technology design, striving to improve quality of life and inclusivity. When organisations prioritise human-centric design, they create more inclusive and accessible environments, ultimately leading to better outcomes for all stakeholders.

The iterative process of XAI begins with a critical evaluation of the initial results, using "why not" and "what if" questions to identify areas for improvement. This involves synthesising insights gained from the first iteration and highlighting key findings, challenges, and opportunities. Actionable recommendations for refining the AI system are then developed, focusing on aspects such as model performance, data quality, interpretability, fairness, and robustness. These recommendations should be prioritised based on their potential impact, feasibility, and alignment with stakeholder objectives, considering factors like resource constraints, timeframes, and dependencies.

Clear goals and objectives must be established for the next iteration, defining what the AI system aims to achieve and how success will be measured. Performance metrics and benchmarks are crucial for evaluating the effectiveness of the enhancements. By systematically refining the AI system through these iterative evaluations, stakeholders can ensure improved performance, transparency, and trustworthiness in the AI's deployment and operation.

Promoting the responsible use of technology involves fostering ethical, safe, and mindful engagement with technological tools. Organisations and communities must collaborate to mitigate potential risks, ensuring that technology drives positive change and societal progress. Expanding practitioners' toolboxes with XAI tools enhances AI systems' transparency, interpretability, and trustworthiness. This empowers data scientists, machine learning engineers, and domain experts to develop and maintain AI technologies that are not only effective but also understandable and reliable. By integrating XAI tools, organisations can foster greater understanding and acceptance of AI technologies across various domains.

5.1 Integrating a human-centred approach into Explainable AI (XAI)

Human-centred explainable artificial intelligence (XAI) represents a paradigm shift in designing and implementing technical systems, foregrounding the human experience in every aspect of technological development. Unlike traditional approaches, which prioritise technical efficiency and performance, human-centred XAI emphasises user intuition, accessibility, and overall experience. This reorientation demands a comprehensive understanding of user needs and preferences, ensuring that the resulting environments are functional and enhance user satisfaction and quality of life.

The dataset used in this study consisted of complex image data, which presented unique challenges and opportunities for developing a robust AI framework. However, this framework is not limited to image data alone. It can be adapted and verified using other types of datasets, such as tabular or frequency data, to evaluate its versatility and effectiveness across different data modalities.

By embedding empathy and user-centric principles into the design process, organisations can create more inclusive and accessible technological solutions, improving outcomes and broader societal benefits.

5.2 Iterative refinement from explainable AI educated outcome

The iterative nature of XAI involves a rigorous and ongoing evaluation and improvement process. The first iteration is foundational, where initial outcomes are scrutinised through critical "why not" and "what if" questions. This reflective analysis is essential for identifying shortcomings and areas for enhancement. Insights gleaned from this process form the basis for actionable recommendations to refine the AI system in subsequent iterations.

The study by Q. Vera Liao and Kush R. Varshney raises additional questions that need further exploration to validate XAI outcomes. These questions highlight the complexity and depth required to understand and implement Explainable AI (XAI) comprehensively.

Critical areas for improvement typically include model performance, data quality, interpretability, fairness, and robustness. Recommendations should be prioritised based on their potential impact, feasibility, and alignment with stakeholder objectives. This prioritisation must also account for practical constraints such as resource availability, timeframes, and interdependencies among different enhancements.

From this study, it shows that setting clear goals and performance metrics for each iteration is crucial. These benchmarks serve as indicators of success, guiding the evaluation of the AI system's effectiveness post-enhancement. By adopting this iterative and reflective approach,

stakeholders can systematically improve the AI system, enhancing its performance, transparency, and trustworthiness.

5.3 Responsible use of technology

Promoting the responsible use of technology is critical in ensuring that advancements in AI and other technological fields are ethically and safely integrated into society. Organisations and communities must collaborate to address potential risks and ensure that technology contributes positively to societal progress.

Different domain experts bring unique perspectives, which can significantly enhance the training and trustworthiness of an AI model. This study employs the insights of several individuals from diverse backgrounds to capture a broad range of perspectives during the model training process. Integrating these varied viewpoints makes the training model's outcome more reliable and trustworthy. This involves creating frameworks and guidelines that encourage ethical behaviour and mindful engagement with technological tools.

5.4 Expanding practitioners' toolbox with XAI

Integrating XAI tools into practitioners' toolbox is vital for enhancing AI systems' transparency, interpretability, and trustworthiness. These tools empower data scientists, machine learning engineers, and domain experts to develop, deploy, and maintain comprehensible and reliable AI systems. The proposed framework represents a significant step towards semi-automated processes in the domain of HCAI. However, there is ample room for further research and development to transition this framework into a fully automated system.

In conclusion, the re-framing of technical spaces through human-centred XAI and the iterative refinement process represent significant advancements in making AI systems more user-friendly, transparent, and trustworthy. By addressing ethical considerations and promoting responsible technology use, organisations can foster a more inclusive and positive technological landscape, ultimately enhancing the societal impact of AI innovations.

APPENDICES

Use LIME with a custom-trained model for image classification.

```
import numpy as np
from skimage.segmentation import mark_boundaries
from lime import lime_image
from keras.preprocessing import image
from keras.models import load_model

# Load your custom-trained model
custom_model = load_model('path_to_your_custom_model.h5')

# Define a function to preprocess input images for your model
def preprocess_input_img(img_path):
    img = image.load_img(img_path, target_size=(299, 299))
    x = image.img_to_array(img)
    x = np.expand_dims(x, axis=0)
    x = x / 255.0  # Normalize pixel values
    return x

# Sample image path
img_path = 'path_to_your_image.jpg'

# Preprocess the input image
img = preprocess_input_img(img_path)

# Create LIME explainer
explainer = lime_image.LimeImageExplainer()

# Define a function to predict with your custom model
def custom_predict(images):
    # Assuming custom_model.predict() returns the prediction probabilities
    return custom_model.predict(images)

# Explain prediction
explanation = explainer.explain_instance(img[0], custom_predict, top_labels=5, hide_color=0, num_samples=1000)

# Show explanation
temp, mask = explanation.get_image_and_mask(explanation.top_labels[0], positive_only=True, num_features=5, hide_rest=True)
plt.imshow(mark_boundaries(temp / 2 + 0.5, mask))
```

Use DiCE with a custom-trained model for image classification.

```
pip install dice-ml
from dice_ml.explainer_interfaces.dice_gaussian import DiceGaussian
from dice_ml.utils import helpers # Import helper functions

# Load your trained model
# Replace "your_model" with the code to load your trained model
# For example:
# from keras.models import load_model
# model = load_model('path_to_your_model.h5')

# Initialize DiceGaussian explainer
explainer = DiceGaussian(dice_expansion_method='uniform')

# Sample data point for which counterfactuals are to be generated
# Replace "sample_point" with your actual data point
sample_point = {'feature1': value1, 'feature2': value2, ...}

# Define the outcome you want the counterfactual to achieve
# Replace "desired_outcome" with the desired outcome
desired_outcome = value

# Generate counterfactuals
counterfactuals = explainer.generate_counterfactuals(sample_point, total_CFs=5, desired_class=desired_outcome)

# Print counterfactuals
for cf in counterfactuals.final_cfs_df:
    print(cf)
```

Use LRP with a custom-trained model for image classification.

```
import numpy as np
import matplotlib.pyplot as plt
from keras.applications.vgg16 import VGG16, preprocess_input, decode_predictions
from keras.preprocessing import image
import innvestigate

# Load the custom model
model = (“include the custom model”)
model.summary()

# Select the image to explain
img_path = 'path_to_your_image.jpg'

# Load and preprocess the image
img = image.load_img(img_path, target_size=(224, 224))
x = image.img_to_array(img)
x = np.expand_dims(x, axis=0)
x = preprocess_input(x)

# Create an analyzer for Layer-wise Relevance Propagation (LRP)
lrp_analyzer = innvestigate.analyzer.relevance_based.relevance_analyzer.LRPZ(model)

# Apply LRP to the image
explanation = lrp_analyzer.analyze(x)

# Plot the explanation
plt.imshow(explanation.squeeze(), cmap='jet', alpha=0.5)
plt.imshow(img, alpha=0.5)
plt.axis('off')
plt.show()
```

REFERENCES

Adjei Domfeh, E., Weyori, B., Appiahene, P., Mensah, J., Awarayi, N., Afrifa, S., 2022. Human-Centered Artificial Intelligence, a review. https://doi.org/10.22541/au.166013641.15972664/v1

Angelov, P.P., Soares, E.A., Jiang, R., Arnold, N.I., Atkinson, P.M., 2021. Explainable artificial intelligence: an analytical review. WIREs Data Mining and Knowledge Discovery 11, e1424. https://doi.org/10.1002/widm.1424

Daniya, T., Vigneshwari, S., 2022. Deep Neural Network for Disease Detection in Rice Plant Using the Texture and Deep Features. The Computer Journal 65, 1812–1825. https://doi.org/10.1093/comjnl/bxab022

DARPA, n.d. Explainable Artificial Intelligence [WWW Document]. Defense Advanced Research Projects Agency. URL https://www.darpa.mil/program/explainable-artificial-intelligence (accessed 2.26.24).

Deng, R., Tao, M., Xing, H., Yang, X., Liu, C., Liao, K., Qi, L., 2021. Automatic Diagnosis of Rice Diseases Using Deep Learning. Front. Plant Sci. 12. https://doi.org/10.3389/fpls.2021.701038

Duan, Y., Edwards, J.S., Dwivedi, Y.K., 2019. Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. International Journal of Information Management 48, 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Ehsan, U., Wintersberger, P., Liao, V., Watkins, E., Manger, C., III, H., Riener, A., Riedl, M., 2022. Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI. pp. 1–7. https://doi.org/10.1145/3491101.3503727

Früh, A., Haux, D., 2022. Foundations of Artificial Intelligence and Machine Learning, Weizenbaum Series. Weizenbaum Institute for the Networked Society - The German Internet Institute, Berlin. https://doi.org/10.34669/WI.WS/29

Holzinger, A., Saranti, A., Angerschmid, A., Retzlaff, C.O., Gronauer, A., Pejakovic, V., Medel-Jimenez, F., Krexner, T., Gollob, C., Stampfer, K., 2022. Digital Transformation in Smart Farm and Forest Operations Needs Human-Centered AI: Challenges and Future Directions. Sensors 22, 3043. https://doi.org/10.3390/s22083043

Htun, K.W., Htwe, C.S., 2018. Development of Paddy Diseased Leaf Classification System Using Modified Color Conversion 6.

Hu, F., Hillary, H., 2023. Developing industrial indoor rice production: AI using for CEA and rapid growth technology. https://doi.org/10.13140/RG.2.2.11901.74724

Iqbal, J., Hussain, I., Hakim, A., Ullah, S., Yousuf, H., 2023. Early Detection and Classification of Rice Brown Spot and Bacterial Blight Diseases Using Digital Image Processing.

Islam, T., Sah, M., Baral, S., Roy Choudhury, R., 2018. A Faster Technique on Rice Disease Detectionusing Image Processing of Affected Area in Agro-Field, in: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). Presented at the 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 62–66. https://doi.org/10.1109/ICICCT.2018.8473322

Kahar, M.A., Mutalib, S., Abdul-Rahman, S., 2015. Early Detection and Classification of Paddy Diseases with Neural Networks and Fuzzy Logic.

Liao, Q.V., Varshney, K.R., 2022. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. https://doi.org/10.48550/arXiv.2110.10790

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.-R., 2019. Layer-Wise Relevance Propagation: An Overview, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 193–209. https://doi.org/10.1007/978-3-030-28954-6_10

Pugliese, R., Regondi, S., Marini, R., 2021. Machine learning-based approach: global trends, research directions, and regulatory standpoints. Data Science and Management 4, 19–29. https://doi.org/10.1016/j.dsm.2021.12.002

Rahman, C.R., Arko, P.S., Ali, M.E., Iqbal Khan, M.A., Apon, S.H., Nowrin, F., Wasif, A., 2020. Identification and recognition of rice diseases and pests using convolutional neural networks. Biosystems Engineering 194, 112–120. https://doi.org/10.1016/j.biosystemseng.2020.03.020

Rawat, P., Pandey, A., Panaiyappan.K, A., 2023. Rice Leaf Diseases Classification Using Deep Learning Techniques, in: 2023 International Conference on Networking and Communications (ICNWC). Presented at the 2023 International Conference on Networking and Communications (ICNWC), pp. 1–8. https://doi.org/10.1109/ICNWC57852.2023.10127315

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. https://doi.org/10.48550/arXiv.1602.04938

Sethy, P.K., Barpanda, N.K., Rath, A.K., Behera, S.K., 2020. Image Processing Techniques for Diagnosing Rice Plant Disease: A Survey. Procedia Computer Science, International Conference on Computational Intelligence and Data Science 167, 516–530. https://doi.org/10.1016/j.procs.2020.03.308

Shneiderman, B., 2020a. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. International Journal of Human–Computer Interaction 36, 495–504. https://doi.org/10.1080/10447318.2020.1741118

Shneiderman, B., 2020b. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. https://doi.org/10.48550/arXiv.2002.04087

T, S., T, D., 2015. CLASSIFICATION OF PADDY LEAF DISEASES USING SHAPE AND COLOR FEATURES 07.

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C.G., van Moorsel, A., 2020. The relationship between trust in AI and trustworthy machine learning technologies, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20. Association for Computing Machinery, New York, NY, USA, pp. 272–283. https://doi.org/10.1145/3351095.3372834

Wachter, S., Mittelstadt, B., Russell, C., 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. https://doi.org/10.48550/arXiv.1711.00399

Wang, D., Yang, Q., Ashraf, A., Lim, B., 2019. Designing Theory-Driven User-Centric Explainable AI. https://doi.org/10.1145/3290605.3300831

West, D.M., Allen, J.R., 2018. How artificial intelligence is transforming the world [WWW Document]. Brookings. URL https://www.brookings.edu/articles/how-artificial-intelligence-is-transforming-the-world/ (accessed 7.10.23).

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J., 2019. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges, in: Tang, J., Kan, M.-Y., Zhao, D., Li, S., Zan, H. (Eds.), Natural Language Processing and Chinese Computing, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 563–574. https://doi.org/10.1007/978-3-030-32236-6_51