Arab American University

Faculty of Graduate Studies

Enhancing Agricultural Research in Palestine through GIS, IoT, and Fog Computing Integration

By

Jebreel Ahmad Saeed Abed

Supervisor

Dr. Amjad Rattroot

Co-Supervisor

Dr. Jacqueleen Joubran

This thesis was submitted in partial fulfillment of the requirements for the Master's degree in Computer Science

July /2024

Thesis Approval

Enhancing Agricultural Research in Palestine through GIS, IoT, and Fog Computing Integration

By

Jebreel Ahmad Saeed Abed

This thesis was defended successfully on 6/7/2024 and approved by:

Committee members	Signature
1. Dr. Amjad Rattroot: Supervisor	Dr Amjad RATTROUT ………………………
2. Dr. Jacqueleen Abu Daoud: Co-Supervisor	………………………
3. Dr. Ahmad Ewais: Internal Examiner	………………………
4. Dr. Rashid Jayousi: External Examiner	R. Jayousi ………………………

Declaration

This is to declare that the thesis entitled “Enhancing Agricultural Research in Palestine through GIS, IoT, and Fog Computing Integration” under the supervision of Dr. Amjad Rattroot and Co-Supervisor Dr. Jacqueleen Joubran is our work and does not contain any unacknowledged work or material previously published or written by another person, except where due reference is made in the text of the document.

The Name of The Student: Jebreel Ahmad Saeed Abed

ID: 202011999

Signature: Jebreel Abed

Date: 14/9/2024

Acknowledgments

In the name of Allah, the Most Gracious, the Most Merciful

I would like to express my sincere gratitude to my supervisor, Dr Amjad Rattroot, and Dr. Jacqueleen Joubran for their valuable guidance and advice throughout the preparation of this thesis. Their expertise and unwavering support have played a crucial role in the success of this thesis.

Furthermore, I extend my deepest thanks to Dr. Muhammad Saleh, who stood with me and helped me begin this thesis.

Furthermore, I extend my deepest thanks to Dr. Muhammad Talib Abed, who stood with me and helped me coordinate this message.

Furthermore, I extend my deepest thanks to Mr. Hussein Younis, who helped and supported me throughout my studies.

To my family and friends, I express my heartfelt thanks for your unwavering support and encouragement at every stage of this thesis. Your moral support has been instrumental in my ability to achieve this milestone.

Finally, I eagerly anticipate presenting the findings of this thesis with pride and confidence, with the hope that it will contribute to knowledge and advancement in the field of study.

Once again, I would like to express my gratitude to everyone for their valuable contributions and continuous support.

Abstract

Agricultural research in the Middle East, particularly in Palestine, faces challenges like limited freshwater access, lack of real-time data for fertilizer regulation, and scattered data sources. To tackle these issues, this study suggests an integrated system using GIS, IoT, and fog computing. The system includes a water control and fertilizer regulatory cloud, along with a national geodatabase cloud for research, based at the National Agricultural Research Centre in Palestine. Soil sensors monitor key parameters like N, P, K, pH, moisture, temperature, and EC, combined with weather data to optimize fertilizer use and promote TWW irrigation for sustainable practices. The geodatabase cloud serves as a central data repository for agricultural research, integrating data from the Palestinian Plant Gene Bank to predict plant species distribution in the West Bank using ArcGIS and Random Forest, achieving an accuracy of 0.72. By integrating GIS, IoT, and fog computing, the system aims to improve data collection, storage, and analysis for informed decision-making and sustainable agricultural strategies in the region.

Table of contents

List of Tables

List of Figures

List of Abbreviations

TWW: Treated Waste Water

GIS: Geographic information systems

IoT: Internet of Things

P: Phosphorus

N: Nitrogen

K: Potassium

NARC: National Agricultural Research Centre

COM: Component Object Model

SDM: Spatial Data Mining

ML: Machine learning

AI: Artificial Intelligence

k-NN: k-nearest neighbors

CART: the Classification and Regression Tree

SRM: the Structural Risk Minimization

NN: Neural Networks

KDD: Knowledge Discovery in Databases

PANGEA: Platform for automatic coNstruction of orGanizations of intElligent agents

IDS: Intrusion Detection System

FIS: Fertilizer and irrigation system

Chapter 1: Introduction

This chapter is served as an introduction to the study. It provides background information on the topic and presents the research problem and questions that will be posed. The aims and objectives of the study are outlined. The significance and motivation behind the study is discussed, emphasizing its importance. Additionally, the chapter delves into the research scope, defining the boundaries and extent of the study. At Last, the organization of the thesis is described, and providing an overview of how the subsequent chapters are structured.

1.1 Background Information

Research in the agricultural sector is one of the most demanding fields in the world and in the Middle East. Most countries in the Middle East are exposed to many problems in this field such as the lack of fresh water, real-time fertilizer regulation, and scattered data.

The water sector in Palestine faces a weakness in maintaining water supplies. Because the Israeli occupation has prevented most Palestinians from building underground wells and depriving them from sea and river water, Palestinians suffer from potable and agricultural water scarcity. (BURHAN, Awad, & RUTROT, 2022) Therefore, freshwater control, and wastewater reuse will be developed because water is one of the most critical issues for Palestinian farmers. (Keraita, Blanca, & Pay, 2008) Additionally, treated wastewater (TWW) should be utilized as much as possible, since it is considered an uninterrupted and inexpensive source of water and a source of income in some countries. (Alkhamisi & Ahmed, 2014)

Fertilizer regulation is an essential part to Agriculture. The most vital fertilizers are Phosphorus (P), Nitrogen (N) and Potassium (K) which help crop yields, whether grains, straw, fruits, or others. The proportion of each fertilizer varies according to the type of crop and soil. (Area-Ren) So, the system should measure them because of their side effects of using large amounts. Excessive use of these fertilizers can affect several problems, whether on the soil or humans, so we want to control the amount of fertilization for each of them separately. (Alkhamisi & Ahmed, 2014)

Agricultural research in Palestine needed a central database to build several research depending on data collected locally, thus the researchers needed references to get needed data to upgrade agricultural researches and Agriculture in Palestine and the Middle East. To execute this these ideas, some techniques were and will be used GIS, IoT, cloud computing, and machine learning.

Geographic information systems (GIS) are integrated systems used to process mapping, databases, spatial analysis, and management of data. Therefore, the use of this technique has a significant impact on the proposal to obtain an integrated geographical database. (Maguire, 1991)

The Internet of Things (IoT) is one of the most popular technologies in the world, as it is considered a global infrastructure for physical things connected to the Internet. (Evangelos A, Nikolaos D, & Anthony C, 2011) It is considered a global network that connects human-to-human, human-to-things, and object-to-things. (Aggarwal & Das, 2012)

Cloud computing is a technique used to store and control data on the Internet in general. One of the most important research areas at this time that follows this technique is the fog computing technique, which combines IoT and cloud computing so that it can be used to control different systems of IoT at the same time. (Aazam, Zeadally, & Harras, 2018)

Machine learning (ML) is a rapidly evolving field within artificial intelligence (AI) that focuses on developing algorithms and statistical models that enable computers to perform tasks without explicit instructions. ML algorithms build models based on sample data, known as "training data," to make predictions or decisions without being explicitly programmed to perform the task. This literature review provides an overview of the historical development, theoretical foundations, methodologies, applications, recent advancements, and challenges in machine learning. (Mahesh, 2020)

The contribution appears when we build a geographical database centre of the agricultural research in Palestine to save the data (at NARC); so, the research will be applied in NARC and can access all people who are interested in.

1.2 Research Problem and Motivation

Palestine is a country with many natural water sources; however, it suffers from a shortage of water supplies, due to many factors, most notably the Israeli occupation, which controls the main water sources, their volatility, and the recent lack of rainwater. Therefore, it is necessary to improve an irrigation and fertilization system to reduce the use of fresh water in agriculture, moreover treated wastewater must be used in agriculture while ensuring that productivity is maintained. (Saleh, 2018)

Also, the data obtained by researchers is at risk of being lost for several reasons, the most important is the lack of a computerized system to store agricultural data in Palestine. Therefore, we propose a computerized system to save and display data in a database cloud in a way that is easy to handle. We also integrated the database to include geographical data.

1.3 Research Question and Objectives

When considering the problems of freshwater shortage and the lack of a database of agriculture in Palestine, several questions are raised:

- Is it feasible to computerize an integrated irrigation and fertilizer system, and what challenges may arise?
- Can a database be established for Palestine's agricultural sector, and what obstacles might be encountered?
- How can the database be linked with GIS for enhanced functionality and data management?
- What strategies can be employed to achieve the highest accuracy in machine learning applications?
- How can ArcGIS and machine learning be effectively integrated, and what process should be followed?

Also, it is important to highlight that some challenges arise related to collecting data in Palestine since the data is in separate places and most of the existing data is Excel files, not shape files as the ArcGIS program deals with shape files. However, this research tries to achieve the objectives as follows:

1. Setting up a pioneering integrated geographical database to serve agriculture in Palestine, specifically to smooth data management.
2. Increase the transparency of, and direct access of, critical data to researchers for their studies.
3. Introduce measures that will reduce the usage of freshwater in farming to ensure sustainability.

4. Use gathered data intelligently, by recognizing when to make an informed or strategic choice.
5. Provide real-time updates on agricultural data for effective and timely decision-making, thereby resulting in increased productivity.

1.4 Thesis Organization

The following sections of this thesis are organized in separate chapters. The literature review in Chapter 2 is the backbone of this research; it declares the important past studies on the subject matter. Methods of research are discussed in Chapter 3, and the strategies for the proposed irrigation and fertilizer system, and central geo-database construction are specified. Chapter 4 deals with the details of the development of an integrated model combining machine learning predictions in ArcGIS. Chapter 5 presents the results of the prediction model of Fuzzy Logic. Chapter 6 encapsulates the summary of recommendations and further avenues to explore.

Chapter 2: Background Information

In this chapter, an overview of relevant literature is presented. It explores existing research and identifies gaps.

2.1 The Programming Language for GIS (ArcGIS)

ArcGIS is a family of client, server, and online geographic information system (GIS) software developed and maintained by Esri. ArcGIS was first released in 1999 and originally was released as ARC/INFO, a command line-based GIS system for manipulating data. ARC/INFO was later merged into ArcGIS Desktop, which was eventually superseded by ArcGIS Pro in 2015. ArcGIS Pro works in 2D and 3D for cartography and visualization, including machine learning (ML). Esri also provides server-side ArcGIS software for web maps, known as ArcGIS Server. (Booth & Mitchell, 2001)

ArcMap serves as the primary component within the ArcGIS software suite, primarily employed for the creation, editing, analysis, and visualization of spatial datasets utilizing a 2-dimensional (2D) render window. To enhance the capabilities of ArcGIS, users can utilize Component Object Model (COM)-based programming languages and leverage ArcObjects, a developer kit based on COM technology specifically designed for ArcGIS. (Mendas & Delali, 2012)

Sameer M. Shadeed et al. used ArcGIS to build a map of domestic rainwater harvesting map for the West Bank. They categorized them into five areas (very high, high, moderate, low, and very low). As shown in figure 1. (Shadeed, Judeh, & Almasri, 2019)

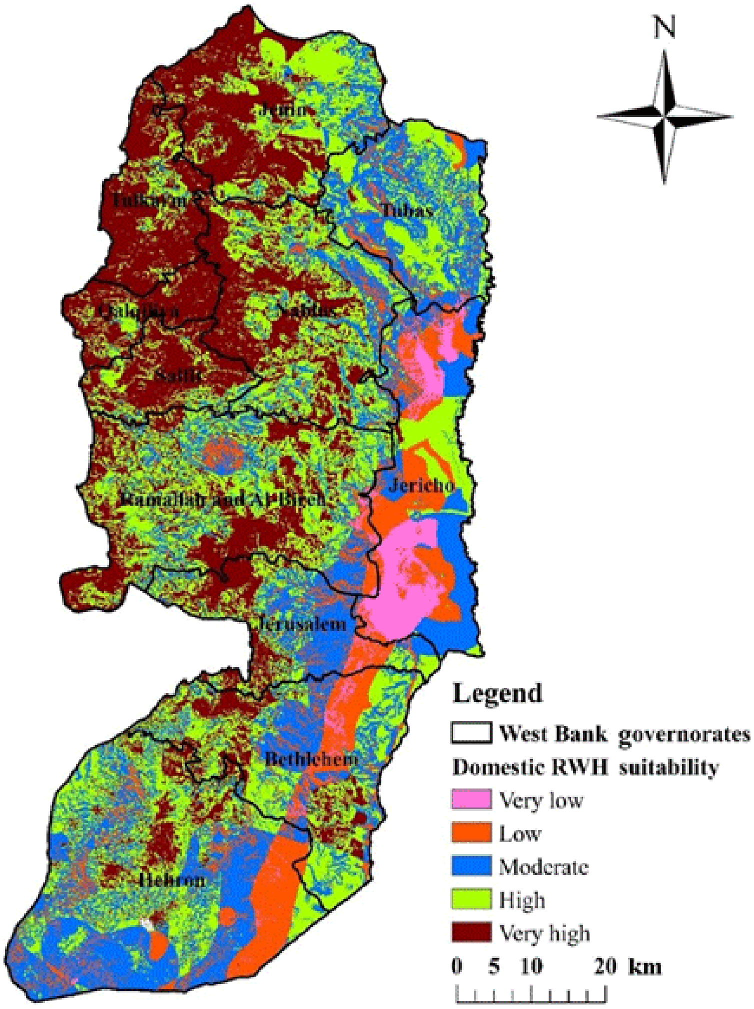

Figure 1 Domestic rainwater harvesting suitability map for the West Bank. ***(Shadeed, Judeh, & Almasri, 2019)***

2.2 Spatial Data Mining (SDM) and Geographic Information System (GIS)

Spatial Data Mining (SDM) technology encompasses advanced techniques, methods, and tools to analyze and extract meaningful patterns and insights from spatial or geographic data, to inform decision-making and gain a deeper understanding of spatial phenomena. (Goyal, Sharma, & Joshi, 2017)

Geographic Information System (GIS) is a computer-based system that captures, stores, manages, analyzes, and visualizes spatial or geographic data. It involves data input from various sources, data storage in specialized databases, data analysis using tools and techniques for spatial queries, statistics, and modeling. It also involves data visualization through maps, other graphical representations, data output for communication and further analysis. GIS has many applications in urban planning, natural resource management, environmental monitoring, transportation, public health, and emergency management. It provides a powerful toolset for managing and analyzing spatial data, facilitating decision-making, and gaining insights into geographic patterns and relationships. (Maguire, 1991)

The paper discusses the characteristics of spatial data, GIS data sources, data formats, and data representation. It also highlights the challenges of dealing with large volumes of data in GIS databases. It proposes an architecture for data integration using a data warehouse approach to address this challenge. The paper further discusses various data mining algorithms used in GIS applications and emphasizes the role of spatial association rule mining in big data of GIS. The paper proposes future implementation and testing of the proposed solution in agriculture. (Zaragozí, et al., 2012)

They mention that GIS stores data from various sources in different formats in geodatabases, representing spatial features. These geodatabases are constantly growing due to data generated from satellite images and other sources related to natural resources, traffic monitoring, tourist monitoring, health management, and biodiversity conservation. (Perumal, Velumani, Sadhasivam, & Ramaswamy, 2015)

2.3 Decision Tree

Decision trees are a fundamental and widely used method in machine learning and statistics for classification and regression tasks. Their intuitive structure, easy interpretation, and capability to handle various types of data have cemented their importance in the field. This literature review delves into the historical development, theoretical foundations, methodologies, applications, advancements, and challenges associated with decision trees.

The decision tree algorithm has been known roots in the early 1960s, with it is notable contributions from the fields of statistics and computer science. One of the seminal works is presented by Breiman et al. (1984) with the introduction of the Classification and Regression Trees (CART) methodology. Around the same period, Quinlan (1986) developed the ID3 algorithm, which laid the groundwork for subsequent algorithms like C4.5 and C5.0. (Breiman, Classification and regression trees, 2017) (Yan-Yan & Ying, 2015)

Decision trees remain a cornerstone in machine learning due to their simplicity, interpretability, and versatility. They are foundational to more complex ensemble methods that mitigate some of their inherent weaknesses. Ongoing research continues to enhance their efficiency, accuracy, and applicability across various fields, ensuring their continued relevance in the evolving landscape of data science.

2.4 Random Forest

An ensemble learning method, has become a cornerstone in machine learning since its inception. This technique, developed by Leo Breiman in 2001, combines the predictions of multiple decision trees to enhance accuracy and robustness. Random Forests have found applications across diverse fields, including bioinformatics, finance, and image recognition, demonstrating their versatility and effectiveness.

The concept of Random Forests emerged from the ensemble learning domain, which seeks to improve predictive performance by aggregating the results of multiple models.

Breiman's pivotal paper, "Random Forests," outlined a methodology that builds multiple decision trees using randomly selected subsets of data and features, then calculate the averages of their predictions to mitigate overfitting and variance issues inherent in single decision trees. (Breiman, Random forests. Machine learning, 2001)

Random Forest operates by constructing a multitude of decision trees during training and outputting the mean prediction (regression) or mode of the classes (classification) of the individual trees. Random Forests have solidified their position as a reliable and versatile tool in the machine learning arsenal. Their ability to handle diverse types of data and deliver robust predictions with minimal overfitting makes them an invaluable method across numerous domains. Ongoing research and advancements continue to address their limitations, ensuring that Random Forests remain at the forefront of machine learning techniques.

2.5 Support Vector Machine (SVM)

Support Vector Machines (SVMs) are powerful supervised learning models used for classification, regression, and outlier detection. Introduced by Vladimir Vapnik and his colleagues in the early 1990s, SVMs are grounded in statistical learning theory and have become a key technique in the machine learning community due to their robustness, accuracy, and ability to handle high-dimensional data.

The concept of SVMs originated from Vapnik's work on statistical learning theory, culminating in the development of the first SVM algorithm in 1992. (Cortes & Vapnik, 1995) The SVM algorithm was designed to find the optimal hyperplane that separates different classes in the feature space, maximizing the margin between the classes. This theoretical foundation is based on the Structural Risk Minimization (SRM) principle, which aims to balance model complexity and generalization capability.

Support Vector Machines have established themselves as a powerful and versatile tool in the machine learning landscape. Their theoretical foundations, coupled with robust performance across various domains, make them a valuable asset for classification and regression tasks. Ongoing research and advancements continue to address their limitations, ensuring that SVMs remain relevant in the rapidly evolving field of machine learning.

2.6 Neural Networks (NN)

Neural Networks (NN) are a foundational element of modern artificial intelligence, modeled after the human brain's interconnected neuron structure. They are used for a variety of tasks, including classification, regression, image and speech recognition, and more. This literature review explores the development, theoretical foundations, methodologies, applications, recent advancements, and challenges associated with neural networks.

The concept of neural networks dates back to the 1940s, with McCulloch and Pitts (1943) proposing a computational model for neural networks. However, it wasn't until the 1980s that neural networks gained significant traction, primarily due to the development of backpropagation. The resurgence of interest in neural networks in the 2000s and 2010s, often referred to as the "deep learning revolution," was driven by increased computational power, the availability of large datasets, and algorithmic advances. (Yongjun, et al., 2021) (Zhang, Rao, & Agrawala, 2023)

Neural networks have transformed numerous fields with their ability to learn and generalize from data. Their versatility and power have led to groundbreaking achievements in areas like computer vision, natural language processing, and healthcare. Continued research and advancements are addressing existing challenges, ensuring that neural networks remain at the forefront of artificial intelligence and machine learning.

Chapter 3: Literature Reviews

In this chapter, an overview of relevant literature is presented. It explores existing research and identifies gaps in the literature that the study aims to address.

Goyal et al. (2017) provide an overview of SDM technology in the context of GIS data. The explosive growth of spatial data has created a demand for innovative approaches to analysing geodatabases, which contain heterogeneous data from multiple sources in different formats. (Goyal, Sharma, & Joshi, 2017)

Zaragozí, et al. aim to contribute to understanding the driving factors behind farmland abandonment. It proposes a new approach that combines GIS and knowledge discovery in databases (KDD) techniques to identify the most critical variables for studying the process. The study findings can be helpful for policymakers and land managers to develop effective strategies for addressing farmland abandonment and its associated impacts. (Zaragozí, et al., 2012)

Perumal et al. (2015) aim to provide an overview of GIS data formats, data representation models, data sources, data mining algorithmic approaches, SDM tools, issues, and challenges. They propose an architecture to address the challenges of GIS data and view GIS as a big data problem. The paper highlights the need for semantic integration of GIS datasets, the challenge of volume and data formats, and proposes an architecture using a data warehouse approach and ontology in GIS. Future work involves implementing and testing the proposed architecture and presenting SDM tools with their merits and limitations. (Perumal, Velumani, Sadhasivam, & Ramaswamy, 2015)

Li et al. (2013) discusses a study conducted from Yushu City in China using geostatistical modules in ArcGIS and a weighted fuzzy clustering algorithm to investigate the distribution and variability of soil nutrients in 576 soil samples from 3 and 7 field sites in-depth. The results are compared with the second national soil survey. Four interpolation methods, namely ordinary Kriging, inverse distance weighted, local polynomial, and global polynomial, are used to predict the spatial variation of soil nutrients. (Li, Chen, Zeng, & Ye, 2013)

Karimipour et al. (2005) discuss the increasing production of digital geospatial data and the role of GIS in managing and analysing such data. The need for geospatial data mining and knowledge discovery is emphasized, focusing on its applications in environmental data management, specifically water quality management. The paper uses geospatial data mining to present a case study that models the correlation between industrial pollution and water quality indicators in Iran's Western and Eastern Azerbaijan Provinces. The results highlight the importance of spatial analysis in water quality management and suggest that geospatial data mining can be used for industrial site selection in master planning stages. The paper concludes by discussing the challenges, such as the need for qualified and complete datasets. It suggests future improvements, including using temporal GIS for spatiotemporal analysis of water resources data. (Karimipour, Delavar, & Kinaie, 2005)

Singleton et al. (2021) discusses the emergence of Data Science and its potential role in Geography, particularly with the growth of spatial "Big Data." The authors argue that while Data Science has expanded to consider more geographic problems, there is a need for closer coupling and assimilation of Geography and Data Science to develop new methodological and epistemological frameworks. The article proposes the establishment of Geographic Data Science within Geography, which would benefit from the critically reflective perspective of Geography and the methodological contributions of Data

Science. The authors suggest a research agenda toward Geographic Data Science that includes systems engineering, new methodological development, and addressing acute challenges of epistemology. And they conclude that there are benefits to integrating Geography and Data Science, both in practical terms and in sustaining the relevance of Geography within a rapidly changing socio-technological landscape. (Singleton & Arribas-Bel, 2021)

Goap et al. (2018) intelligent irrigation architecture is proposed, which is based on the Internet of Things (IoT) and a hybrid machine learning approach to predict soil moisture. The algorithm utilizes recent sensor data and weather forecasted data to indicate the soil moisture levels in the upcoming days, resulting in improved accuracy and reduced error rates. The prediction approach has been integrated into a standalone system prototype that is cost-effective and customizable for specific scenarios. The system also has an auto mode, making it an intelligent system. In this article, they use a database server to create an integrated system for irrigation only. They also use a sensor for each of the mentioned signs, where there is one sensor that measures 3 of the mentioned signs in addition to two other signs to get better results. (Goap, Sharma, Shukla, & Krishna, 2018)

Villarrubia et al. (2017) aim to create virtual organizations of agents that can communicate and monitor crops. Using a low-cost sensor architecture, farmers can optimize the growth of their crops by managing resources in real time. Since the hardware has limited processing and communication capabilities, the PANGEA (Platform for automatic coNstruction of orGanizations of intElligent agents) architecture is used to overcome this limitation. The system is designed to collect heterogeneous information from sensors measuring temperature, solar radiation, humidity, pH, moisture, and wind. The system can merge the data from these sensors and produce a response tailored to the context. To validate the system, a case study was conducted where farmers were provided with a tool to monitor the condition of their crops on a smart screen using a low-cost device. This study offers an innovative solution to help farmers improve crop growth and resource management. As shown in figure 2. (Villarrubia, De Paz, De La Iglesia, & Bajo, 2017)

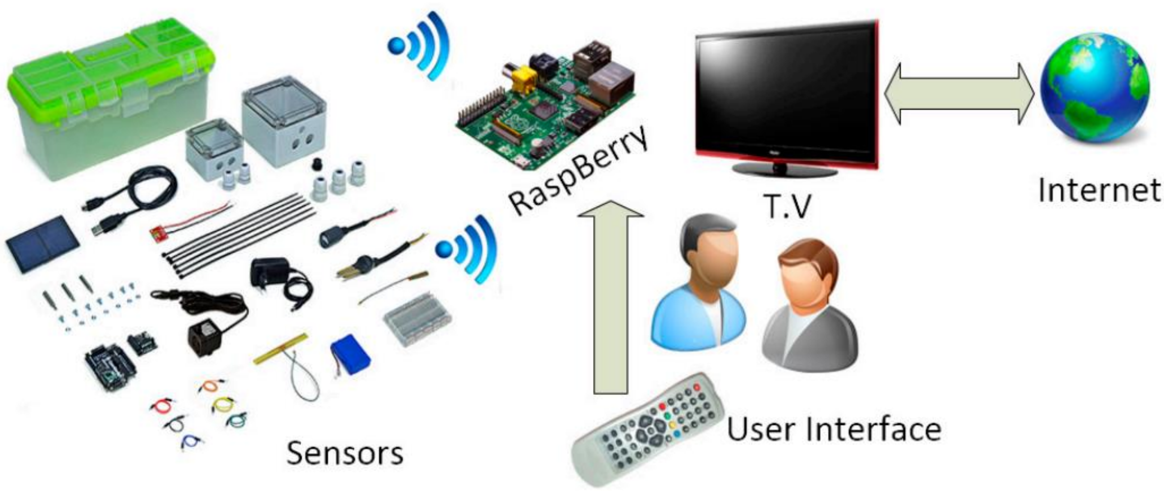

Figure 2 Remote control platform of the irrigation system. ***(Villarrubia, De Paz, De La Iglesia, & Bajo, 2017)***

Jisha rt al. (2019) proposed an IoT system to track the water level in the tank to reduce the waste of water and electricity by using a sensor connected to a smartphone, as well as reading the degree of soil moisture to see if it needs water through sensors connected to the same phone. (Jisha, Vignesh, & Deekshit, 2019)

Kunal et al. (2019) focus on security concerns at each architecture level to prevent any unauthorized access or alteration of data. The paper also addresses potential security issues like authentication, integrity, secure storage, key management, and intrusion detection system (IDS) in fog devices and cloud computing. An all-encompassing secure, dependable framework can be applied in various areas of human life to securely and efficiently gather relevant information. For that, fog computing can be used by a cloud server to give the user access to modify the data that pertains to him without the intervention of the main cloud, as this data can be obtained from the cloud server. As shown in figure 3. (Kunal, Saha, & Amin, 2019)

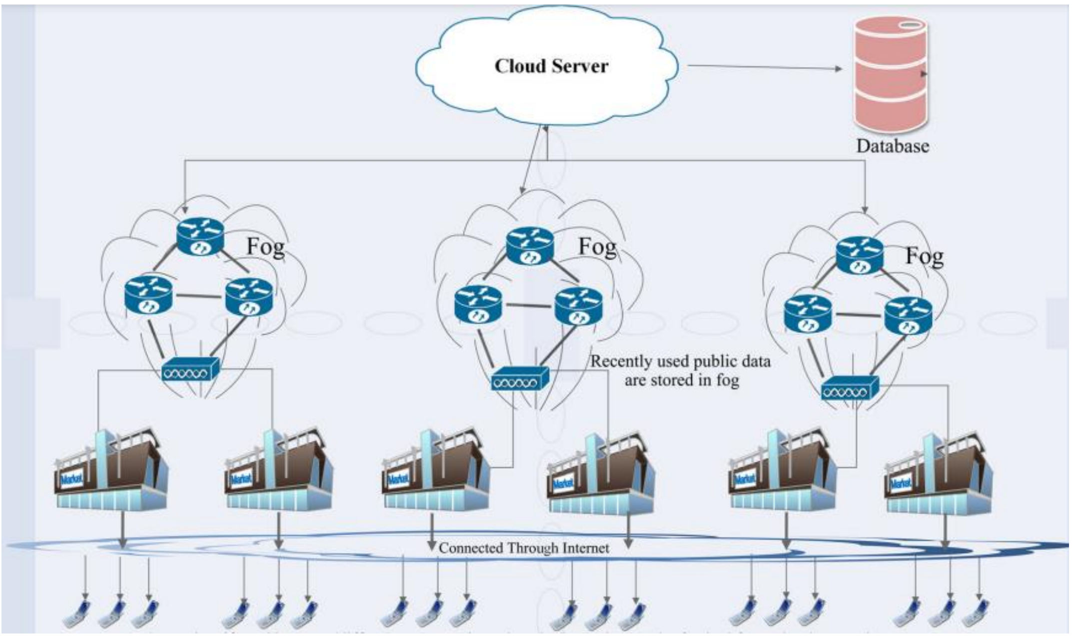

Figure 3 Fog node with cloud ***(Kunal, Saha, & Amin, 2019)***

Houshia et al. (2022) used the treated water to irrigate grains and legumes to examine its effect on protein, and they found it has a relatively small effect. Thus, treated wastewater can irrigate forage crops to obtain high-quality forage and reduce freshwater consumption. (Houshia, et al., 2022)

Integrating spatial data mining, GIS, and fuzzy logic provides a promising approach for extracting knowledge and optimizing the use of wastelands in agriculture. The findings significantly affect sustainable land management and various agricultural development and environmental conservation stakeholders. (Faridi, Verma, & Mukherjee, 2018)

Team (2000) used the R language to implement data mining algorithms for hepatitis prediction, for which they used K-means, K-medoids, Agglomerative, divisive, and fuzzy validations algorithms. (Team, 2000)

Chapter 4: Dataset And Data Pre-processing

This chapter focuses on the dataset used in the study. It provides a description and analysis of the dataset, highlighting its characteristics. The data pre-processing techniques employed are explained, including data cleaning and extracting, converting images to RGB format, and image resizing. Additionally, data augmentation techniques are discussed, such as increasing defect pattern frequencies and using a mixed defect patterns generator. The chapter concludes with a discussion on categorical encoding techniques.

4.1 Dataset Description and Analysis

Information on 1491 plant species was obtained from the Department of Genetic Resources Research (National Plant Gene Bank) at the Palestinian National Center for Agricultural Research. This data contains data in Arabic and English. This data consists of 36 columns, namely: sample number, collection number, collection institute code, species, genus, species, subspecies, local name, English name, sample condition, dangerous condition, date of collection, governorate, village, longitude, latitude, height (meters), source of collection, soil composition, degree of salinity, the utilization, parts used, traditional uses in Arabic, traditional uses in English, sample type, conservation type, description in Arabic, description in English, method of propagation, life cycle, structure, the environment in Arabic, the environment in English, agricultural pests in Arabic, agricultural pests in English. As shown in Table 1.

Table 1 Description features of National Plant Gene Bank.

#	Features	Description
1	Species	Contains plant species (Poaceae, Fabacea, Liliaceae, Rosaceae, … etc.).
2	Genus	Contains the genus of the plant (Triticum, Hordeum, Cicer, Vicia, … etc.).
3	Type	Contains the type of plant (durum, astivum, turgidum, vulgare, … etc.).
4	Subtype	Contains plant subtype (polymorpha, eliezeri, dicoccoides, … etc.).
9	crop name	Contains the name of the crop in Arabic (قمح، شعير، حمص، فول، ... الخ.).
10	local name	Contains the local name of the crop in Arabic (ناب الجمل, كحلا, نورسي, ... الخ.).
11	English name	Contains the name of the crop in English (wheat, barley, alfalfa, … etc.).
12	sample case	Contains sample case (cultivar, wild, landraces/local variety, … etc.).
13	dangerous condition	It indicates whether the sample is at risk or not.

14	date of collection	Contains the date of collection (2007-2020).
15	Governorate	Contains the name of the governorate from which it was collected (Jenin, Ramallah, Tubas, … etc.).
16	Village	Contains the name of the village in the governorate from which it was collected (Tammon, Qabatya, Tayaseer, … etc.).
17	Longitude	It shows the longitude of the area from which the sample was collected.
18	Latitude	It shows the latitude of the area from which the sample was collected.
19	height (meters)	It shows the height above sea level of the area from which the sample was collected.
20	source of collection	Shows the nature of the area from which the sample was taken (crops, orchard, store, fallow land, … etc.).
21	soil composition	Contains the soil composition (rocky, clay, calcareous soil ... etc.).
22	degree of salinity	Contains the degree of soil salinity (low, medium, and high).

23	The utilization	Contains plant uses in general (food, fodder, medicinal, wood … etc.).
24	parts used	It contains the parts that are used (flowers, stem, leaves, seeds … etc.).
25	traditional uses in Arabic	Contains the traditional uses of the plant in Arabic (مشغولات يدوية بالقش، أطباق شعبية، غذاء ... الخ.).
26	traditional uses in English	Contains the traditional uses of the plant in English (Straw handicrafts, traditional dishes, food...etc.).
27	sample type	Shows the type of sample that was taken (often contains seeds).
28	conservation type	Shows the type of sample that was preserved (often contains seeds).
29	description in Arabic	Contains a brief description of the plant in Arabic (قائم، مفترش، زاحف ... الخ.).
30	description in English	Contains a brief description of the plant in English (Erect, Extended, Prostrate … etc.).
31	method of propagation	Contains plant propagation method (often contains seeds).

32	life cycle	Contains plant life cycle (annual, perennial, biennial … etc.).
33	Structure	Contains plant structure (herbaceous, arboreal, climber, creeper … etc.).
34	environment in Arabic	Contains the environment of the plant from which it was collected in Arabic (حقل زراعي، محطة ابحاث زراعية، ارض هامشية ... الخ.).
35	environment in English	Contains the environment of the plant from which it was collected in English (Field, Research station, Marginal habitat … etc.).
36	agricultural pests in Arabic	Contains diseases and pests that affect plants in Arabic (الأصداء، التفحمات، الذبول ... الخ.).
37	agricultural pests in English	Contains diseases and pests that affect plants in English (Rusts, Smuts, Fusarium wilt … etc.).

4.2 Data Preprocessing

The data taken is not pure, so it needs to be reconstructed. The data is reconstructed in several steps: Converting language, Inserting the location if needed, Data cleaning, Data selection, and Converting extension of the data file.

4.3 Converting language

This data contains data in Arabic that needs to be converted to English, so we have converted the language.

We used Python code to convert the Excel file columns in Arabic to English so that the data can be used locally and globally. as shown in the figure 4. When we converted the language using Python, we had a problem with understanding the Arabic language incorrectly, so we converted it manually.

```python
import pandas as pd
from translate import Translator
def translate_arabic_to_english(text):
    translator = Translator(to_lang="en", from_lang="ar")
    try:
        translation = translator.translate(text)
        return translation
    except Exception as e:
        print(f"Translation error: {e}")
        return text  # If translation fails, return the original text
def translate_selected_columns(input_file, output_file, sheet_name, columns_to_translate):
    # Read the specified sheet from the Excel file
    df = pd.read_excel(input_file, sheet_name=sheet_name)

    # Translate the specified columns
    for column_name in columns_to_translate:
        if column_name not in df.columns:
            print(f"Column '{column_name}' not found in the Excel file.")
            continue
        df[column_name] = df[column_name].apply(translate_arabic_to_english)
    # Write the translated data to a new Excel file
    df.to_excel(output_file, index=False)
if __name__ == "__main__":
    input_file = "na.xlsx"  # Replace with the path to your input Excel file
    output_file = "output_file.xlsx"  # Replace with the desired output Excel file path
    sheet_name = "Sheet1"  # Replace with the name of your sheet
    columns_to_translate = ["crop_name", "local_name", "village", "source_of_collection", "soil_composition", "degree_of_salinity",
                            "The_utilization", "parts_used", "sample_type", "conservation_type", "method_of_propagation", "life_cycle",
                            "structure"]# Replace with the names of your columns

    translate_selected_columns(input_file, output_file, sheet_name, columns_to_translate)

```

Figure 4 Converting language

4.4 Data cleaning and selection

This data contains data that cannot be used, so we eliminated it. So, we have 1491 items and when we cleaned them, we received 821 items. That is because many reasons for eliminating them, the most important of them is we cannot detect the location of the data.

4.5 Data cleaning and selection

This data contains data that cannot be used, because of many causes. For example, the data is not complete. So, we eliminated it. And select the spices feature as a basic feature to show and process because there is a minimum number of items.

Chapter 5: Methodology

The system will be divided into two sub-systems to facilitate the explanation of the system. Due to the large number of systems associated with this system, it is difficult to explain, so we will explain a new system of irrigation and fertilization as a sub-system that can be applied to all of these systems. We will also explain the establishment of a geographical database for the Palestinian Seed Bank as part of the main system. So, the system will be divided as follows:

5.1 Central Geo-database creation

The purpose of this research is to use the data available in the National Agricultural Research Centre of the Ministry of Agriculture to solve the problem of the lack of a central database and the problem of data loss to obtain a clear visualization of the data. Because some of the data contain spatial data, we have created a central geographical database to obtain as much benefit as possible from this data. Figure 5 shows the general perception of the research.

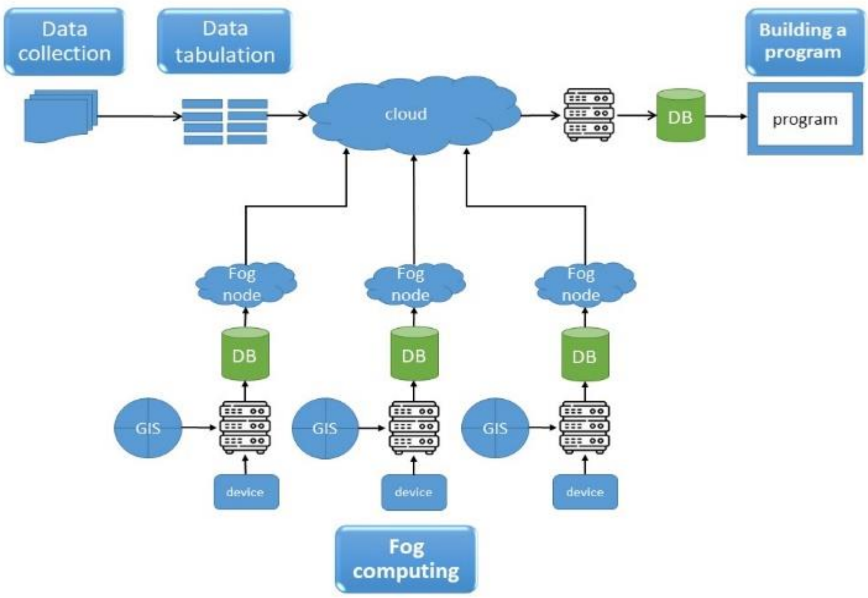

Figure 5 Central Geo-database module

To create a central geo-database, the following steps must be followed will be completed:

1. **Connect Devices to Servers**: All devices everywhere will be connected to the servers to get the data.
2. **Save Location and Map Regions**: The server will save the location of all regions and map it.
3. **Save Data in Real-time Database**: The server will save the data in a real-time database.
4. **Regulate Database on Fog Node**: The database will be regulated on the fog node.
5. **Protect Data on Cloud**: The fog node will protect the data on the cloud.
6. **Collect Rubbish Data**: Rubbish data will be collected from any location in the enterprise.
7. **Clean and Tabulate Data**: Collected data will be cleaned and tabulated to show in the database and saved in the cloud.
8. **Save Data in Cloud**: The cloud will save the main server and database data.
9. **Access Data via Website**: The data in the database can be accessed using a website that provides all required statistics.

5.2 Fertilizer and irrigation system (FIS)

Water in Palestine and the Middle East is in danger, so we built a model for using treated wastewater to irrigate crops with the optimal use of fertilizers necessary for the soil. In this model, we used the technology of the Internet of Things, cloud computing, and fog computing, and we linked it to the geographical location of the sensors used to know whether they are valid. Figure 6 illustrates this.

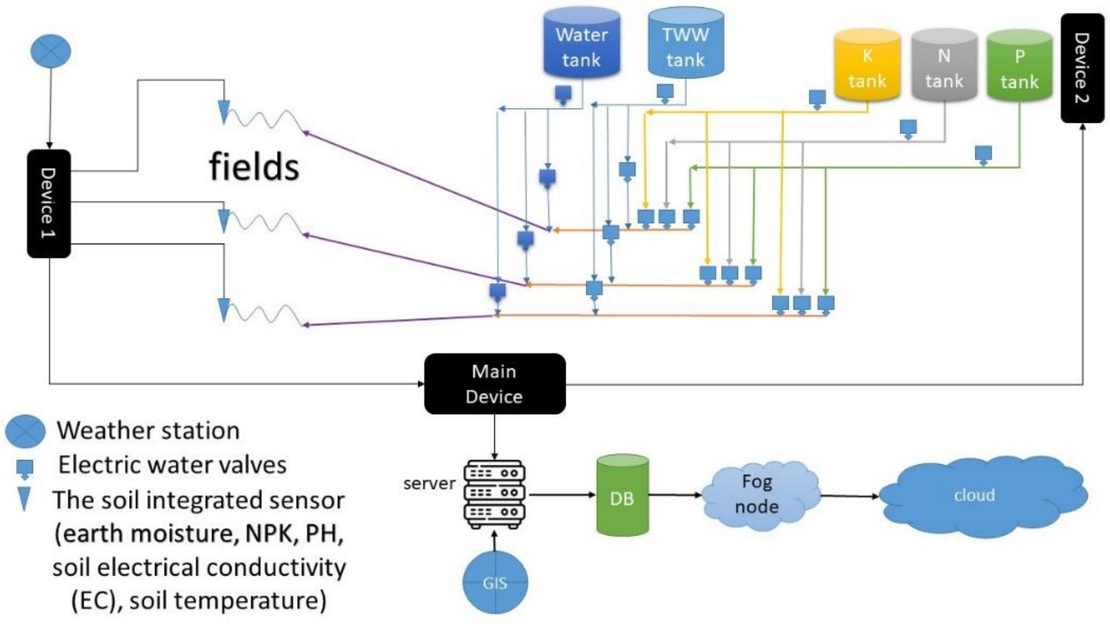

Figure 6 FIS Module

A fertilizer and irrigation system will be created. The following steps will be completed:

1. **Soil-integrated Sensors**: Measure N, P, K, pH, soil moisture, temperature, and electrical conductivity (EC) in the crop.
2. **Weather Station**: Get weather data.
3. **Connect Sensors and Weather Station**: All soil-integrated sensors and the weather station will be connected to the device in the field to read all data.
4. **Send Data to Main Device**: The data will be sent to the main device to decide the amount and type of water and fertilizer.
5. **Analyze Data and Control Valves**: Based on the analyzed data, the main device will give orders to another device that can control the valves of fresh water, TWW, and liquid fertilizer (N, P, and K).
6. **Save Data on Server**: The main device will save the data on the server.
7. **GIS Mapping**: The GIS will save the location of sensors and fields and map it in the same server to maintain location transparency.

8. **Regulate Data in Real-time Database**: The server will regulate the data in a real-time database to maintain concurrency transparency.
9. **Save Database on Fog Node**: The database will be saved on a fog node to add another fog node.
10. **Connect Fog Node to Main Cloud**: The fog node will connect to the main cloud.
11. **Represent Data in Cloud Program**: The cloud will be represented as a program to show the data.

5.3 Data and Fuzzy Logic

Fuzzy logic is one of the Expert systems mechanisms used in cases of uncertainty to make essential decisions about machines. Because the system was not implemented and there was more than one attribute, we used Fuzzy logic to take the appropriate action in irrigation cases based on rules from experts in this field. (Faridi, Verma, & Mukherjee, 2018)

The percentage of fertilization varies for each field based on the type of crop and the type of soil, so orders are given to the fertilization valves accordingly, and they do not need any artificial intelligence mechanisms to apply them. (Li, Chen, Zeng, & Ye, 2013)

5.3.1 Data structure

The data should be structured in the database in a regular shape. So, each sensor should have a unique number called sensor ID to differentiate between them, N, P, and K values that have a float number between (1~1999 mg/kg(mg/L)), PH value that has a float number between (0~14PH), EC value that has a float number between (0-20mS/cm), earth moisture value that has an integer number between (0~100%), soil temperature value that has an integer number between (-40~80℃), and a region value which have decimal degrees. As shown in Table 2.

Table 2 Attributes description and range.

Attribute name	Attribute description	Range
Soil Temperature	Indicates heat flux in the soil and heat exchanges between the soil and atmosphere.	-40~80℃
Soil Moisture	a measure of water stored in the soil.	0~100%
EC	Indicates the ability of soil to conduct (transmit) or attenuate electrical current.	0~20mS/cm
NPK	a measure of the amount of nitrogen, phosphorus, and potassium in fertilizer	1~1999 mg/kg(mg/L)
Soil PH	a measure of the acidity or alkalinity of the soil.	0~14PH

Fuzzy logic will be used to decide between turning on/off the water valve, so we build rules of the attributes that affect irrigation in MATLAB using the Fuzzy function.

5.3.2 Input data

The attributes should be divided separately to detect the decision of valve status. As following:

5.3.2.1 Soil temperature

The soil temperature can be divided into three levels. Low takes a range number between (-40~10), a medium that takes a range number between (8~32), and a high that takes a range number between (30~80). Figure 7 shows that in MATLAB.

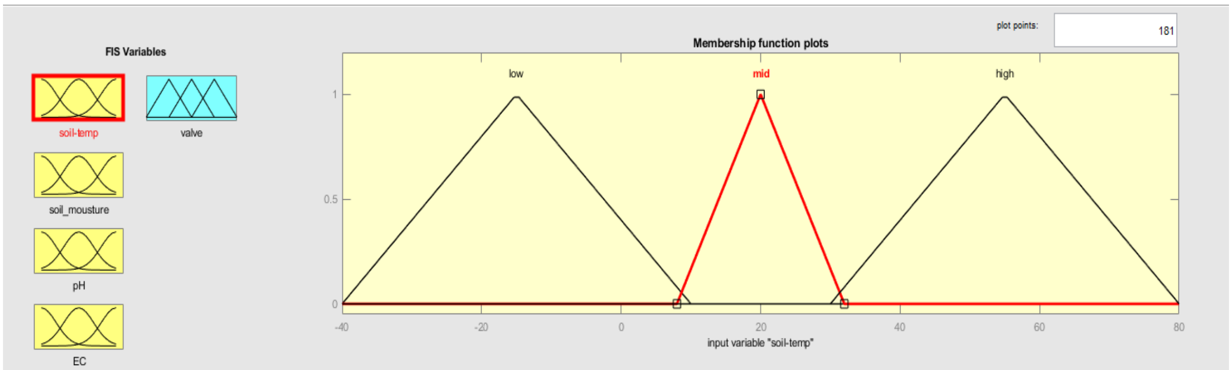

Figure 7 Soil temperature range divided

5.3.2.2 Soil moisture

The soil moisture can be divided into two levels. Low takes a range number between (0~75) and usually takes a range number between (65~100). Figure 8 shows that in MATLAB.

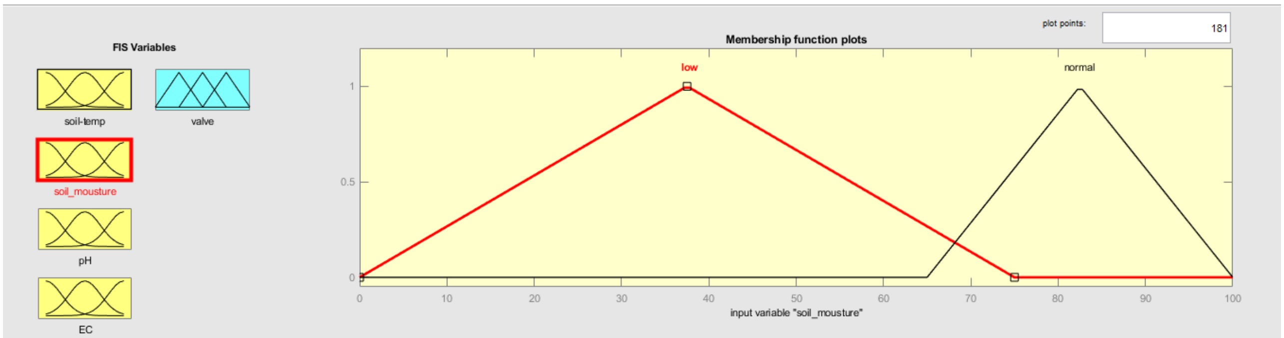

Figure 8 Soil moisture range divided

5.3.2.3 PH

The PH can be divided into four levels. acidic takes a range number between (0~5.5), a neutral takes a range number between (5.5~8), a slightly alkaline takes a range number between (8~9), and a highly alkaline takes a range number between (9~14). Figure 9 shows that in MATLAB.

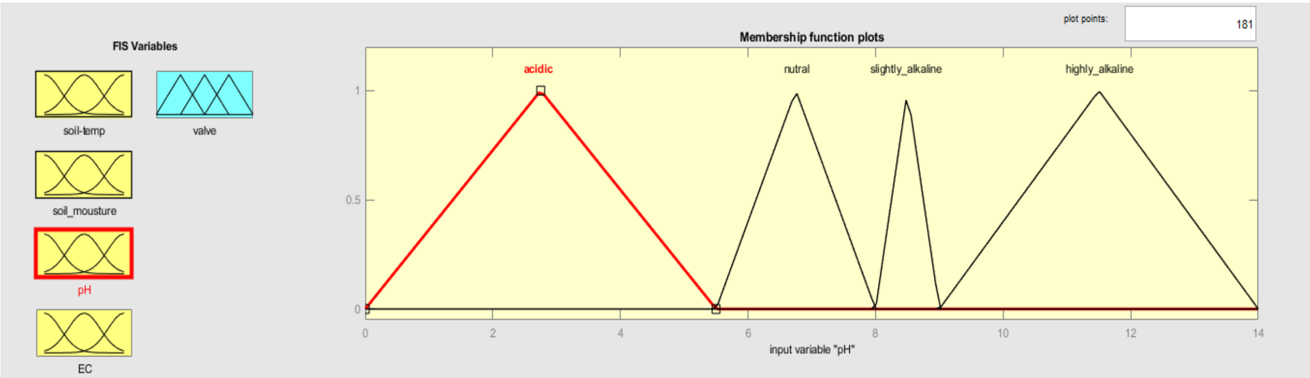

Figure 9 PH range divided

5.3.2.4 EC

The EC can be divided into three levels. Non-saline takes a range number between (0~5), slightly saline takes a range number between (4~8), and strongly saline takes a range number between (7~20). Figure 10 shows that in MATLAB.

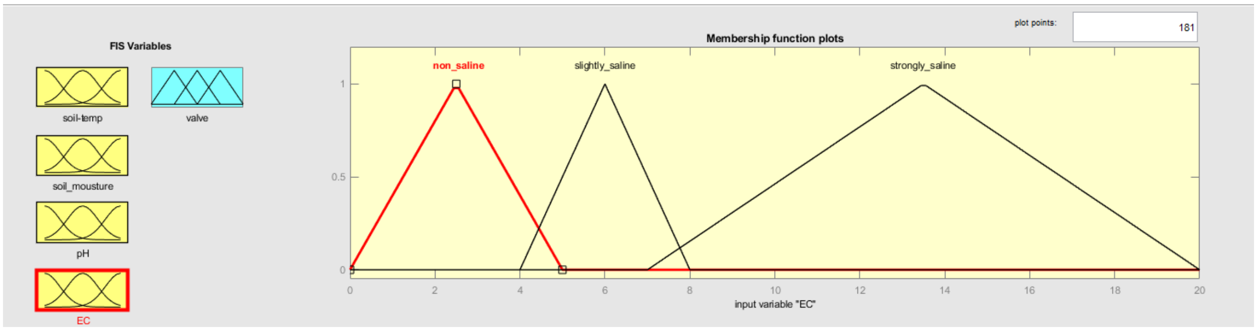

Figure 10 EC range divided

5.3.3 Output data

The output data can be divided into two levels. Valve off takes a range number between (0~0.6), and valve on takes a range between (0.4~1). Figure 11 shows that in MATLAB.

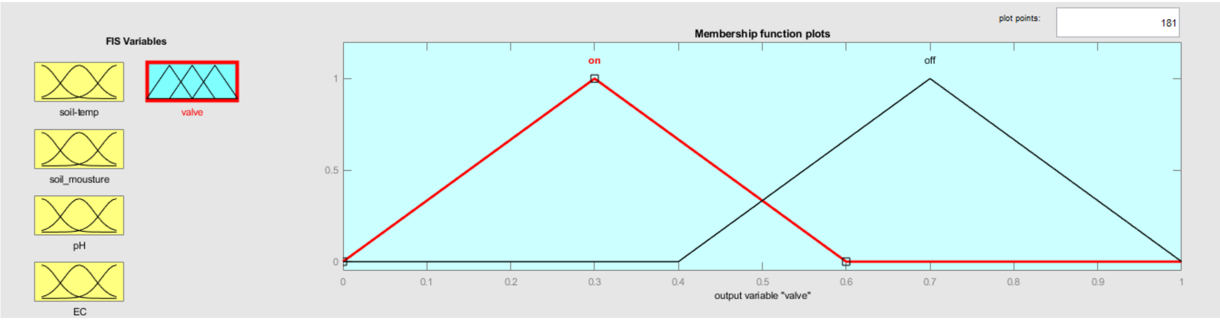

Figure 11 Output range divided

5.3.4 Building rules

There are nonresearched merges between all the attributes used. So, the roles build depend on experts in water and soil from the Ministry of Agriculture in Palestine. The rules in MATLAB are built as shown in Figure 12.

The number of rules equals multiplying the number of divisions in each attribute, as shown in equation 1.

$$Number\ of\ rules = 3 * 2 * 4 * 3 = 72\ rule \quad (1)$$

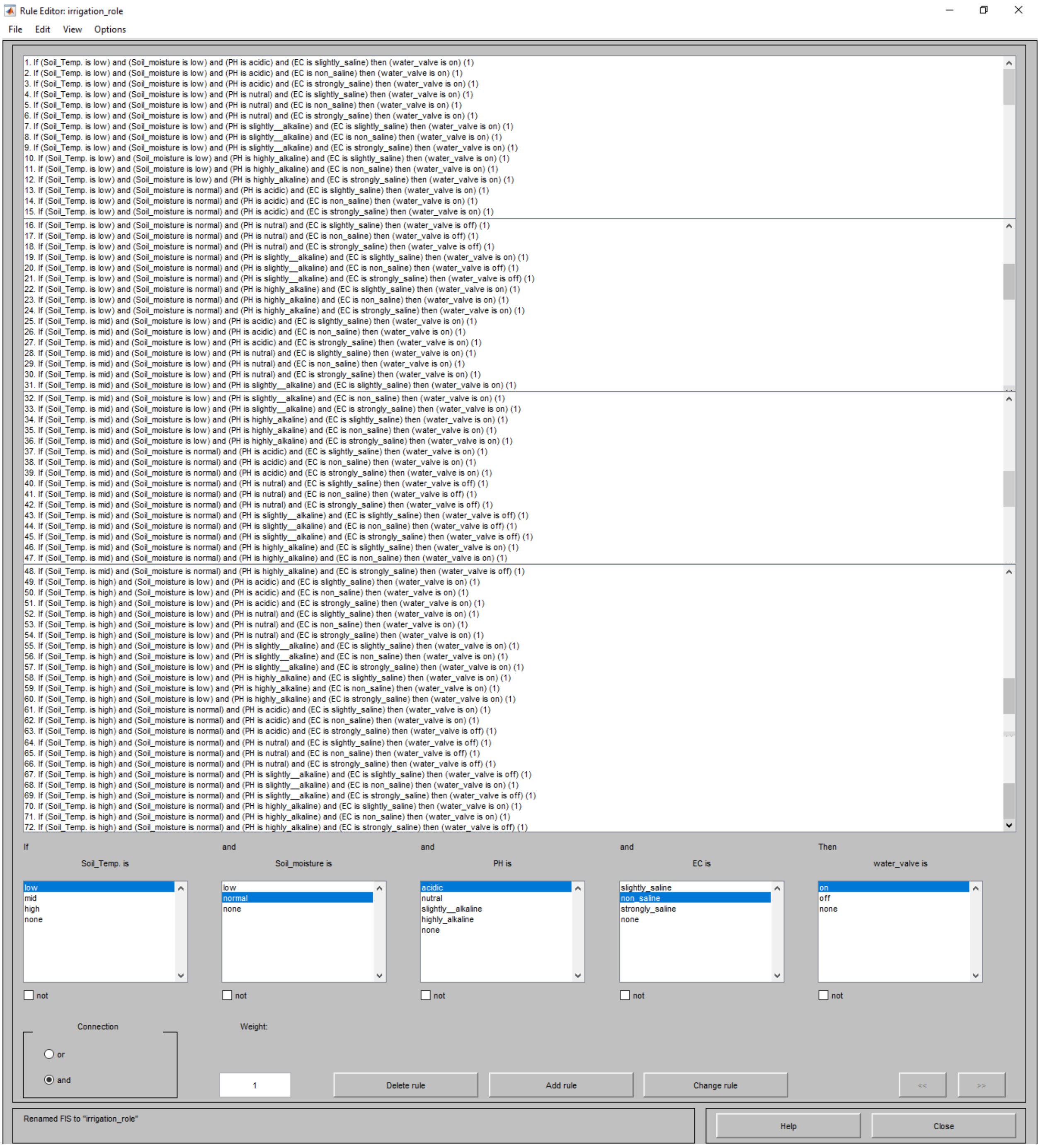

Figure 12 Building rule in MATLAB

5.4 ArcGIS building with machine learning

ArcGIS, developed by Esri, is a powerful geographic information system (GIS) software suite used for creating, managing, analysing, and visualizing spatial data. With its extensive tools and capabilities, ArcGIS empowers users to understand the geographic context of their data, make informed decisions, and solve complex problems related to location and geography.

One of the standout features of ArcGIS is its integration with Python, a versatile programming language widely used in data science, geospatial analysis, and automation tasks. This integration opens up endless possibilities for extending ArcGIS functionalities, automating repetitive tasks, and building custom geoprocessing workflows.

In this introduction to ArcGIS with Python, we will explore the fundamentals of utilizing Python scripting within the ArcGIS environment. From basic data manipulation to advanced spatial analysis, this integration offers a seamless and efficient way to leverage the power of both ArcGIS and Python for spatial data management and analysis tasks. Whether you're a GIS professional, a researcher, or a developer, mastering ArcGIS with Python can greatly enhance your ability to extract valuable insights from spatial data and streamline your workflows.

5.4.1 Showing the data point in the map

To show the point in the map we should use the display XY Data form. So, we enter the x field as longitude and the y field as latitude. Because of the data in degree, we should use the Geographic Coordinate System: GCS_WGS_1984. As shown in Figure 13**.**

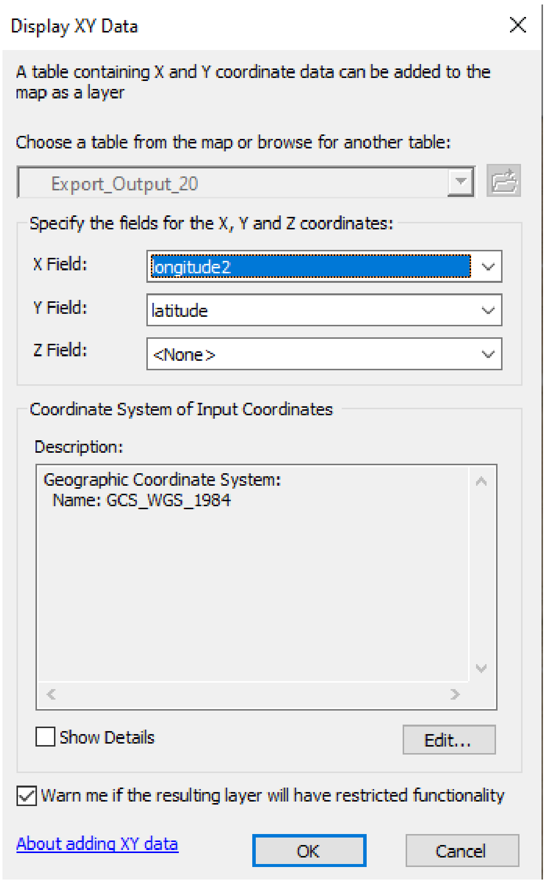

Figure 13 Display XY Data

5.4.2 Categorize point

To distinguish the species of seed we use the symbology properties – Categories – Unique values and add all values of species and distinguish them with colour Ramp. As shown in Figure 14.

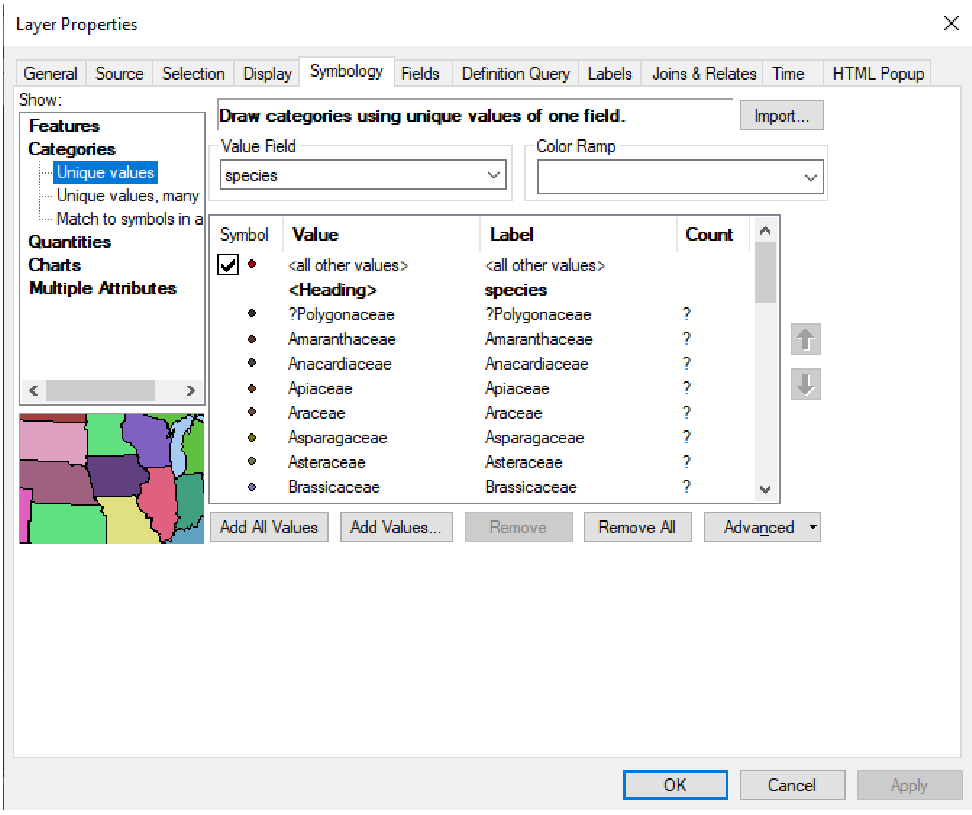

Figure 14 categories of all species

5.4.3 Summarize the attributes

We show this is many attributes for study. So, we need to decrease them. To make this we summarize the species using a tool named summarize. As shown in Figure 15. The summarized tool output table contains species and counts of them. As shown in Figure 16.

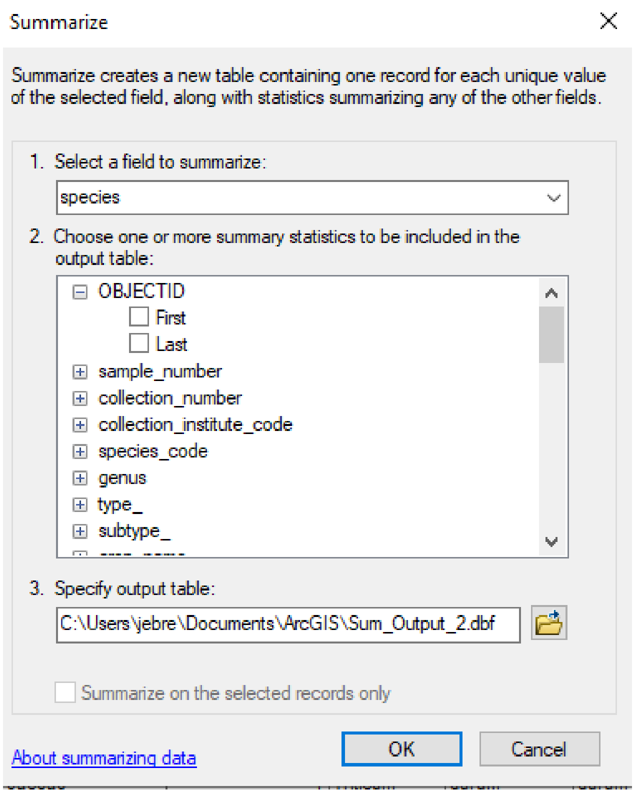

Figure 15 Summarize tool

	species	Count_species
0	?Polygonaceae	1
1	Amaranthaceae	1
2	Anacardiaceae	26
3	Apiaceae	11
4	Araceae	1
5	Asparagaceae	3
6	Asteraceae	105
7	Brassicaceae	3
8	Caesalpiniaceae	8
9	chenopodiaceae	1
10	Chenopodiaceae	3
11	Cistaceae	3
12	Cucurbitaceae	93
13	Ephedraceae	1
14	Ericaceae	3
15	Euphorbiaceae	1
16	Fabaceae	132
17	Fagaceae	5
18	Lamiaceae	24
19	Lauraceae	2
20	Leguminosae	5
21	Liliaceae	3
22	Malvaceae	8
23	Moringaceae	1
24	Oleaceae	3
25	Papilionaceae	98
26	Pedaliaceae	51
27	Pinaceae	1
28	Poaceae	153
29	Ranunculaceae	3
30	Rhamnaceae	6
31	Rosaceae	40
32	Santalaceae	1
33	Smilacaceae	2
34	Solanaceae	10
35	Styracaceae	7
36	Tamaricaceae	1

Figure 16 summarizes the species table

5.4.4 Selection attribute and make it layers

We select the maximum 5 counts of species named: Asteraceae, Cucurbitaceae, Fabaceae, Papilionaceae, and Poaceae. We use the selection tool to select attributes selected each one alone. As shown in Figure 17. After selecting the attribute, we make it a layer. As shown in Figure 18.

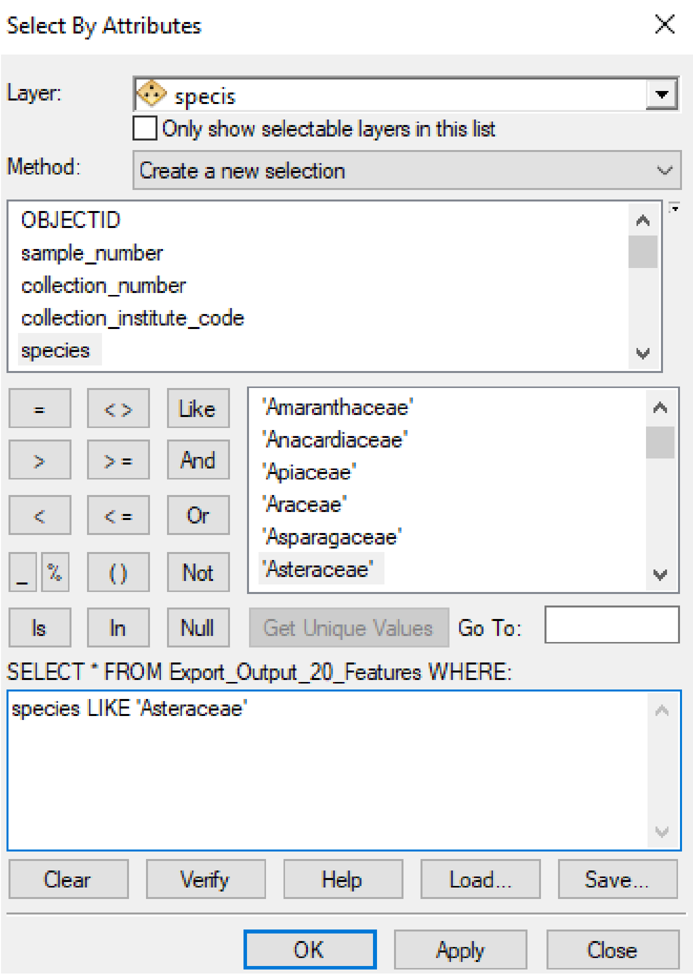

Figure 17 Selection attribute

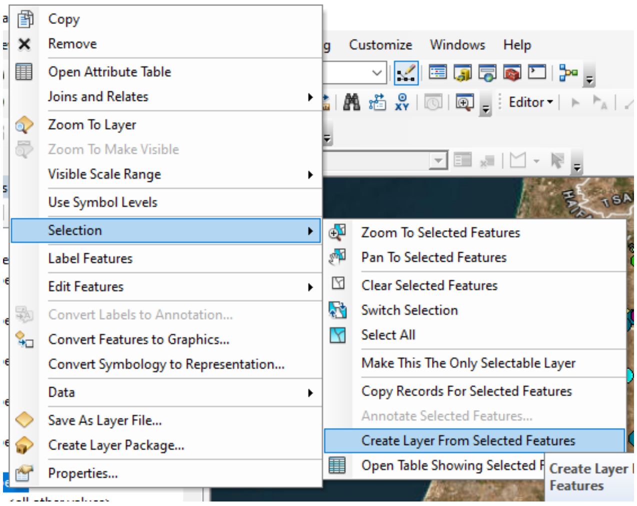

Figure 18 making selection as a layer

5.4.5 Another data

We have a data of 7 weather stations in West-Bank this data include rain fall, temperature, and the heights of the stations.

In another hand; we have a raster data of the heights of all the west bank.

5.4.6 Select the months of rain fall and temperature and summarize them

We select the top five months of rain fall and temperature that effect of growing the plants, that is from month 11 to month 3. As shown in figure 19. And summarize the sum value of rainfall of stations that collected in the selected year months of the stations. As shown in figure 20. And summarize the average value of temperature that collected in the selected year months of the stations. As shown in figure 21.

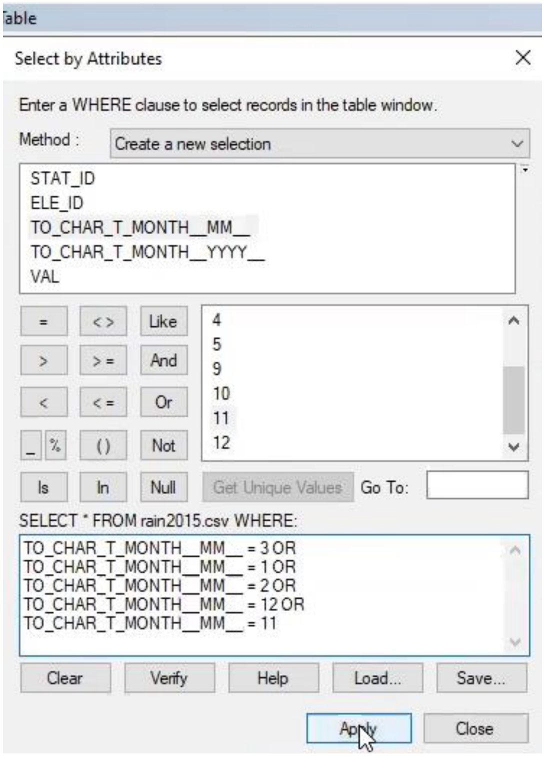

Figure 19 selection months

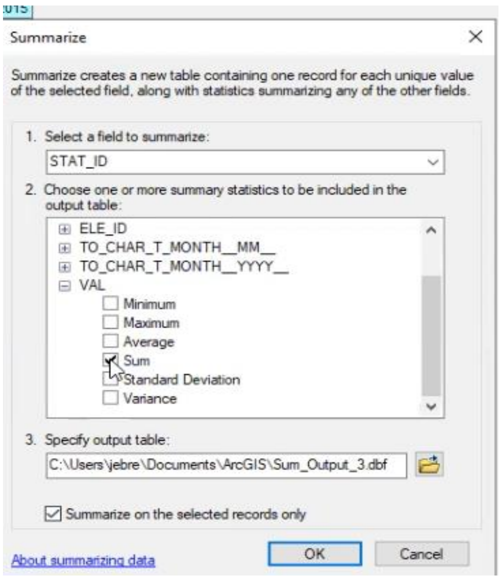

Figure 20 Summarize the sum value of rainfall of stations

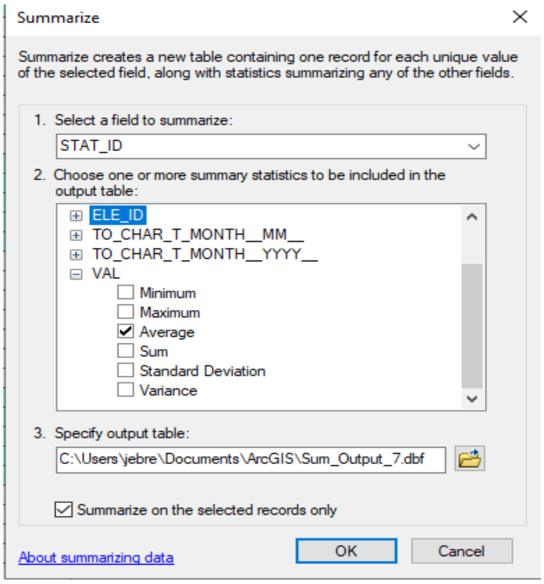

Figure 21 Average temperature

5.4.7 Join data from stations and sum rainfall and average temperature

We join sum rainfall values to station table data. As shown in figure 22. And we join average temperature values to station table data. As shown in figure 23.

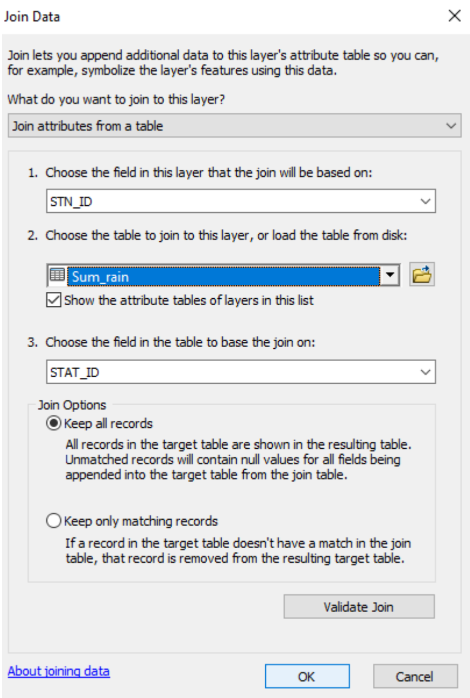

Figure 22 join sum rainfall

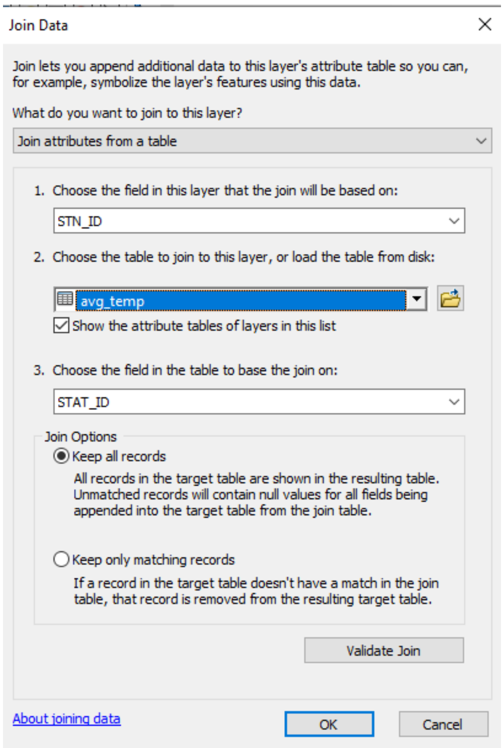

Figure 23 join average temperature

5.4.8 Marge selected species and Extract value to point

To collect the species, we use marge tool. As shown in figure 24. To add raster value, we extracted the merge spices with raster file of Westbank. As shown in figure 25.

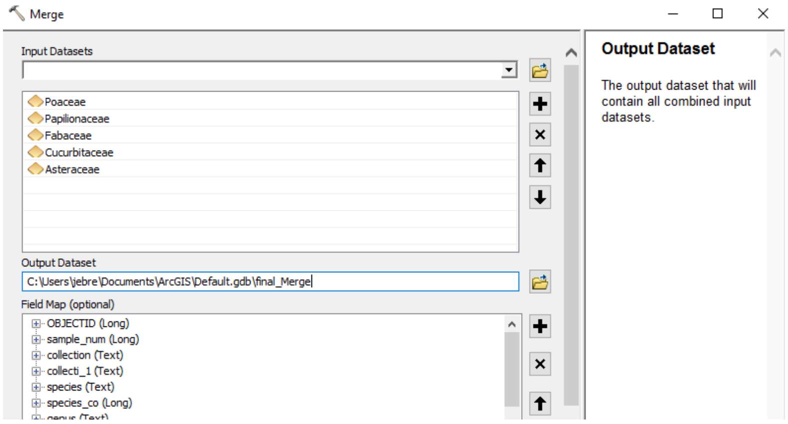

Figure 24 Marge selected species

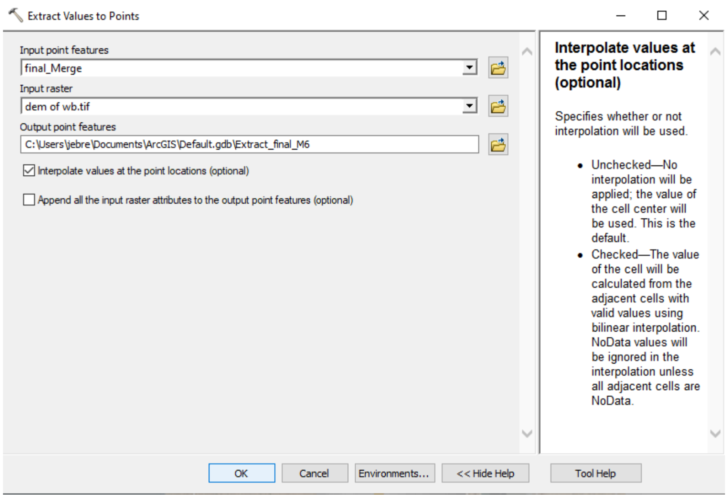

Figure 25 Extract value to point

5.4.9 Regression in excel

Excel program have tool of regression so we used it to predict value of sum rainfall and average temperature depend on raster value. As shown in figure 26. The regression result of sum rainfall is significant (0.0186 < 0.05). As shown in figure 27. Also, the regression result of average temperature is significant (0.0139 < 0.05). As shown in figure 28.

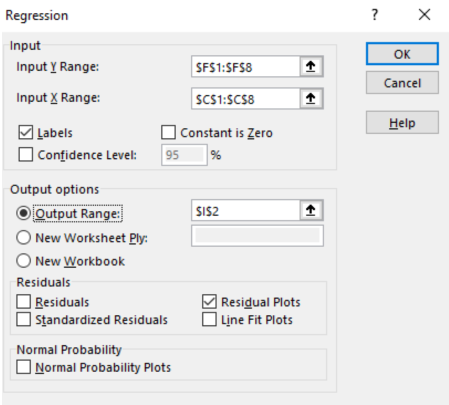

Figure 26 Regression in excel

SUMMARY OUTPUT			Y = B0 + B1 * X					
Regression Statistics								
Multiple R	0.837831943							
R Square	0.701962365							
Adjusted R Square	0.642354838							
Standard Error	113.6272404							
Observations	7							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	152046.9255	152046.9255	11.77640477	0.018601816			
Residual	5	64555.74878	12911.14976					
Total	6	216602.6743						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	331.8131806	64.19048771	5.169195505	0.003557272	166.8062788	496.8200823	166.8062788	496.8200823
RASTERVALU	0.481734127	0.140378647	3.43167667	0.018601816	0.120879328	0.842588927	0.120879328	0.842588927

Figure 27 regression of sum rainfall

Regression Statistics								
Multiple R	0.856105234							
R Square	0.732916171							
Adjusted R Square	0.679499406							
Standard Error	1.987347205							
Observations	7							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	54.19063152	54.19063152	13.72071412	0.013939256			
Residual	5	19.74774456	3.949548912					
Total	6	73.93837607						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	21.23367086	1.122695455	18.91311732	7.6132E-06	18.34769031	24.1196514	18.34769031	24.1196514
RASTERVALU	-0.009094539	0.002455231	-3.704148231	0.013939256	-0.015405911	-0.00278317	-0.015405911	-0.00278317

Figure 28 regression of average temperature

5.4.10 Build the equation of rainfall and temperature prediction

We can use the regression to build an equation using the general low. (Seber & Lee, 2012) As shown in **equation (2)**. Depend on the general low we build an equations of sum rainfall and average temperature prediction. As shown in **equations (3) & (4)**.

$$Y = B0 + B1 * X \tag{2}$$

$$RAINFALL\_PRE = 331.8131806 + 0.481734127 * RASTERVALU \tag{3}$$

$$TEMPRETURE\_PRE = 21.23367086 - 0.009094539 * RASTERVALU \tag{4}$$

5.4.11 Add field of sum rainfall and average temperature

We add two field to calculate the sum rainfall and average temperature prediction to table of selected species. As shown in figure 29.

Figure 29 Add field

5.4.12 Calculate the sum rainfall and average temperature prediction

We used the field calculator feature in ArcGIS to calculate the sum rainfall and average temperature prediction of all selected species depend on regression equations. As shown in figure 30 and 31.

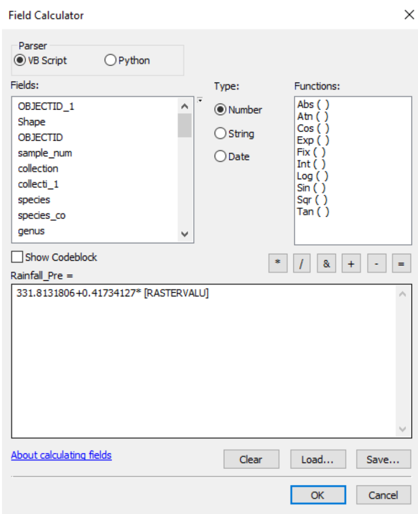

Figure 30 calculate the sum rainfall

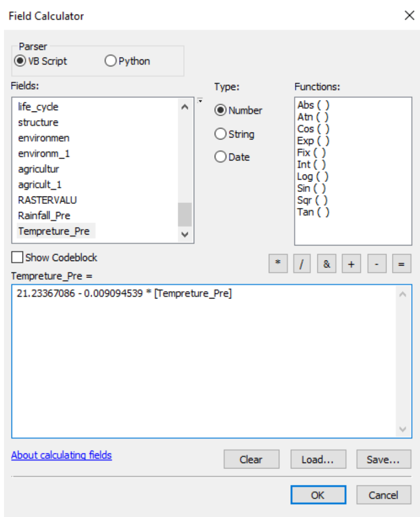

Figure 31 calculate average temperature

5.4.13 Build python code of prediction model

We used the python code to predict the spices of plant, using feature height, rainfall, and temperature. We used SVM, Decision tree, Random Forest, and Nural Network machine learning to prediction. We build a code and tuning the parameter of all machine learning and get the following accuracy as shown in figure 32. The accuracies are very bad, so we need to add feature.

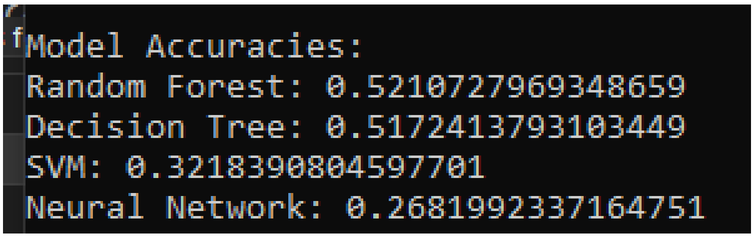

Figure 32 Prediction model accuracies

5.4.14 Adding new feature

We get a geology of west bank data. And add the data in ArcGIS. As shown in figure 33. We joined the data based on geology Westbank data. As shown in figure 34.

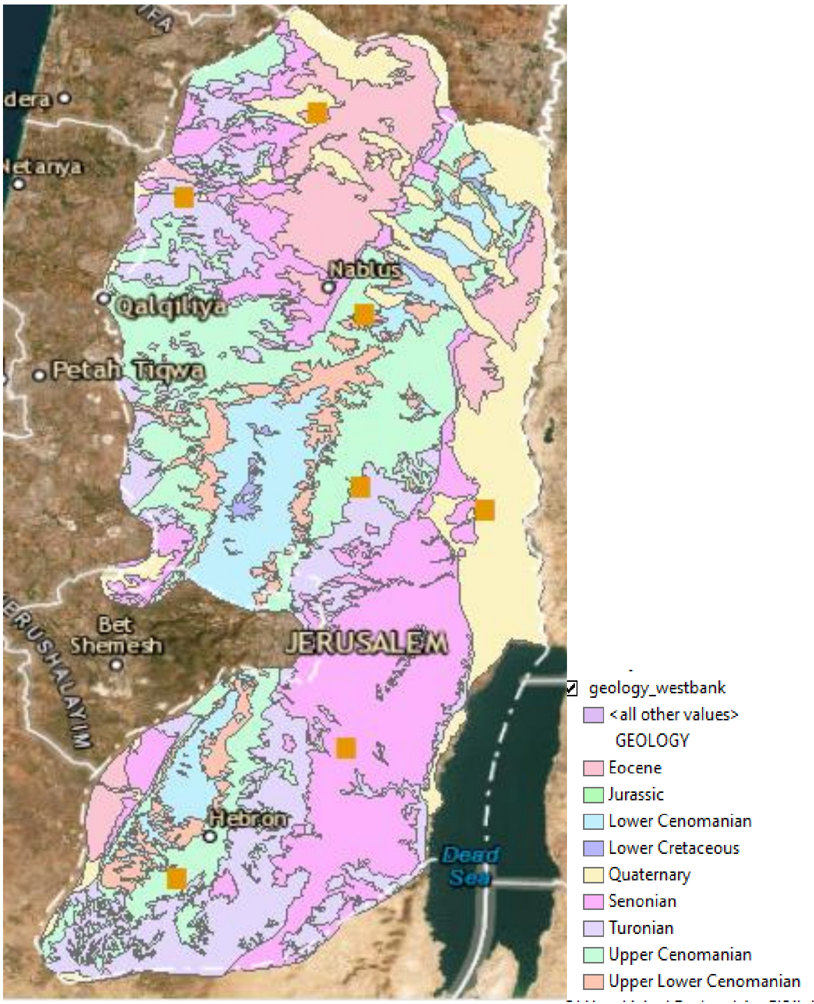

Figure 33 Geology data

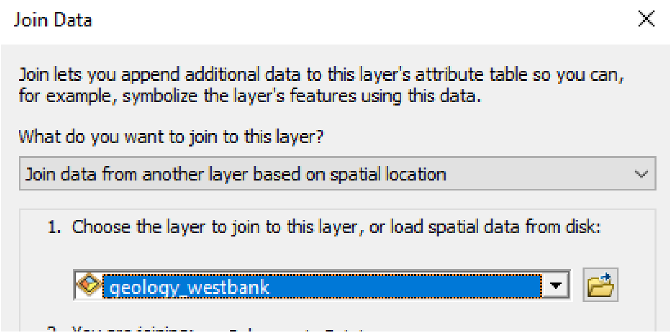

Figure 34 joined geology data

5.4.15 Rerun code

We rerun the python code after add new feature (geology) and we get better accuracies. The best accuracy is 0.67 in random forest, so we choose it. as shown in figure 35.

```
Model Accuracies:
Random Forest: 0.6704980842911877
Decision Tree: 0.6551724137931034
SVM: 0.24904214559386972
Neural Network: 0.39080459770114945
```

Figure 35 model accuracies after add new feature

5.4.16 Retuning the parameter of random forest

We using the code in figure 36 to select the best parameter give the best performance in random forest. The result give n_estimator (60), max_feature (sqrt), max_debth (12), criterion (entropy), min_samples_split (2), and min_samples_leaf (1). And we get better accuracy (0.70).

```
# Define a parameter grid for GridSearchCV with more parameters
param_grid = {
    'n_estimators': [10,20,30,40,50,60,70,80,90,100],
    'max_features': ['auto', 'sqrt', 'log2'],
    'max_depth': [10,12,15,20,25,30,40,45,50],
    'criterion': ['gini', 'entropy'],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}

# Use GridSearchCV to find the best parameters
grid_search = GridSearchCV(estimator=clf, param_grid=param_grid, cv=kf, n_jobs=-1, verbose=2)
grid_search.fit(X_resampled, y_resampled)
```

Figure 36 tuning parameter of random forest

5.4.17 Code of select random state

To select the best random state, we use the following code in python. As we use from 0 to 10000 test random state and choose the best random state to get the best accuracy. As shown in figure 37. The code run successful and give the better result (0.72) when random state (208). as shown in figure 38.

```python
# Variables to store the best random state and its corresponding accuracy
best_random_state = None
best_accuracy = 0.0

# Try different random states and evaluate model performance
for random_state in range(10000):
    X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled, test_size=0.35, random_state=random_state)

    grid_search = GridSearchCV(estimator=clf, param_grid=param_grid, cv=kf, n_jobs=-1)
    grid_search.fit(X_train, y_train)

    best_clf = grid_search.best_estimator_
    y_pred = best_clf.predict(X_test)
    accuracy = metrics.accuracy_score(y_test, y_pred)

    if accuracy > best_accuracy:
        best_accuracy = accuracy
        best_random_state = random_state

print("Best random state:", best_random_state)
print("Best accuracy:", best_accuracy)
```

Figure 37 select random state code

```
Best random state: 208
Best accuracy: 0.7203065134099617
Press any key to continue . . .
```

Figure 38 Accuracy of selected random state

5.4.18 Create Fishnet

To detect all the region which contain the species or what species can grow in the region, we use the fishnet tool to distribute point in all region of Westbank. As shown in figure 39. and we use the extraction value to point tool. As shown in figure 40.

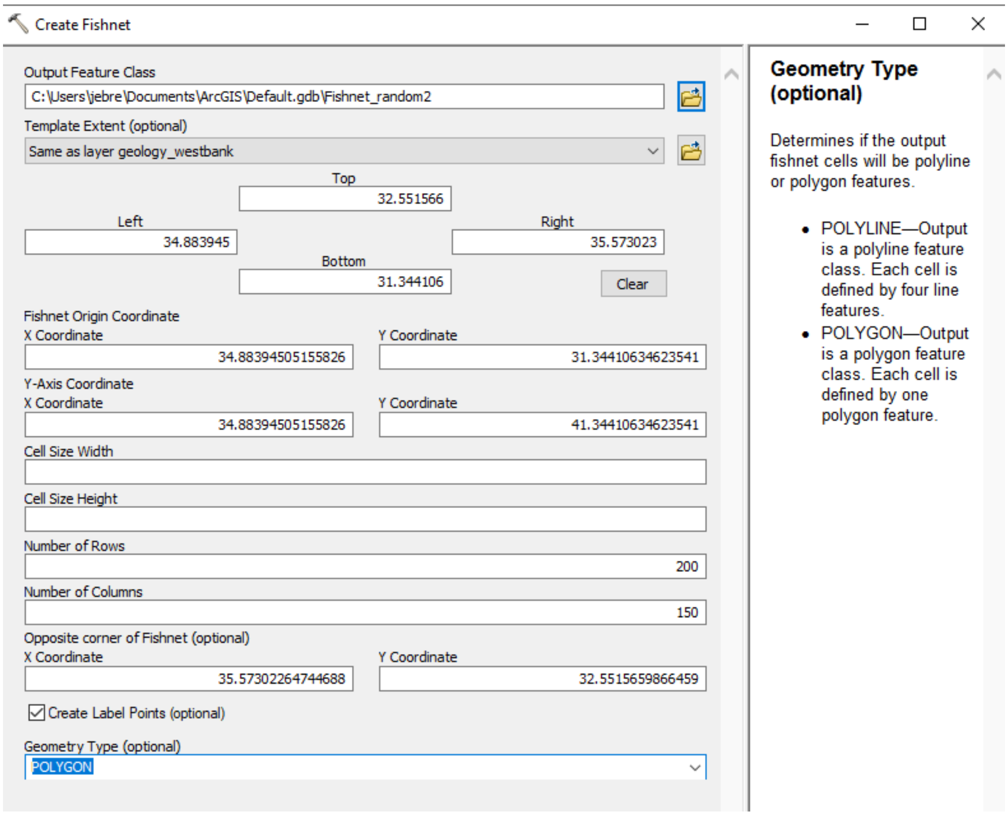

Figure 39 Create Fishnet

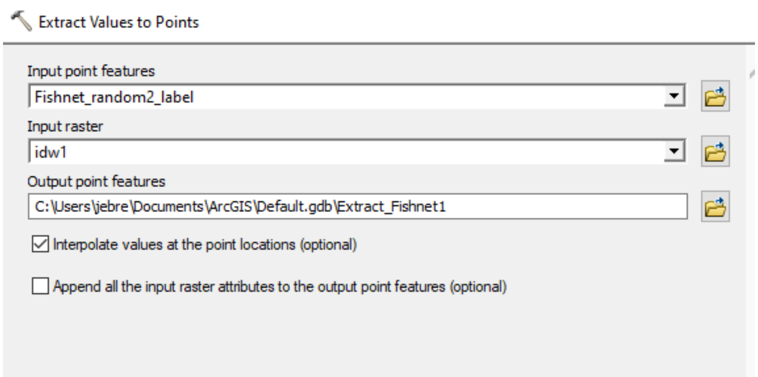

Figure 40 Extract value to point Fishnet

5.4.19 Generate new data

We add the height value and predicted rainfall and temperature of all fishnet point and add a geology type of them. Then we export the data to excel sheet.

5.4.20 Prediction model of species type

We add a python code to predict the species type using the previse code. As shown in figure 41.

```
# Load new data for prediction
new_file_path = r'C:\Users\jebre\Desktop\GIS-DataMining\Newfolder\newdata.xlsx'
new_sheet_name = 'Sheet1'

# Read new data from Excel file
new_df = pd.read_excel(new_file_path, sheet_name=new_sheet_name)

# Print the first few rows to verify new data loading
print("First 5 rows of the new data:")
print(new_df.head())

# Check if 'Species' column exists, if not skip this step
if 'Species' in new_df.columns:
    new_df['Species'] = new_df['Species'].astype('category').cat.codes

# Ensure the new data has the same feature columns
new_df = pd.get_dummies(new_df, columns=['GEOLOGY'])

# Ensure all columns present in training set are in new data
missing_cols = set(X.columns) - set(new_df.columns)
for c in missing_cols:
    new_df[c] = 0

# Ensure the order of columns in new data matches the training data
new_X = new_df[X.columns]

# Predict the response for the new dataset
new_predictions = best_clf.predict(new_X)

# Map the numerical predictions back to the original labels
new_df['Predicted_Species'] = pd.Categorical.from_codes(new_predictions, categories=list(species_mapping.values()

# Print the new dataframe with predictions
print("New data with predictions:")
print(new_df.head())

# Save the new data with predictions to a new Excel file
output_file_path = r'C:\Users\jebre\Desktop\GIS-DataMining\Newfolder\new_data_with_predictions_f.xlsx'
new_df.to_excel(output_file_path, index=False)

print(f"New data with predictions saved to {output_file_path}")
```

Figure 41 Prediction model of species type

5.4.21 Cross validation

The cross-validation results for the Random Forest model show a range of scores across different folds, indicating variability in the model's performance. The cross-validation scores are as follows: 0.66442953, 0.61744966, 0.67785235, 0.62416107, and 0.63758389. The mean cross-validation score is approximately 0.6443, suggesting moderate predictive power during training. The accuracy on the test set is 0.7203, which is higher than the mean cross-validation score, indicating that the model generalizes better to unseen data than during the cross-validation phase. This discrepancy may suggest some level of overfitting or that the model has learned patterns that are more effective for the test data. As shown in figure 42.

```
Cross-validation scores: [0.66442953 0.61744966 0.67785235 0.62416107 0.63758389]
Mean cross-validation score: 0.6442953020134228
```

Figure 42 Cross validation

5.4.22 Confusion matrix:

In the context of your Random Forest classification model, the confusion matrix provides insight into the performance of the model on the test data. Each row of the matrix corresponds to the actual class labels, while each column corresponds to the predicted class labels. For instance, the first row indicates that the model correctly predicted 42 instances of the first class but misclassified 2 instances as the second class, 3 as the third class, 1 as the fourth class, and 2 as the fifth class. Similarly, the diagonal elements of the matrix represent the number of correct predictions for each class. The off-diagonal elements indicate misclassifications, which provide valuable information on where the model is making errors, and can be used to assess and improve the model's performance. As shown in figure 43.

```
Confusion Matrix:
 [[42  2  3  1  2]
 [ 7 39  1 10  0]
 [ 8  3 29  8  2]
 [ 3  2  2 40  6]
 [ 4  1  5  3 38]]
```

Figure 43 Confusion matrix

Chapter 6: Findings and Discussions

This chapter focuses on the findings of the study and provides a detailed discussion. It begins by describing the development tools and system requirements used. The experimental setup is explained, followed by the presentation of experimental training results. Model evaluation is conducted, and any memory limit issues encountered during experiments are addressed. Finally, a discussion of the results is presented, comparing them with past research.

6.1 Result of fuzzy logic to control irrigation

The result shows the direction of opening or closing the valve of water. By the way, toning all attributes that affect irrigation can take the result as shown in Figure 44.

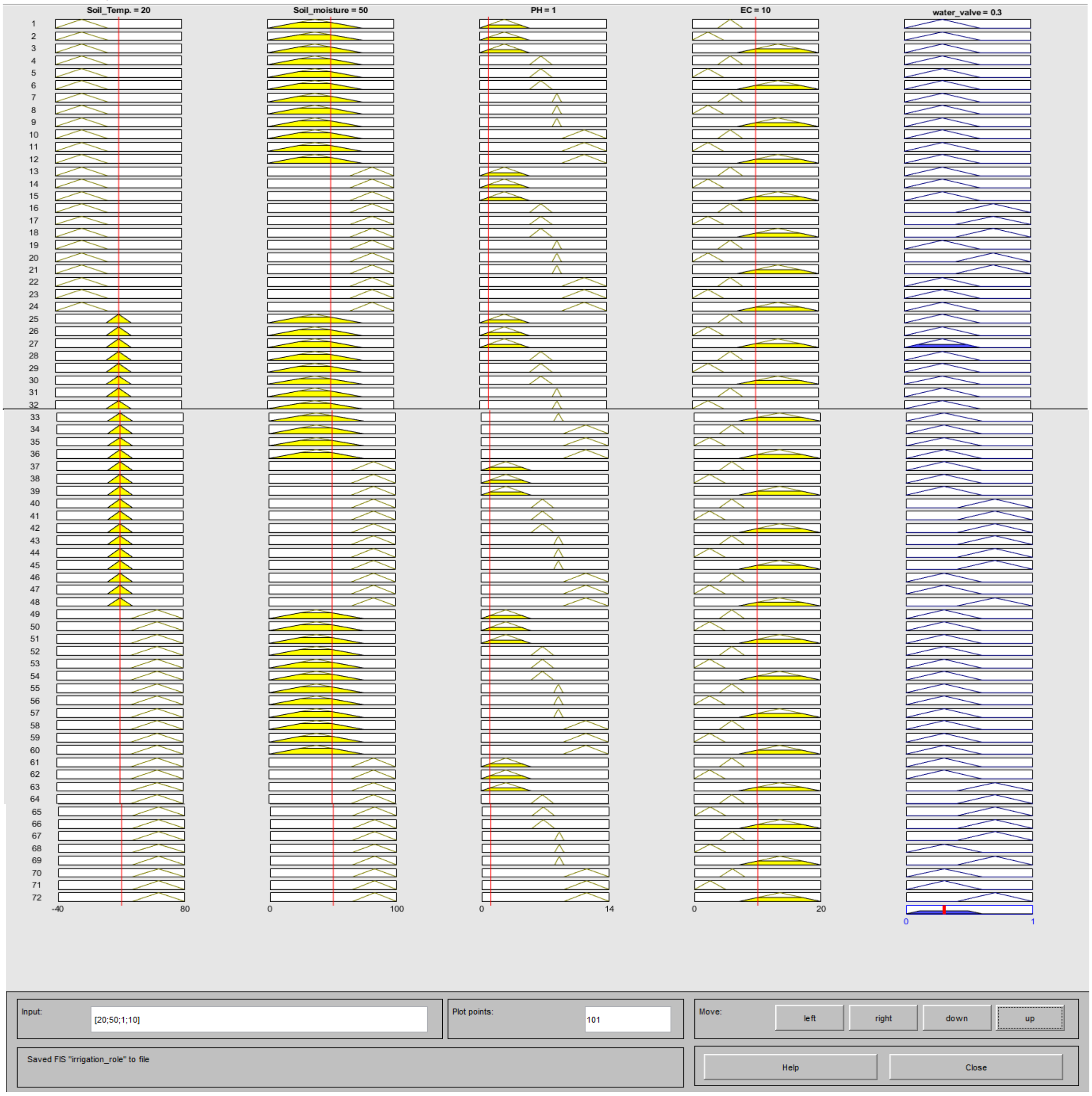

Figure 44 Result of Fuzzy Logic

6.2 Prediction model

To display the prediction model of species using ArcGIS we add the prediction data generated to ArcGIS and join the data with fishnet layer. And we categorize the data depend on species. As shown in figure 45.

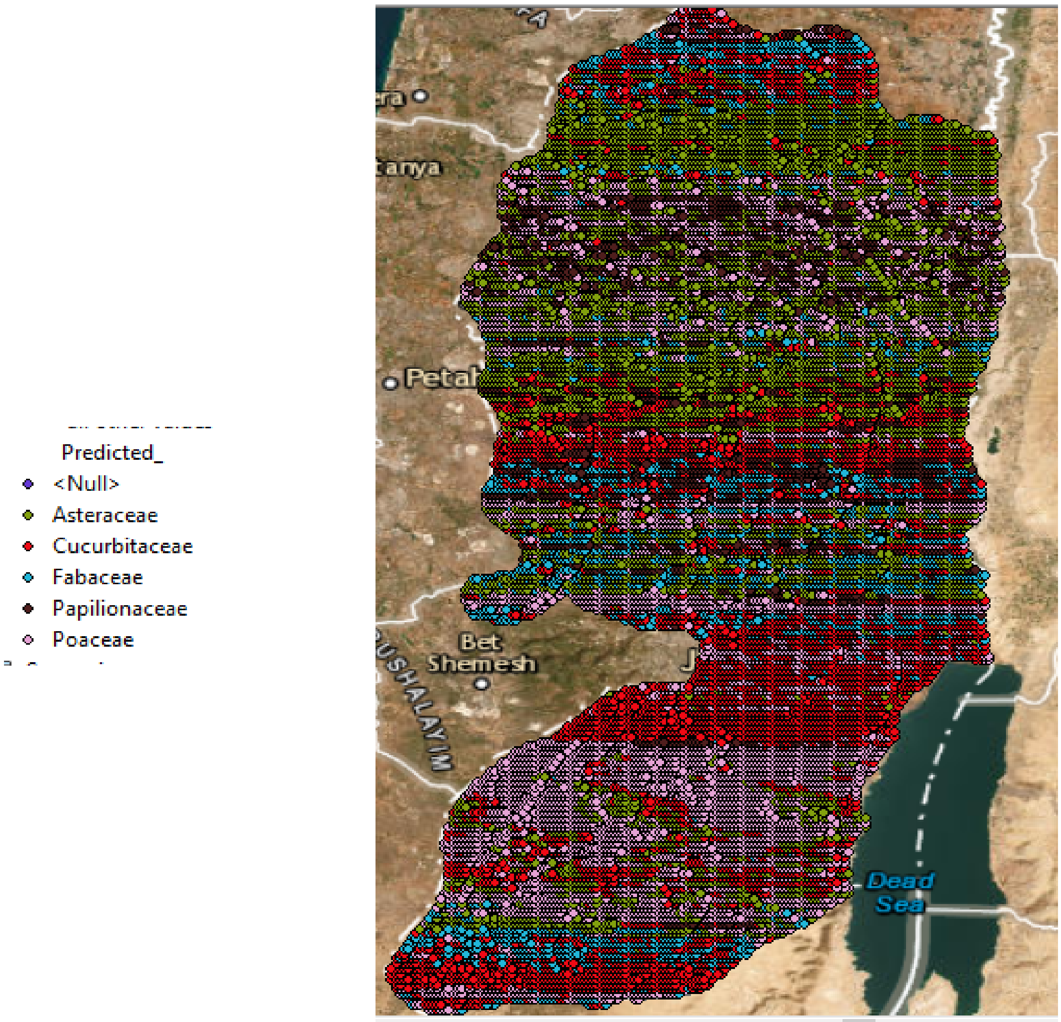

Figure 45 Prediction model map

Chapter 7: Conclusions

This work addresses critical issues in agriculture, particularly the scarcity of fresh water and the need for precise fertilizer regulation, by leveraging the capabilities of IoT, cloud computing, databases, and machine learning. The proposed system aims to utilize treated wastewater, significantly reducing the reliance on fresh water. Simultaneously, it will optimize fertilizer usage by responding to real-time data, potentially leading to reduced maintenance time, lower fertilization costs, and decreased human effort.

Furthermore, this project aims to resolve the problem of data fragmentation in Palestinian agricultural research. By utilizing cloud computing, databases, and GIS, the system will ensure that data is stored securely in multiple locations and presented in a user-friendly manner. The importance of this system extends beyond mere data storage, as the database could serve as a valuable reference for Palestinian researchers.

The Fertilization Information System (FIS) is initially simulated using fuzzy logic due to the absence of a comprehensive dataset. As the study progresses, machine learning can be employed to enhance decision-making accuracy. However, the proposed system has yet to be established, primarily due to a lack of funding. Decision-makers are encouraged to support the system's development by providing the necessary financial and technical resources.

Additionally, data from the Palestinian seed bank has been used to predict suitable regions for specific plant species using ArcGIS, Python, and Excel. The prediction model, based on Random Forest machine learning, achieved an accuracy of 0.72. There is potential for further improvement, allowing future researchers to enhance the model's accuracy.

References

Aazam, M., Zeadally, S., & Harras, K. (2018). Fog computing architecture, evaluation, and future research directions. *IEEE Communications Magazine 5, 5*, 46-52.

Aggarwal, R., & Das, M. (2012). RFID security in the context of" internet of things. *In Proceedings of the First International Conference on Security of Internet of Things*, 51-56.

Alkhamisi, S. A., & Ahmed, M. (2014). Opportunities and challenges of using treated wastewater in agriculture. *Environmental Cost and Face of Agriculture in the Gulf Cooperation Council Countries: Fostering Agriculture in the Context of Climate Change*, 109-123.

Area-Ren, C. (n.d.). Analysis Of Soil NPK, Ph And Electrical Conductivity At Adham Area-Renk, Upper Nile State. *Upper Nile State*.

Booth, B., & Mitchell, A. (2001). *Getting started with ArcGIS.*

Breiman, L. (2001). Random forests. Machine learning. (45), 5-32.

Breiman, L. (2017). Classification and regression trees. *Routledge*.

BURHAN, F., Awad, M., & RUTROT, A. (2022). PREDICTION FOR NON-REVENUE AND DEMAND OF URBAN WATER USING HYBRID MODELS OF NEURAL NETWORKS AND GENETIC ALGORITHMS.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*(20), 273-297.

Evangelos A, K., Nikolaos D, T., & Anthony C, B. (2011). Integrating RFIDs and smart objects into a UnifiedInternet of Things architecture. *Advances in Internet of Things 2011*.

Faridi, M., Verma, S., & Mukherjee, S. (2018). Integration of gis, spatial data mining, and fuzzy logic for agricultural intelligence. *In Soft Computing: Theories and Applications: Proceedings of SoCTA 2016, 1*(Springer Singapore), 171-183.

Goap, A., Sharma, D., Shukla, A., & Krishna, C. (2018). An IoT based smart irrigation management system using Machine learning and open source technologies. *Computers and electronics in agriculture*(155), 41-49.

Goyal, H., Sharma, C., & Joshi, N. (2017). An integrated approach of GIS and spatial data mining in big data. *Int. J. Comput. Appl 169, 11*, 1-6.

Houshia, O., Al-Rimawi, F., Zaid, O., Zaid, M., Imoor, J., Casey, R., & Daraghmeh, H. (2022). Influence of Treated Wastewater on the Percentage of Protein Content during Fodder Intercropping. *International Journal of Agronomy 2022, 1*, 1822328.

Jisha, R., Vignesh, G., & Deekshit, D. (2019). IOT based water level monitoring and implementation on both agriculture and domestic areas. *In 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 1*(IEEE), 1119-1123.

Karimipour, F., Delavar, M., & Kinaie, M. (2005). Water quality management using GIS data mining. *Journal of Environmental informatics, 2*(5), 61-71.

Keraita, B., Blanca, J., & Pay, D. (2008). Extent and implications of agricultural reuse of untreated, partly treated and diluted wastewater in developing countries. *CABI Reviews*, pp. 15-pp.

Kunal, S., Saha, A., & Amin, R. (2019). An overview of cloud-fog computing: Architectures, applications with security challenges. *Security and Privacy, 4*(2), e72.

Li, C., Chen, G., Zeng, G., & Ye, J. (2013). The study of soil fertility spatial variation feature based on GIS and data mining. *In Computer and Computing Technologies in Agriculture VI: 6th IFIP WG 5.14 International Conference, CCTA 2012, Zhangjiajie, China, October 19-21, 2012, Revised Selected Papers, Part II*(6), 2011-220.

Maguire, D. J. (1991). An overview and definition of GIS. *Geographical information systems: Principles and applications 1, 1*, 9-20.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR). [Internet], 9*(1), 381-386.

Mendas, A., & Delali, A. (2012). Integration of MultiCriteria Decision Analysis in GIS to develop land suitability for agriculture: Application to durum wheat cultivation in the region of Mleta in Algeria. *Computers and electronics in agriculture, 83*, 117-126.

Perumal, M., Velumani, B., Sadhasivam, A., & Ramaswamy, K. (2015). In Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI. *Spatial data mining approaches for gis–a brief review., 2*(Springer International Publishing.), 579-592.

Saleh, M. (2018). Prediction of Pipes Break in Water Distribution System Using Data Mining Tools "Case Study Nablus Municipality". *Diss. Arab American University-Jenin*.

Seber, G., & Lee, A. (2012). Linear regression analysis. *John Wiley & Sons*.

Shadeed, S., Judeh, T., & Almasri, M. (2019). Developing GIS-based water poverty and rainwater harvesting suitability maps for domestic use in the Dead Sea region (West Bank, Palestine). *Hydrology and Earth System Sciences 23, 3,* 1581-1592.

Singleton, A., & Arribas-Bel, D. (2021). Geographic data science. *Geographical Analysis, 1*(53), 61-75.

Team, R. (2000). R language definition. *Vienna, Austria: R foundation for statistical computing, 1*(3), 116.

Villarrubia, G., De Paz, J., De La Iglesia, D., & Bajo, J. (2017). Combining multi-agent systems and wireless sensor networks for monitoring crop irrigation. *Sensors, 8*(17), 1775.

Yan-Yan, S., & Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry, 27*(2), 130.

Yongjun, X., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., . . . J, Q. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation, 2*(4).

Zaragozí, B., Rabasa, A., Rodríguez-Sala, J., Navarro, J., Belda, A., & Ramón, A. (2012). Modelling farmland abandonment: A study combining GIS and data mining techniques. *Agriculture, ecosystems & environment*(155), 124-132.

Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836-3847.

الملخص

تواجه البحوث الزراعية في الشرق الأوسط، وخاصة في فلسطين، تحديات عديدة، بما في ذلك ندرة المياه، وعدم كفاية البيانات في الوقت الحقيقي لتنظيم الأسمدة، ومصادر البيانات المتفرقة. وتتأثر الزراعة الفلسطينية بشدة بسبب القيود المفروضة على الوصول إلى المياه العذبة بسبب القيود الجيوسياسية، مما يستلزم حلولاً مبتكرة لإدارة المياه واستخدام الأسمدة. تقترح هذه الدراسة تطوير نظام سحابي تنظيمي للتحكم في المياه والأسمدة، إلى جانب سحابة قاعدة البيانات الجغرافية الوطنية للبحوث الزراعية، والتي يستضيفها المركز الوطني للبحوث الزراعية(NARC) في قباطية، جنين، فلسطين. يدمج النظام أنظمة المعلومات الجغرافية (GIS)، وإنترنت الأشياء(IoT) ، وتقنيات حوسبة الضباب لتعزيز جمع البيانات وتخزينها وتحليلها. ومن خلال الاستخدام الأمثل للأسمدة (الفوسفور والنيتروجين والبوتاسيوم) وتشجيع استخدام مياه الصرف الصحي المعالجة في الري، يهدف النظام المقترح إلى ضمان ممارسات زراعية مستدامة. بالإضافة إلى اقتراح أسلوب لتنظيم البيانات الزراعية يمكن تعميمه على كافة أنواع البيانات.

تم استخدام البيانات الأولية من بنك الجينات النباتية الفلسطيني من المركز الوطني للأبحاث لإنشاء خريطة تنبؤية للأنواع النباتية في جميع مناطق الضفة الغربية أو الأماكن التي يمكن زراعتها فيها. استخدمنا برنامج ArcGIS وخوارزمية التعلم الآلي Random Forest في بيئة Python للتنبؤ وحصلنا على نتيجة تنبؤ تبلغ 0.72.