Synergetic Innovation in Gherkin Cultivation Based on Machine Learning Algorithms

Dayarathna T.W.K.D.A.

B.Sc. (Hons) Degree in Information Technology

(Specialization in Software Engineering)

Department of Software Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

Aug 2024

Actual Harvest Prediction and Factor Analysis

Dayarathna T.W.K.D.A.

IT20254216

The dissertation was submitted in partial fulfilment of the requirements
for the B.Sc. Special Honors degree in Information Technology
(Specialization in Software Engineering)

Department of Software Engineering

Sri Lanka Institute of Information Technology
Sri Lanka

Aug 2024

1 DECLARATION

I declare that this is my own work, and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Name	Student ID	Signature
Dayarathna T.W.K.D.A.	IT20254216	

The above candidate is carrying out research for the undergraduate Dissertation under my supervision.

Signature of the supervisor   Date

8/23/2024

………………………….   ……………..…………

(Mr. Dharshana Kasthurirathna)

2 ABSTRACT

This thesis aims to develop an advance machine learning model to predict the actual harvest of yield of gherkins and analyze the factors that directly affect the harvest. The key factors that focused for the prediction process are soil pH levels, nutrient composition (N, P, K), temperature, rainfall, and fertilizer combinations. The study utilized the Random Forest algorithm due to the best accuracy compared to the other regression models including Linear Regression, Decision Tree and SVM., The Random Forest model achieved 0.94 training accuracy after the optimization. Additionally, the thesis introduced a mobile application to the farmers and the officers who are dealing with the gherkin cultivation. The application developed using Flask backend and a React-Native fronted and allows users to input data and obtain yield predictions in a user-friendly manner. The model was tested against the actual harvest values of Yala-Off season and the mobile application was tested with positive feedback from farmers and company users. The study concluded that data-driven approaches in agriculture can significantly improve productivity and resource optimization.

Keywords: Machine Learning, Harvest Prediction, Gherkin Cultivation, Random Forest, Factor Analysis, Fertilizer Optimization, Mobile Application, Sustainable Agriculture, Data-Driven Decision Making

3 ACKNOWLEDGEMENT

This research study being a mixture of technology and agriculture required the guidance and assistance of not only technology experts but also agriculture professionals. The sincere gratitude to the external supervisors; Mr. Kumara Perera, Mr. Nalaka Jayawardana, Mr. Gayan Fernando, and Mr. Ananda Pathirage from HJS Condiments Limited for the invaluable assistance and guidance.

4 TABLE OF CONTENTS

5 LIST OF FIGURES

6 LIST OF TABLES

7 LIST OF ABBREVIATIONS

Abbreviation	Description
SVM	Super Vector Machine
PCA	Principal Component Analysis
MOP	Muriate of Potash
TSP	Triple superphosphate

1 INTRODUCTION

1.1 Background

Gherkins belong to the cucumber family, *Cucumis sativus*. They are grown and eaten all over the world, since they are a popular preserved fast-food. Gherkins have a history going as far back as more than several thousand years. The forerunner of the gherkin can be determined from the wild specimens, in the foothills of the Greater Himalayas in what is now modern India. This creeper type plant with small yellow flowers belongs to the genus Cucumis and is known was believed to be domesticated around 4000 years ago. [1] Starting from India cucumbers found their way to the Middle East and later Mediterranean regions became incorporated in the diet of the ancient world power of Greeks and the Romans.

Thus, pickling cucumbers may have originated as a way of keeping them in good condition and thus available for consumption throughout the year or whenever there was convection with fresh produce. The very name “gherkin” is said to originate from the Dutch word “gurken” which means a small, pickled cucumber The gherkin however had become a rage in Europe, in the Medieval ages and it was from here that the growing of gherkins spread in the continent. The gherkins were later brought into the Americas during the 16th century with the European explorers who adopted taking them. [2]

As more people are inclined toward taking pickled foods and are demanding low calorie foods, the global gherkin market is expected to soar further. With consumer awareness of the nutritional qualities of gherkins on the increase there is a further prospect of increased global production.

Commonly known as small fruits, gherkins contain essential nutrients oblivious of their size. It is low in calories therefore recommending it for people who would wish to undergo some dietary intention. It also contains minimum calories, a 100-gram serving of gherkin has only 12 calories, so it is perfect for weight watchers.

They can be regarded as the source of vitamins and minerals in the body systems. They are very rich in vitamin K which is essential in blood clotting and in the formation of bones. A 100-gram portion of Gherkin is stated to contain 20 % of Vitamin K daily requirement, Vitamin A at 10 %, Vitamin C at 5 %.

There are also minerals; potassium in gherkins is essential in controlling blood pressure and magnesium that has roles for muscles and nerves. Gherkin has a high fiber content that helps in digestion, overcoming feelings of hunger, and other health benefits.

Even the process of pickling contributes to the advantage of taking gherkins. Pickles contain vitamin C; potassium and they also have some number of probiotics which help in the proper digestion. Such probiotics can enhance gut health, protect against infections, and possibly play a role in mental wellbeing by modulating the gut-brain connection.

All in all, it may be concluded that gherkin is a rather useful product from the point of view of its nutritional value, as a part of a daily diet.

India itself is now the largest producer and exporter of gherkins and contributes to most of the world demand. It is cultivated mainly in the states of Karnataka, Tamil Nadu and Andhra Pradesh in India and most of the products are exported to Europe and North America.

France, Germany and Poland are the famous producers of gherkins, and they also have been practicing the pickling of gherkin for several centuries. France has its cornichons, small sharp gherkins which may be seasoned with tarragon, for example. The normal form of gherkins consumed in Poland is the brine-pickled type, and the country has a large production of this relish for local use and export.

The largest market for gherkins is the United States of America because pickled cucumbers are an integral part of the AmeriThe U. S. too has a well-established pickling industry most of the companies that process pickled gherkin are engaged in processing a full range of pickled gherkin products including dill pickles and sweet bread-and-butter pickles.

Over the last few years, there was a trend in consumption of quality, natural products such as artisan and organic gherkins. This trend is especially prominent in North America and Europe where the buying public become discriminating against the type of pickles they buy; articulating a preference for locally made, artisanal pickles with niche flavors.

Today Sri Lanka has succeeded in becoming one of the players in the international gherkin market, owing to which it is possible to talk about Sri Lanka as one of the exporters of quality gherkins, courage by the availability of optimum climate required for the cultivation of gherkin and good soil texture exported of gherkin has been on the increase in the recent decades in addition to the increasing global demand for pickles. Sri Lankan Gherkin pickles have a great demand in the international industry because of the excellent natural taste and aroma of the products. The naturally placed geographic location, year-round sunlight and most favorable atmospheric conditions are causing that. Therefore, Sri Lanka placed a significant exporting brand in the foreign market. Major proportion of gherkin is produced in the central and the Northern provinces of Sri Lanka and mostly produced by the small and middle farmers.

Majority of the gherkin export in Sri Lanka is done through contract farming in which export company deals directly with the growers. It has been a good understanding for both the farmer and the exporter since it has provided the farmer with a market for his produced gherkins and a source of supply for the exporter. The Sri Lankan gherkins are big, crisp and have excellent taste, due to which they are much in demand in export markets.

In the agribusiness market, the HJS Condiments Limited in Sri Lanka is the largest exporter of gherkin. The firm has expanded to be an exporter because its products can be sold in Europe, North America and the Middle East. HJS has a direct interface with more than 10000 farmers all over the country and provides all essential input, technical support besides reasonable and remunerative price for establishing and carrying on a profitable business. Because of this, through the network of suppliers, HJS has established means the company has been able to supply gherkins in a continuous way and nonetheless meet the quality demands of the world market.

The export of gherkins for Sri Lanka has been very much rewarding for the economy since it has brought flow of foreign exchange and touch of hope for many poor families that are in the farming zones of the country. In the recent past the industry has also produced an additional product that is gherkins in brine or vinegar which is locally processed and canned before exporting to other markets. Not only does it help in increasing the revenue generation of the gherkin industry but on the side, it also helps in employment generation of the related sectors such as processing and packaging.

However, some of the challenges facing the Sri Lankan gherkin export industry as we shall shortly demonstrate include fluctuating market prices in the global market, other exporters of gherkin and effects of climatic changes on agricultural production. However, if, Sri Lanka is decisive to develop more its current form of technologic farming, increase resource for more research & development session, implement more and more of sustainable facility in agricultural sector like mechanizing and vertical farming, it can maintain and increase portion among the global gherkin market.

Therefore, the exporting business of gherkin has greatly grown and become a stable business in Sri Lankan mainly with better economic returns and the integrity of the farmers in the rural areas. Therefore, while the global market demand for pickled products is increasing the quality of the products of Sri Lanka and environmentally sustainable technologies will be the key to success in the future.

1.2 Literature Survey

N. Suresh et al. (2021) [3] and N. Rale et al. (2019) [4] both work on the applicability of machine learning models, more precisely the Random Forest algorithm, toward crop yield estimation and optimized cultivation practices. Suresh et al. have shown the efficiency of Random Forest in crop yield prediction through the analysis of various climatic and agronomic parameters. This will clearly be very close to the aims of your research on synergetic innovation in gherkin cultivation, where a key component in prediction is to achieve optimization in production outcomes; Rale et al [4]. discuss more generally how machine learning can be applied not only to predict crop yields but also to optimize cultivation practices by integrating climate forecasts. The approach to integrating varied sources of data appeals to the focus on using predictive models to guide gherkin farmers to achieve expected levels of harvest. The studies underline the critical role of machine learning in modern agriculture, thus forming a base for developing more sophisticated prediction systems. Such insights will be very valuable in refining your harvest prediction model; more so, getting to understand how various factors that affect the gherkin cultivation interact with each other to come up with reliable and accurate predictions.

Some of the conflicting opinions on the employment of machine learning and factor analysis for agricultural forecasting can be understood from the works of Kandan et al. (2021) [5] and H. Zeng et al. (2019) [6]. Kandan et al. describe a crop yield forecasting model which includes climatic and agricultural factors, and their paper uses applied ML to enhance the forecast models. This approach, in fact, owes its applicability to the own research agenda; similar methods could be used to predict gherkin output by integrating environmental and agronomic information. The fact that the study also focuses on the integration of multiple data sources to improve the prediction performance is something that you seek to develop for the prediction of harvest. On the other hand, Zeng et al [6]concentrate on the ultimate factor analysis methods of parallel factor analysis methods and signal processing. Even though their work is not in the field of agriculture, the methods they discuss can be used in your factor analysis of agricultural inputs including the specific fertilizers. In combination, these works give insights into the uses of machine learning and factor analysis when given to agricultural forecasting to promote the research.

Han et al. (2016) [7] and S. Wang [8], in their surveys for applications, account for the development of knowledge-based systems and for factor analysis of agricultural situations. Han and his colleagues suggest developing Experiential Knowledge Platform to support the domain experts in making decisions. Discuss, in the light of such an integrated analysis of experiential knowledge, how such a platform can be used for supporting reliable decisions—having relevance for your goal to construct a knowledge base for gherkin cultivation. A platform like this would add value to the gherkin farmers through important insight and recommendations arising out of the aggregation of expert knowledge to improve gherkin-cultivation practice. In his part, Wang discusses the matter of fresh food safety by applying factor analysis to evaluate and enhance the quality of food by collecting data from the storage environment to pinpoint key factors that may have some impacts on the quality of food. As the research food safety, the technique of factor analysis according to Wang is highly relevant to the study. It helps to analyze the various factors that influence gherkin cultivation, ranging from general environmental conditions and agricultural inputs to the most important elements that will be giving optimum harvests. Such studies are the counseling of systems knowledge and factor analysis in agricultural decision-making, providing effective tools and methodologies for the research.

T. Wang, 2020 [9], and Z. Ping and T. Yuanhua, in 2015 [10], consider the application of factor analysis and large-scale knowledge base system construction, which will be helpful for the research on synergetic innovation in gherkin cultivation. Wang's work gauges the quality of fixed assets using factor analysis, proving the feasibility of this method in assessing the reliability and performance of such tangible assets in a manufacturing sector. Although this research is not directly related to agriculture, the factor analysis model discussed can be tailored in assessing the impact of inputs into agricultural production on the quality of the yield of gherkins. This will help in doing an in-depth analysis of factors that most affect gherkin cultivation to guide farmers on better practices. Ping and Yuanhua, discuss methods for constructing and digitizing large knowledge base systems. Knowledge representation and data modeling has been presented in their work, where they shed light on how knowledge bases must be structured so that data retrieval and decision making can be executed with ease. Their insights are, therefore, of direct application toward your goal of developing a knowledge base system for gherkin farmers. Integrate their methodologies to build a holistic, user-friendly knowledge base that would assist farmers in optimizing cultivation practices.

Integrate the approaches described in the following studies offer a strong tool to back decisions in gherkin cultivation with the intention of arriving at more effective and sustainable agricultural practices.

1.3 Research Gap

When referring to the agricultural field, numerous studies have explored various methodologies and solutions for crops to yield predictions of various crops to improve decision making. But the amount of research done on gherkin cultivation is negligible. Therefore, there was a significant gap in the conducted research compared to other theses. However, the gaps in existing studies are identified for the proposed system about gherkin cultivation in Sri Lanka, using machine learning models and factor analysis together with a knowledge-based system, for the improvement of actual harvest predictions and optimization of cultivation practices.

The study by N. Suresh et al. (2021) (Study A) tries to prove the prediction capability of the Random Forest algorithm for crop yield in view of different climatic factors. In a way, although it handles in a generalized context, the specific prediction about crops can be done by considering local agricultural practices and environmental conditions. In this relationship, the proposed system tries to fill in this gap by applying the Random Forest technique in a more specialized manner in gherkin cultivation, taking various local factors like soil type, regional climate, and farming practice to improve its prediction accuracy of harvests.

The research of N. Rale et al. (Study B) fits the meteorological data to predict crop cultivation, which is very important for understanding the impact of changes in weather patterns on agriculture. On the other hand, the practical applications of these predictions in specific crop cultivation and management do not come up, including the real-time adjustment possibilities in cultivation practice. The system now provides for the prediction of actual gherkin harvesting, and second, an integration of a knowledge-based system that provides actionable insights and actionable recommendations on how to optimize gherkin cultivation considering real-time data.

M. Kandan et al. (Study C) focus on implementing a yield prediction system using climatic and agricultural parameters. While it noted that several factors need to be considered in various yield predictions, it did not consider their interactions or how they could be jointly optimized to raise specific harvest objectives. This is in the aspect of this factor analysis being included within the system to incorporate an in-depth understanding of the interaction and impact of these various variables with the actual harvesting of gherkins and the recommendations on optimizing them.

In the work of H. Zeng et al. (Study D), a detailed comparison is drawn between complex parallel factor analysis and parallel factor analysis with respect to applications in signal processing and data decomposition. This research is very helpful for understanding theoretical aspects of factor analysis but does not apply its methods to agricultural data in crop yield prediction. It does so through the application of factor analysis techniques in identifying, and subsequently optimizing, the key factors governing gherkin yield to provide actionable solutions to enhance cultivation outcomes.

The table below compares the five previously mentioned existing solutions to our proposed system in detail.

	Study A	Study B	Study C	Study D	Study E	Proposed solution
Specific for gherkin cultivation	No	No	No	No	No	Yes
Comprehensive harvest prediction models	No	Yes	No	No	No	Yes
Regional adaptation on soil and climate	Yes	Yes	No	No	No	Yes
Integration of factor analysis	No	No	Yes	Yes	No	Yes
Real-time data utilizations	No	No	No	No	No	Yes

Table 4.1.1-1 Research Gap based on previous research vs conducted research.

1.4 Research Problem

Gherkin is one of the major agricultural export crops in Sri Lanka, but there is no proper and reliable method for actual yield forecasting in this industry, which is very important to cater to international market needs. The current farming practices tend to give little or no consideration to the assimilation of critical data in this case, including the soil pH levels, the percentage of N-P-K in the soil, rainfall, and temperature, and the combination of fertilizers applied. Because analysis was not systematically approached, farmers end up with less-than-optimal yields.

The existing prediction models and decision support systems are not sufficient to cater to the needs of growing gherkins in Sri Lanka, since usually, an overall factor analysis is not incorporated to predict the optimum combinations of fertilizer applications. Moreover, a knowledge base strong enough to support farmers in making appropriate decisions based on their environmental setting is not available. Thus, the critical task will be to develop a machine learning-based actual harvest prediction model using data related to soil properties, climatic conditions, and the application of fertilizer. This will be aided by a factor analysis approach to find the most appropriate combinations of fertilizers and supplemented by a knowledge-based system to achieve high accuracy in prediction and to optimize resource use for productivity and sustainability in gherkin farming in Sri Lanka.

In brief, the below points can be gain as the main research problems.

How can harvest prediction be made more accurate?
What is the best method to identify the optimal fertilizer combination?
How can a knowledge base help farmers make better decisions?
What challenges exist in using data-driven farming methods?
How can improved predictions benefit Sri Lanka's gherkin industry?

2 RESEARCH OBJECTIVES

2.1 Main Objectives

The overall aim of the research project is to develop an integrated system for actual harvest prediction in gherkin cultivation, including factor analysis and a knowledge base in support of farmers' decisions to optimize their practice of agriculture. The research focuses on making a very accurate prediction and bringing out actionable insights that will help decision-making in the field.

2.2 Specific Objectives

1. Develop a farmer friendly application
2. Accurate predictions of Actual Harvest
3. Done a factor analysis to suggest the suitable fertilizer combination for the specific cultivation

3 SYSTEM OVERVIEW

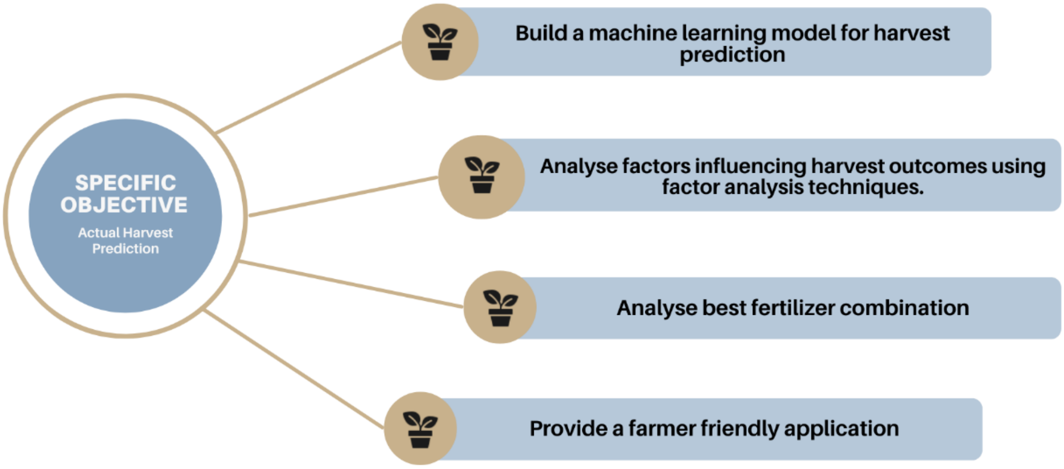

SPECIFIC OBJECTIVE
Actual Harvest Prediction
Build a machine learning model for harvest prediction
Analyse factors influencing harvest outcomes using factor analysis techniques.
Analyse best fertilizer combination
Provide a farmer friendly application

Figure 3-2.2-1 Objectives summary diagram

3.1 Overall System Overview

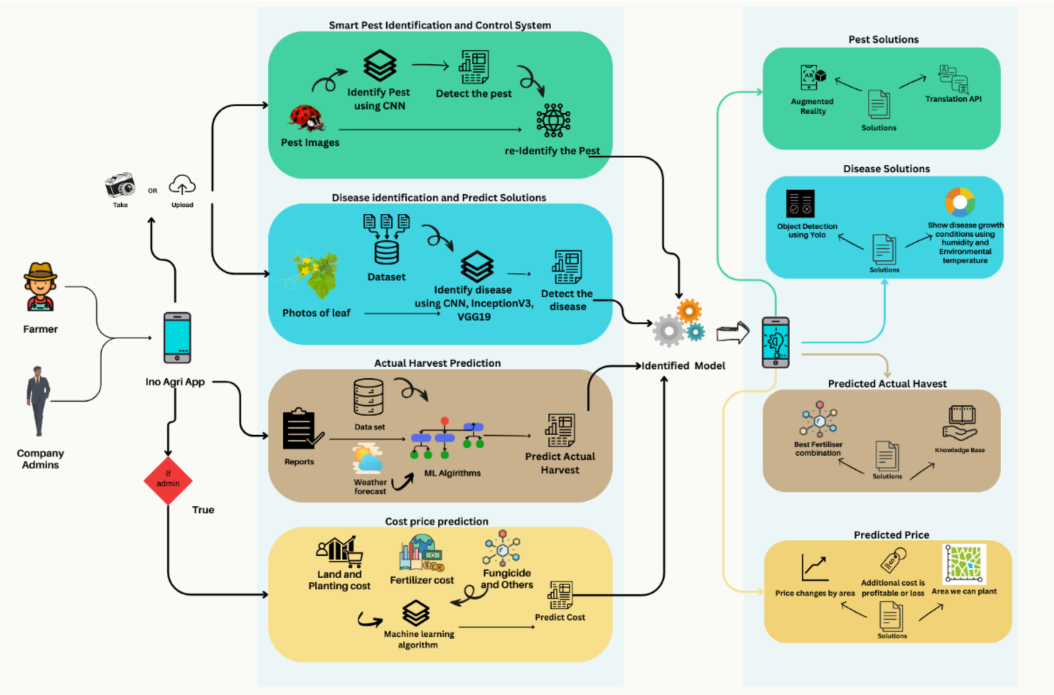


Figure 3.1-1 Overall System Diagram

According to the figure 3.1-1, the farmers and the company admins are the users who directly engage with the mobile application. The application is consistent with four main functionalities. Those are smart pest identification, disease identification, actual harvest prediction and cost price prediction. All these functionalities are developed using machine learning models and they give the accurate predictions using the past datasets.

The smart pest identification and the disease identification models were used image processing algorithms for object detection and identification. The images of pests and the diseases were used to train the models. The harvest prediction and the cost price prediction models were trained using supervised learning algorithms. Labeled datasets were used to train the models and optimization techniques were used to improve the training accuracy of the models.

3.2 Component Specific Overview

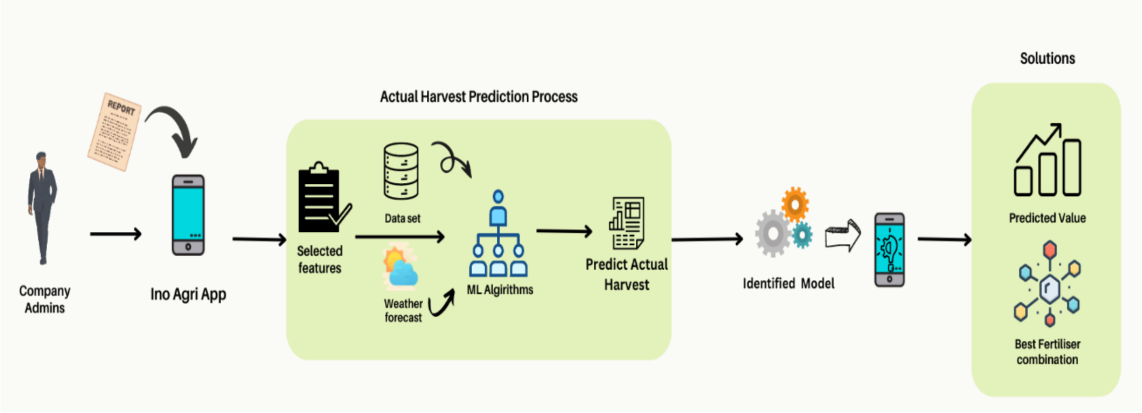


Figure 3.2-1 Component Specific System Diagram

According to figure 3.2-1, the company admins can use the harvest prediction functionality of the mobile application. First, the admins need to enter all the details of the cultivation including cultivation details, soil details and fertilizer details. Then the model will predict and display the actual harvest value for the yield. Then the expected harvest should be entered and the best fertilizer that should be applied to get the expected harvest will be displayed to the user.

4 METHODOLOGY

4.1 Methodology

4.1.1 Main Task

The goal of this study was to acquire and augment a set of features for utilizing a machine learning approach to obtain the actual harvest of gherkin production. Hypotheses to be tested in the present study concerned the effects of the extent of land area under cultivation, the pH of the soil, and the kinds of chemicals, minerals, and fertilizers essential for plant growth including Ca, Mg, K, N, P, and Zn, as well as Urea,TSP, MOP and CaNO3. Climatic variables were also incorporated including rainfall and temperature characteristics. The actual and expected harvest values are also included among the datasets recorded during the study of the variants.

Hayleys' subject matter experts helped complete the process of selecting data by determining key factors affecting the harvest. Four machine learning algorithms were considered for model training: Linear Regression, Decision Tree, Random Forest and Support Vector Machine (SVM). As highlighted earlier in this paper, the best algorithm that was found was the Random Forest, which had an initial accuracy of 0.89.

To improve the outcome of the model even more, hyperparameters were also optimized based on certain parameters including numbers of estimators = 10, 50 and 100, as well as a criterion in which options included squared_error, absolute_error and poisson. After optimization, there is a great improvement of the accuracy with the model getting to 0.93.

Besides the forecast of the harvest, factor analysis was used to find out the most suitable fertilizer combination to get the forecasted harvest. The emphasis was made to study how Urea, TSP, MOP and CaNO3 affect the actual harvest prospects. Analyzing the difference between realized and potential harvest, the model suggested which type of fertilizer is beneficial to be used to minimize this difference.

The development of the predictive model was to generate a practical application which may be useful to the farmers. This app is intended to assist farmers to be in a position of making correct decision making concerning the type of fertilizer or fertilizers to apply depending on the physical condition of the soil and general environmental factors. Furthermore, the knowledge base was developed in a form of farmer's version specifically to enhance its understandability.

In this study, the volume of machine learning approach in enhancing the yield of gherkin cultivation has been established alongside offering a unique solution. Credibility of using a knowledge base and practical recommendations can be seen in how farmers can most effectively apply fertilizer and increase yields. In addition, it has also presented the possibilities of commercialization with the aim of delivering data analytics and advice to farmers in Sri Lanka and other countries to enhance better agricultural and economic future.

4.1.2 Sub Tasks

The sub-tasks for this research project were classified into the following fundamental phases to achieve the goal of harvest prediction model and its applicability.

The first sub-task was featuring selection where the initial aim was confined to determining the strongly correlated features that characterize actual harvest. These were pH, land size, and major soil nutrients like, Calcium, Magnesium, Potassium, Nitrogen, Phosphorus and Zinc. Other factors about the type of fertilizer that should be used including Urea, TSP (Triple Super Phosphate), MOP (Muriate of Potash) and calcium nitrate (CaNO3). The external factors including rainfall and temperature were also considered. The choice of these features was made in consultation with specialists to make them relevant and valid.

The second sub-task that was identified was model development- This entailed developing several machine learning models. The following models were applied: Linear Regression, Decision Tree, Random Forest, SVM where the aim was to estimate the actual harvest. Random Forest was the most accurate and hence it was chosen from the rest.

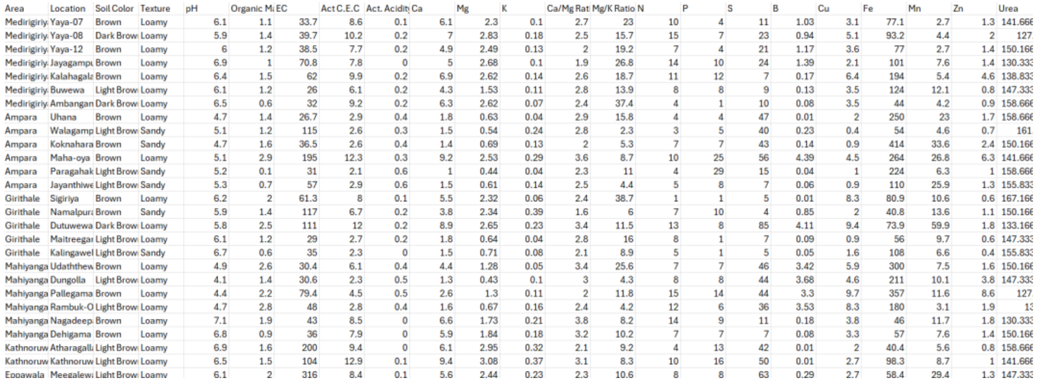
Hyper Tuning was the subsequent step which was done with an intention of turning the accuracy further high with the selected random forest model. The number of estimators (n_estimators) and other parameters relating to criterion for splitting nodes were modified. Consequently, while the accuracy dropped from 0.94 when predicted versus actual sunspot numbers, the model's actual harvest performance was enhanced from 0.94 to 0.98.

The subsequent step was factor analysis which proved to be important in determining the right types of fertilizers to apply to foresee the real chance to close the discrepancy between expected and actual yields. This analysis made it possible for the system to advise farmers concerning fertilizer amendment to increase their yields.

Finally, the application development task, which involves designing a user-friendly touch farmer mobile application fully implemented to facilitate data entry and use of the developed machine learning model to generate useful useable recommendations.

4.1.3 Data Collection

The data was collected mainly by engaging with HJS Condiments Limited. The datasets were created from multiple sources including the soil testing lab reports, weather forecasting reports, harvest detail reports and the fertilizer reports. Agronomists from HJS helped in explaining which factors were important for the yield of the crop. Some of the required fields were soil pH, area of land, calcium, magnesium, potassium, nitrogen, phosphorus, zinc, urea, TSP, MOP, calcium nitrate, rainfall, temperature, actual and expected yield. The data set was preprocessed using machine learning techniques including one hot encoding and removed unnecessary columns which were not directly affected to the harvest.



Area	Location	Soil Color	Texture	pH	Organic Ma	EC	Act C.E.C	Act. Acidity	Ca	Mg	K	Ca/Mg Rat	Mg/K Ratio	N	P	S	B	Cu	Fe	Mn	Zn	Urea
Medirigiriy	Yaya-07	Brown	Loamy	6.1	1.1	33.7	8.6	0.1	6.1	2.3	0.1	2.7	23	10	4	11	1.03	3.1	77.1	2.7	1.3	141.666
Medirigiriy	Yaya-08	Dark Brow	Loamy	5.9	1.4	39.7	10.2	0.2	7	2.83	0.18	2.5	15.7	15	7	23	0.94	5.1	93.2	4.4	2	127.
Medirigiriy	Yaya-12	Brown	Loamy	6	1.2	38.5	7.7	0.2	4.9	2.49	0.13	2	19.2	7	4	21	1.17	3.6	77	2.7	1.4	150.166
Medirigiriy	Jayagampu	Brown	Loamy	6.9	1	70.8	7.8	0	5	2.68	0.1	1.9	26.8	14	10	24	1.39	2.1	101	7.6	1.4	130.333
Medirigiriy	Kalahagala	Brown	Loamy	6.4	1.5	62	9.9	0.2	6.9	2.62	0.14	2.6	18.7	11	12	7	0.17	6.4	194	5.4	4.6	138.833
Medirigiriy	Buwewa	Light Brow	Loamy	6.1	1.2	26	6.1	0.2	4.3	1.53	0.11	2.8	13.9	8	8	9	0.13	3.5	124	12.1	0.8	147.333
Medirigiriy	Ambangan	Dark Brow	Loamy	6.5	0.6	32	9.2	0.2	6.3	2.62	0.07	2.4	37.4	4	1	10	0.08	3.5	44	4.2	0.9	158.666
Ampara	Uhana	Brown	Loamy	4.7	1.4	26.7	2.9	0.4	1.8	0.63	0.04	2.9	15.8	4	4	47	0.01	2	250	23	1.7	158.666
Ampara	Walagamp	Light Brow	Sandy	5.1	1.2	115	2.6	0.3	1.5	0.54	0.24	2.8	2.3	3	5	40	0.23	0.4	54	4.6	0.7	161.
Ampara	Koknahara	Brown	Sandy	4.7	1.6	36.5	2.6	0.4	1.4	0.69	0.13	2	5.3	7	7	43	0.14	0.9	414	33.6	2.4	150.166
Ampara	Maha-oya	Brown	Loamy	5.1	2.9	195	12.3	0.3	9.2	2.53	0.29	3.6	8.7	10	25	56	4.39	4.5	264	26.8	6.3	141.666
Ampara	Paragahak	Light Brow	Sandy	5.2	0.1	31	2.1	0.6	1	0.44	0.04	2.3	11	4	29	15	0.04	1	224	6.3	1	158.666
Ampara	Jayanthiwe	Light Brow	Sandy	5.3	0.7	57	2.9	0.6	1.5	0.61	0.14	2.5	4.4	5	8	7	0.06	0.9	110	25.9	1.3	155.833
Girithale	Sigiriya	Brown	Loamy	6.2	2	61.3	8	0.1	5.5	2.32	0.06	2.4	38.7	1	1	5	0.01	8.3	80.9	10.6	0.6	167.166
Girithale	Namalpura	Brown	Sandy	5.9	1.4	117	6.7	0.2	3.8	2.34	0.39	1.6	6	7	10	4	0.85	2	40.8	13.6	1.1	150.166
Girithale	Dutuwewa	Dark Brow	Loamy	5.8	2.5	111	12	0.2	8.9	2.65	0.23	3.4	11.5	13	8	85	4.11	9.4	73.9	59.9	1.8	133.166
Girithale	Maitreegar	Light Brow	Loamy	6.1	1.2	29	2.7	0.2	1.8	0.64	0.04	2.8	16	8	1	7	0.09	0.9	56	9.7	0.6	147.333
Girithale	Kalingawel	Light Brow	Sandy	6.7	0.6	35	2.3	0	1.5	0.71	0.08	2.1	8.9	5	1	5	0.05	1.6	108	6.6	0.4	155.833
Mahiyanga	Udaththew	Brown	Loamy	4.9	2.6	30.4	6.1	0.4	4.4	1.28	0.05	3.4	25.6	7	7	46	3.42	5.9	300	7.5	1.6	150.166
Mahiyanga	Dungolla	Light Brow	Loamy	4.1	1.4	30.6	2.3	0.5	1.3	0.43	0.1	3	4.3	8	8	44	3.68	4.6	211	10.1	3.8	147.333
Mahiyanga	Pallegama	Brown	Loamy	4.4	2.2	79.4	4.5	0.5	2.6	1.3	0.11	2	11.8	15	14	44	3.3	9.7	357	11.6	8.6	127.
Mahiyanga	Rambuk-O	Light Brow	Loamy	4.7	2.8	48	2.8	0.4	1.6	0.67	0.16	2.4	4.2	12	6	36	3.53	8.3	180	3.1	1.9	13
Mahiyanga	Nagadeepa	Brown	Loamy	7.1	1.9	43	8.5	0	6.6	1.73	0.21	3.8	8.2	14	9	11	0.18	3.8	46	11.7	1.8	130.333
Mahiyanga	Dehigama	Brown	Loamy	6.8	0.9	36	7.9	0	5.9	1.84	0.18	3.2	10.2	7	7	7	0.08	3.3	57	7.6	1.4	150.166
Kathnoruw	Atharagalla	Light Brow	Loamy	6.9	1.6	200	9.4	0	6.1	2.95	0.32	2.1	9.2	4	13	42	0.01	2	40.4	5.6	0.8	158.666
Kathnoruw	Kathnoruw	Light Brow	Loamy	6.5	1.5	104	12.9	0.1	9.4	3.08	0.37	3.1	8.3	10	16	50	0.01	2.7	98.3	8.7	1	141.666
Eppawala	Meegalewa	Light Brow	Loamy	6.1	2	316	8.4	0.1	5.6	2.44	0.23	2.3	10.6	8	8	63	0.29	2.7	58.4	29.4	1.3	147.333

Figure 4.1.3-0-1 Figure of dataset

4.1.4 Tools and Technologies

Category	Tools and Technologies
Backend and Frontend	Python, Flask, React Native
Machine Learning Libraries	Sklearn, NumPy, Pandas, Seaborn
Tools for Hyperparameter Tuning	GridSearchCV
Factor Analysis Tools	Principal Component Analysis (PCA), Correlation Matrix
Development Environment	Jupyter Notebook, VS Code
UI/UX Design Tool	Figma
Database	MongoDB
Version Controlling	Github
Collaboration Tool	Microsoft Teams
Project Management	Microsoft Planner

Table 4.1.4-1 Tools and Technologies

4.2 Commercialization Aspects of the Application

The focuses of the commercialization of the mobile application are the two-sided use of the application: for company administrators and farmers. For the benefit of company admins, the system makes it possible to input the information required at foretelling harvest yields. These are the key end-users of the prediction function as only these admins possess important details of soil quality, weather prospects, and the right amount of fertilizer to apply. By these inputs, the actual harvest is forecasted and the best combination of fertilizers to bring the gap is recommended. This predictive analysis is important most especially to the company admins who will have to ensure farmers increase their production levels and achieve expected yields.

With regards to the design of the application, for ease of use and to ensure that the company admin can input data and access the results efficiently, the application is developed with an appropriately simplified user interface. The non-technical features also make the solutions easily implementable to blend with day to day working with little to no training to the users. As usually admins collect huge data related to the soil health, usage of fertilizers and the climate conditions, the application helps them in simplifying the decision-making pertaining to the agriculture management.

From a commercialization point of view, the proposed app presents a platform for affiliation with agricultural firms and bodies with interest in the farmers. By realizing this system, companies will be able to improve their support services and give farmers information and profound solutions on how they can improve the yield in their farms. Additionally, it is versatile; it can be implemented as a standalone app and as an integrated solution into existing agriculture platforms hence suitable for large agricultural companies and smallholder farmers. Focus on ease of use and accurate recommendations gives the application not only a friendly interface but also utility in furthering the efficiency of agriculture.

4.3 Implementation and Testing

4.3.1 Implementation

In the implementation phase of the research that was undertaken and some of the steps included. First, cleaning was done on the generated dataset to make it suitable to be used in developing machine learning models. This process entailed the elimination of repetitive values as well as inputs that called for truncation of longer series to fit a consistent length of data fields Expectation, and the conversion of Object-Type data to numerical data through Use of One-Hot Encoder. For example, where format specific columns like “grade” existed, different sizes; small, medium, large were given numerical representations of 0, 1, 2 respectively. Also, columns unnecessary for harvest were deleted and the Hayleys’ subject experts’ most important features for harvest were only kept.

Once pre-processing steps are gone through, four machine learning algorithms, namely Linear Regression, Decision Tree, SVMs and Random Forests were used to predict the actual harvest. Random Forest model given the best accuracy percentage as 0.94. Additional tuning was also done by taking the best number of estimators and selecting best split criterion to enhance the model to an accuracy of 0.98.

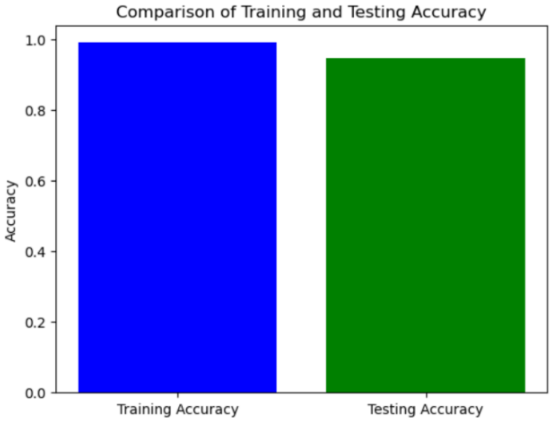


Figure 4.3-1 Comparison of training and testing accuracy of Random Forest model

The most productive fertilizer was then tested using factor analysis in the expected harvest other than the actual one. This was also important to offer tangible guidelines on how to increase yield of crops.

Lastly, the developed Random Forest model was exported as a saved pickle file and then incorporated into a smartphone application. The backend was designed using Flask to run the prediction model and the frontend was designed using React Native for better user interfaces. Such combo cooperation enabled company admins to input data and get a prediction in real-time via the mobile application.

4.3.2 Testing

When testing the actual harvest prediction model, then only the last year yields data was used in the analysis to compare the results to the predictions made. Comparing the predicted harvest values for the model with the actual harvest, it is evident that the predictions obtained from the model are relatively accurate. This validation step was important in assuring That the machine learning model, which was established through the Random Forest algorithm was very accurate. Part of testing required the comparison of the predicted harvest with actual harvest numbers to easily come up with variances and guide on areas that could still be optimized.

Besides evaluating its capability of predicting accurately, the standalone mobile application to the visitors, accessibility and functionality of the final application was inspected particularly by the company admins, who are the key user group of the application. Given that the prediction function in the application is designed for company admins who should have the data inputs anyway, their feedback was instrumental. While testing the software, the admins were able to enter actual values for soil pH, N-P-K, rainfall, temperature and fertilizer mix fertilities to achieve yields to predict the harvest. Before moving to this testing phase, every design was done online, thus giving practical knowledge concerning the usability of the application. Having gathered feedback from the company admins, the emphasis was made on the inputting of data, the readability of the predictions, as well as the usability of the system.

To attain this goal, several loops of usability feedback were integrated to practice its manufacturability as well as its functionality. The authors ensured that the user interface of the application was simple and easy to use thus minimizing the complexity of data input and prediction viewing for non-technical users. If the admins had some concerns during testing, for they might include small UI pop-ups on the application or performance problems, then the same would be treated at other iterations of the application. Such approach to the application testing enabled the company to improve on the application and thus meeting the need of the company.

5 RESULTS AND DISCUSSION

5.1 Results

The identification of variable influencing actual harvest output, and the trained machine learning model was further evaluated using dataset obtained from this company, with key parameters including percentage pH, N-P-K, Rainfall, Temperature and Fertilizer. Finally, the Random Forest model that was chosen out of all the models due to its relatively high performance was tweaked and obtained 0.94 of accuracy. This was much higher than the other models tried out such as Linear Regression, Decision, Tree and SVM models.

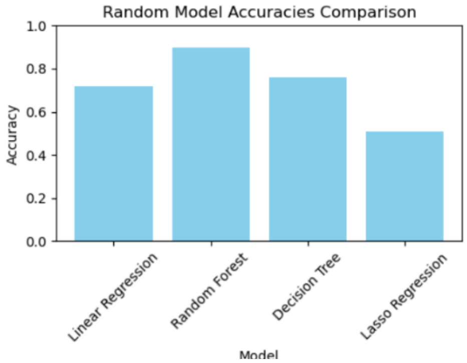


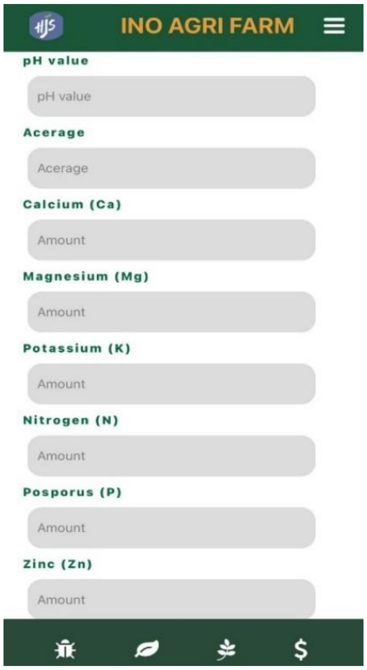
Figure 5-1 Comparison of the models

Further, factor analysis suggested the most appropriate combinations of fertilizers with which expected and actual yields could be aligned. Recommendations regarding the quantities of the specific fertilizers like Urea, TSP, MOP and CaNO3 were given to increase yields in the coming crops.

Most of the envisaged ideas were implemented in the application for company admins; they managed to implement the prediction model. They were also able to enter real time data to get timely harvest estimations by the admins. Actual values obtained through testing were nearly equal to the level of the predicted harvest which confirms the effectiveness and accuracy of the created scheme.

5.2 Discussion

This research work is proposed to design an accurate model for anticipating actual harvest values in the gherkin farming process while suggesting ideal fertilizer blends using ML algorithms along with factor analysis. The Random Forest model that has been used for the prediction was found to have a good training accuracy of 0.94 adequate to capture the interactions between the input features including the soil pH level, nutrient availability, weather and use of fertilizers. These criteria were selected with the advice of experts; there was a stress on understanding a particular domain as critical to boosting accuracy in models taught by a machine.



INO AGRI FARM

pH value

pH value

Acerage

Acerage

Calcium (Ca)

Amount

Magnesium (Mg)

Amount

Potassium (K)

Amount

Nitrogen (N)

Amount

Posporus (P)

Amount

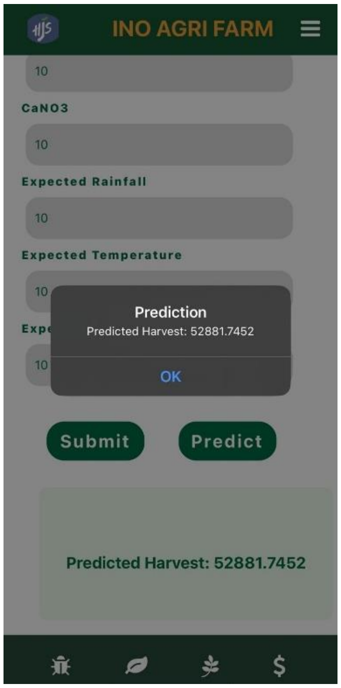
Zinc (Zn)

Amount

Figure 5.2-1UI interface of input form

This high accuracy show that the model can predict results, and this can easily be seen when using more than one variable to determine agricultural yield. pH, Nitrogen (N), Phosphorus (P), Potassium (K), and Urea, TSP, and MOP features helped significantly enhance the model’s performance. The incorporation of temperature as well as rainfall also added a lot of value to the accuracy of the model since these climatic conditions have a strong influence on productivity of gherkins.

This paper also featured a factor analysis where the recommendation of the most suitable fertilizer combination that could enhance the yield was determined. The analysis was beneficial in revealing workable recommendations about changing fertilizers to close the gap between estimated production and actual production. This is well illustrated in a commercial agricultural setting because the proper management of fertilizers tends to enhance output within a certain cost limit. These changes could assist farmers in the application of resources hence playing a role in developing sustainable agriculture.



10

CaNO3

10

Expected Rainfall

10

Expected Temperature

10

Expe

10

Prediction

Predicted Harvest: 52881.7452

Predicted Harvest: 52881.7452

Figure 5.2-2 Prediction UI

Adding value to the general system is the feature that creates the mobile application for company admins for entering data into the program and getting real-time predictions. The developed application can allow company admins to input normal variables such as the type of soil and climatic conditions which have the capacity to determine the expected harvest values together with the recommendations on fertilization. Another strength premised on this research is no loss of presence of the mobile technology and even machine learning integration into practical usefulness for industries.

More importantly, applying the model on actual datasets enhanced the perceived accuracy of the model. The differences between predicted harvests based on the developed model and actual harvests from the recent yields were in a small range further supporting the understandings that the proposed model could be used for decision-making. But feedback received from company admins during the thread was helpful in the decision making when it came to the modification of the application especially in term of the user interface and ease of managing the application. Improvements stemming from these findings could optimize its utility and attract more users to apply the identified technique.

However, certain issues that are loopholes in the model need to be considered so that they do not have decisive influences in the model. These include limitations such as inadequate disaggregate information of pest incidences which affect the harvest yield but not embodied in the current model. Moreover, the analysis based on the fertilizer management has contributed to the understanding of some important and relevant factors; however, for obtaining more reliable and accurate results, the model should include static environmental factors, for example, advanced weather prediction and real-time soil condition data analysis.

The following attributes make this application ideal for commercialization especially within the large-scale agricultural company such as Hayleys. Features that allow for immediate predictions and realistic suggestions using correct data might enhance national and organizational resource utilization, enhance production rates, and boost profitability. Furthermore, the solutions also have potential for scaling the application across different types of crops and regions making the solution applicable for the entire agriculture market.

As for the improvements, there is a few points that need to be considered to get one hundred per cent effective model and be helpful in simulation of real practice. For instance, use of IoT soil and weather sensor in real time will enable the model to update the predictions from the sensors as they are received. Moreover, installing models that allow identifying pest presence and disease occurrence would represent a more accurate solution for crop health and yield stability.

The provision of a functionality of the mobile application as a tool for both company admins and, in the potential future, the farmers themselves, expands the scope of the project. While the company admins input data while making strategic decisions currently, future versions might be for farmers. A simpler interface and easier input capture using internet of Things devices can help the small-scale farmers in that production region in making the key managerial decisions like use of fertilizers in the production of gherkin.

6 CONCLUSION

This research study has been able to show the application of machine learning — the Random Forest specifically in actual harvest yields in gherkin production. Selecting pH, N, P, and K as influential soil properties and important temperature and rainfall information along with proper combined fertilizers and nutrient the model had a training accuracy of 0.94. The predictions given help to offer a valuable tool for the company administrators to make effective and sound decisions for improving agricultural yields and the efficiency of using fertilizers.

An implies a substantial proportion of the study involved factor analysis which assisted in determining appropriate fertilizer combinations that can help to bridge the teaching learning gap in expectation and reality in harvest yields. As this part of the analysis shows, it has the important function of providing solutions to cultivate gherkin while optimizing productivity and reducing wastage. In this way farmers and the admins of the company can use data analysis to optimize the value of fertilizer, meaning reducing input and getting better yields, which in turn creates sustainable farming.

The development of a mobile application for company admins provided use of voice for data entry and simple user-friendly interface for the data input and the prediction about the harvest. For this reason, the integration of this technology in the working and functionality of the prediction model enables it to be active in real time and guarantees users proper decisions based on the most up to date data. The time spent on testing this application on a live client enriched the feedback collected, which in turn enhance the interface of the application.

However, the current study also offers directions for research in the future. More and real time data from pests control measures and better or more precise weather forecasts could increase the practicality as well as reliability of the system. Further development of the application for using by small scales farmers and including more simple user interfaces with accompanying automated data collecting systems will enhance the application's influence on the agricultural field.

7 REFERENCES

[1] S. Pruitt, "The Juicy 4,000-Year History of Pickles," History Education, [Online]. Available: https://www.history.com/news/pickles-history-timeline. [Accessed 28 July 2024].

[2] Studio Letsch & de Clercq, "I've always wanted to write about Gherkin," [Online]. Available: https://www.mirjamletsch.com/journal/gherkin. [Accessed 10 August 2024].

[3] N. Suresh, N. Ramesh, S. Inthiyaz and P. P. Priya, "Crop Yield Prediction Using Random Forest Algorithm," in *7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, 2021.

[4] N. Rale, R. Solanki, D. Bein, J. A. Vasko and W. Bein, "Prediction of Crop Cultivation," in *IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, 2019.

[5] M. Kandan, G. S. Niharika, M. J. Lakshmi, K. Manikanta and K. Bhavith, "Implementation of Crop Yield Forecasting System based on Climatic and

Agricultural Parameters," in *IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT)*, Visakhapatnam, India, 2021.

[6] H. ZENG, Z. LI and Z. ZHOU, "Comparative Study of Complex Parallel Factor Analysis and Parallel Factor Analysis," in *Prognostics and System Health Management Conference*, Qingdao, China, 2019.

[7] K. Han, E. G. Lee, H. Je and M. Y. Yi, "Introducing experiential knowledge platform: A smart decision supporter for field experts," in *International Conference on Big Data and Smart Computing (BigComp)*, Hong Kong, China, 2016.

[8] S. Wang, "Evaluation and application of fresh food safety through storage environment system based on factor analysis," in *International Conference on Networks, Communications and Information Technology (CNCIT)*, Beijing, China, 2022.

[9] T. Wang, "Quality Analysis of fixed assets based on factor analysis model," in *2nd International Conference on Economic Management and Model Engineering (ICEMME)*, Chongqing, China, 2020.

[10] Z. Ping and T. Yuanhua, "Digitizing construction method of large knowledge base system," in *International Conference on Computer and Computational Sciences (ICCCS)*, Greater Noida, India, 2015.