Productivity Analysis of Potato (*Solanum tuberosum)* Yielding in Cumilla, Bangladesh Based on Soil Chemical Parameters Using Machine Learning Approaches.

Submitted in partial fulfillment of the
requirements for the degree

of

Bachelor of Science in Information and Communication Engineering

by

Md. Kamrul Hasan 201-50-012
Anonto Mia 201-50-001
Yeasir Arafat Shohan 201-50-006

Under the Supervision of

Dipto Biswas
Lecturer, Department of Information and Communication Engineering

Daffodil International University

September 2024

© Daffodil International University

APPROVAL

This is to certify that the entitled **"Productivity Analysis of Potato (*Solanum tuberosum*) Yielding in Cumilla, Bangladesh Based on Soil Chemical Parameters Using Machine Learning Approaches"**, submitted by Md. Kamrul Hasan (201-50-012), Anonto Mia (201-50-001) and Yeasir Arafat Shohan (201-50-006) are undergraduate students of the Department of ICE has been examined. Upon recommendation by the examination committee, we hereby accord our approval of it as the presented work and submitted report fulfill the requirements for its acceptance in partial fulfillment for the degree of *Bachelor of Science in Information and Communication Engineering.*

BOARD OF EXAMINERS

Md. Taslim Arefin Associate Professor and Head Department of ICE Daffodil International University	**Chairman**
Prof. Dr. A. K. M Fazlul Haque Professor, Department of ICE Daffodil International University	**Internal Member**
Ms. Tasnuva Ali Assistant Professor, Department of ICE Daffodil International University	**Internal Member**
Dr. Engr. M. Quamruzzaman Professor and Head, Department of EEE World University of Bangladesh	**External Member**

© Daffodil International University

ii

© Daffodil International University

DECLARATION OF AUTHORSHIP

Project Title	Productivity Analysis of Potato (*Solanum tuberosum*) Yielding in Cumilla, Bangladesh Based on Soil Chemical Parameters Using Machine Learning Approaches.
Authors	*Md. Kamrul Hasan, Anonto Mia, Yeasir Arafat Shohan.*
Supervisor	Dipto Biswas, Lecturer, Dept. of ICE, Daffodil International University.

We, First Author, Second Author and Third Author hereby declare that this capstone project titled "**Productivity Analysis of Potato (*Solanum tuberosum*) Yielding in Cumilla, Bangladesh Based on Soil Chemical Parameters Using Machine Learning Approaches**" is entirely our own work, unless otherwise referenced or acknowledged. The content of this project is the result of our own research and efforts, and we have complied with all ethical guidelines and academic standards.

We are aware of the consequences of academic dishonesty and understand that any violation of ethical standards in this project may lead to disciplinary actions as defined by Daffodil International University's (DIU's) policies.

Kamrul

Md. Kamrul Hasan, 201-50-012

Anonto mia

Anonto Mia, 201-50-001

Shohan

Yeasir Arafat Shohan, 201-50-006

7/9/24

Supervisor

Dipto Biswas, Lecturer, Department of Information and Communication Engineering.

© Daffodil International University

iii

© Daffodil International University

ACKNOWLEDGEMENT

First and foremost, we express gratitude to Almighty Allah for this divine gift, which enabled us to successfully finish the final year project.

We are grateful and wish our profound indebtedness to **Dipto Biswas**, **Lecturer**, Department of ICE, Daffodil International University, Birulia, Savar, Dhaka - 1216. Deep Knowledge & keen interest of our supervisor in the field of "*Artificial Intelligence and Machine Learning*" to carry out this project. Endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the **Head of the Department, Md. Taslim Arefin, Dept. of ICE**, for kind help in finishing our project and also to other faculty members and the staff of the Department of ICE, Daffodil International University.

We express our special gratitude to **Dr. Md. Alamgir Siddiky (Chief Scientific Officer)**, **Dr. Md. Mokter Hossain Bhuiyan (Senior Scientific Officer)** expert in Soil Science and **Shamima Sultana** expert in Plant breeding of **Regional Agricultural Research Station, Bangladesh Agriculture Research Institute (BARI) Cumilla**. The immerse help of them for providing us the proper information of soil types, soil chemical parameter in different areas of Cumilla, weather data and validating the data make us to create an acceptable and legitimate dataset to work on. Also, their additional information helps us to increase the quality of our research.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

© Daffodil International University

ABSTRACT

The aim of this project is to apply machine learning methods to evaluate the productivity of potato yields in Cumilla, Bangladesh, based on soil chemical parameters. Collaborating with the **Regional Agricultural Research Station, Bangladesh Agriculture Research Institute (BARI)** in Cumilla, a comprehensive dataset spanning crop yields, soil quality, climate conditions, and region-specific agronomic practices has been developed. This dataset was subjected to 277 machine learning classifiers in order to determine the most effective techniques for predicting potato productivity. Based to the analysis, the top 25 classifiers including AdaBoost, GLMBoost, CNN, Simulated Annealing, Bayesian Optimization, GAN, Multi-Layer Perceptron, Support Vector Machine, FDA, Deep-Q-Network, Linear Regression, LDA, Ensemble Model, Autoencoder provided significant insights into the parameters influencing potato yield, with soil chemical characteristics emerging as key influences. Simulated Annealing, Bayesian Optimization, MLP, and CNN exhibit outstanding results with 90%, 89%, 88%, and 87% accuracy, respectively. despite this, GAN approaches and become the best match for the dataset with 99.96% accuracy. The correctness and applicability of the dataset to the regional agricultural environment were validated during the validation procedure. The relationship between crop yield and soil properties has become clearer because of to these discoveries, which have real-world implications for enhancing agricultural practices in Bangladesh. The verified dataset supports data-driven decision-making and sustainable development objectives by bridging a critical gap in the local agricultural research infrastructure.

Keywords: Machine Learning, Potato (*Solanum tuberosum*) Yielding, Productivity Analysis, Agriculture in Bangladesh, Soil Chemical Parameter.

© Daffodil International University

TABLE OF CONTENTS

© Daffodil International University

© Daffodil International University

List of Figure

© Daffodil International University

List of Tables

© Daffodil International University

Chapter 1

INTRODUCTION

Agriculture is the foundation of the global economy, and staple economical source in Bangladesh [1]. Sophisticated era of agriculture specially demands eco-friendly technologies regarding crop production. Farmers of Bangladesh utilize conventional farming methods, and are reluctant about modern farming technologies [2]. Bangladesh's geographic location makes it ideal for growing a wide variety of crops, and many of the country's regions produce an extensive variety of seasonal crops [3]. However, due to factors like weather, temperature, humidity, and soil quality, not all crops are capable of being cultivated in every region. Usually, the market has a variety of crops during the winter [4]. Bangladeshi farmers focus mostly with cultivating rice, wheat, potatoes, tomatoes, and After rice and wheat, potatoes are one of Bangladesh's major food crops [5]. Potato production is possible in all the agro-ecological areas of Bangladesh. Talking productivity, it comes second right after rice and it comes third when it comes to croplands area, behind rice and wheat. Hence, the minimum effort and cost needed for potato farming assures a high impact of this vegetable on the improvement of the Bangladeshi socioeconomic situation [6].

Nevertheless, there are several difficulties which unfortunately remain that the country has to cope up with in order to intensify the potato production [7]. Causes of these problems can be associated with a lack of knowledge about types of soil, yields of crops, weather, irrigation mistakes, incorrect harvesting, using insecticides improperly and many more. As a consequence, these problems can cause existing problems to become more serious, and raise the costs of farming [8]. As the population is growing every second, the industry of agriculture has to take it under consideration as well. Generally, there are major losses that set in since they range from choosing the crops to the lack of knowledge of technology to the sales of the products [9, 10]. The concept of machine learning is widely popular to solve any kind of problems using computer programming which allows different solutions of a single problem. It is the idea where high level language is used to communicate with the machine, while solving a problem using machine learning there need various algorithms and models [11, 12]. The algorithms and models work depending of the dataset provided by humans. Machine learning can be utilized for efficient agricultural processes (particularly potato growing) predictive modeling of weather and climate, real-time monitoring of climate, cultivation, and yield, precision farming for fertilization, weed detection and control, and variable rate application of

© Daffodil International University

chemicals, as well as disease and pest detection and decision support about operations or products [13, 14, 15, 16]. Machine learning can help farmers develop more efficient and sustainable agricultural systems, increase the resilience of crops and yield quality, and maximize soil management practices [17].

Many countries like America, Russia, China, Brazil, Italy are already using artificial intelligence and machine learning to their agriculture sectors and getting tremendous output [18]. In the past few years, the market of smart agriculture is impressively increasing. Being a developing nation, Bangladesh needs some time to adapt to the advanced use of machine learning in the agricultural industry [19]. As the base of Bangladesh's economy is agriculture so, for the betterment of economic growth as well as the agriculture sector it needs to improve the cultivation methods with the help of science [20, 21]. As per the statistics potato comes 3rd mostly grown crop after rice and wheat [21]. And the demand of potato is huge as it is a seasonal crop. Being a crop harvested in the winter, it risks numerous weather-related diseases. For that reason, it needs to be pampered and grown with extra care [22]. So, here machine learning can play an important role by identifying soil quality, humidity in the air as well as rain prediction, Ph of the field, nutrition's level of the soil, leaf disease detection using image processing and computer vision etc. [23, 24, 25]. Which can make the potato production more efficient and the quality of produced potato will be good and wastage will decrease. There are several types of ML models mostly used in agriculture those are:

- **Regression:** A statistical method for modeling and evaluating the connections between a dependent variable and one or more independent variables is regression analysis.
- **Clustering:** Classifying a collection of objects so that they are more similar to one another than to those in other clusters is the primary goal of clustering. Usually, a distance metric or set of predetermined criteria are used to measure this resemblance.
- **Bayesian models:** Inferring probabilities of unknown values from observable data and past knowledge is the primary goal of Bayesian models.
- **Neural networks:** Learning and modeling complex relationships and patterns in data is the primary goal of neural networks.
- **AdaBoost:** AdaBoost (Adaptive Boosting) aims to increase the accuracy of weak learners (usually decision trees or other simple classifiers) by training them repeatedly on the same dataset while giving previously incorrectly identified cases a larger weight.

© Daffodil International University

- **GLMBoost:** The primary goal of Generalized Linear Model Boosting is to increase predictive accuracy by the stage-by-stage forward sequential combination of weak learners, which are usually regression models.

So, developing machine learning models to predicted potato yields for a particular time of the year or region by analyzing previous data on weather patterns, soil conditions, irrigation techniques, and yields [26]. In order to automatically detect and identify pests, diseases, and other dangers to the crop, computer vision models are trained on photos of both healthy and diseased/infested potato plants [27, 28]. Timely procedures can prevent extensive damage by enabling detection at an early stage. Analyzing soil samples and satellite data using machine learning algorithms allows one to identify the pH, moisture content, and nutrient levels of the soil [29]. Application of fertilizer, timing of irrigation, and other soil management procedures are guided by this information. With the use of numerous atmospheric variables and historical weather data, those algorithms can be trained to produce precise weather forecasts that are region- or farm-specific. This can support farmers in their activities, such as planting, irrigation, and harvesting, more effectively [30, 31, 32, 33].

1.1 Project Background

Significantly studying the state of agricultural productivity today, especially in relation to Cumilla, Bangladesh, this chapter provides background information for the research. The first part of it looks at the current situation and the main obstacles to the best possible potato production in the area, with a focus on the crucial role that soil chemical factors play. Subsequently, the chapter delves into a comprehensive assessment of the literature, scrutinizing current research and methodology concerning soil analysis and the utilization of machine learning techniques in agriculture. In addition to capturing the state of the art in these fields through this review, the chapter also highlights the research gaps and difficulties that this study attempts to solve, supporting the necessity of the novel strategies suggested in this thesis.

1.1.1 Current State and Limitations

1. The paper by Saifullah Omar Nasif et al. [1] in “A SURVEY OF POTATO GROWERS IN BANGLADESH: PRODUCTION AND CHALLENGES” relies solely on farmer interviews through a questionnaire. There is no triangulation of data through other methods like field observations, yield measurements, insect sampling etc. Here data is entirely based on farmer perceptions, which can be subjective. This analysis is mostly descriptive statistics like frequencies and percentages. There is a lack of deeper

© Daffodil International University

statistical analysis to look at correlations, differences across regions/varieties etc. While this study provides a broad overview from the farmer perspective, there are opportunities to strengthen the methodological rigor, statistical analysis, literature grounding, and specific actionable outcomes from the findings. An integrated multi-disciplinary approach could make the research more impactful.

2. Abdullah-Al-Faisal et al. [2] in "Assessment and prediction of seasonal land surface temperature change using multi-temporal Landsat images and their impacts on agricultural yields in Rajshahi, Bangladesh" examined seasonal (summer and winter) LST transitions, distributed changes in Land Use/Land Cover (LULC) (from 2000 to 2020 at 10-year intervals), and projected these two variables (for 2030 and 2040) using Artificial Neural Network (ANN) and Cellular Automata (CA) algorithms using Landsat images. In furthermore, studies including focus groups and key informant interviews (KIIs) were carried out to find out the impact of rising temperatures on agricultural output, and the most productive regions were identified by a suitability analysis based on the analytic hierarchy process (AHP). Using ANN and CA in their investigation, the output was insufficient to be accepted. If they could have used additional models and compared their accuracy, the outcome could have been acceptable.
3. In the research of Debabrata Sarkar et al. [3] in "Modelling agricultural land suitability for vegetable crops farming using RS and GIS in conjunction with bivariate techniques in the Uttar Dinajpur district of Eastern India" the efficiency of the architecture, potential uncertainties, and limitations are not fully explored. The model training process, including the selection of predictive variables, validation of the model, and model optimization method, is not fully explained in the work. The study's findings may not be dependable or repeatable as they may be due to this methodological approach's lack of transparency. These difficulties must be appropriately taken into consideration in order ensure the durability and applicability of the models that are suggested. It is important that the study take into consideration seasonal fluctuations, hydrogeological processes, and long-term climatic patterns.
4. Abhijit Shaha et al. [5] in their research "Crop Diversity and Cropping Patterns of Comilla Region" does not provide a comprehensive discussion on how the findings from the Comilla region could be extrapolated or applied to other regions or at a national level. The study does not discuss the potential implications of using relatively old data or the need for regular updates to capture the dynamic nature of agricultural

© Daffodil International University

systems. There is no mention of measures taken to validate or cross-check the data provided by the local officials, which could lead to potential biases or inaccuracies. Also, the sampling method and the criteria for selecting the Sub-Assistant Agriculture Officers who filled out the questionnaires are not clearly described, which could impact the representativeness of the data.

5. Muktasha Deena Chowdhury et al. [6] in "Problems and Prospects of Potato Cultivation in Bangladesh" fails to get into detail on how the data was analyzed. Better analytical techniques, both quantitative or qualitative, could make the study stronger. Additionally, there doesn't appear to be much previous research on the economics of potato growing in Bangladesh covered in the brief literature review section. Contextualizing the study's aims and conclusions could be improved with a more comprehensive evaluation.
6. Patil et al. [16] in "Smart agriculture using IoT and machine learning", outlined a smart agriculture system integrating IoT sensors and actuators for real time data collection. The authors mentioned the system's superiority in which these objectives can be achieved, including enhanced productivity, resource conservation, precision agriculture, early problem detection, decision support, automation, and remote accessibility. However, it misses implementation details, for example, the machine learning algorithm used, datasets used for training, and performance evaluation, which limits its ability to fully appraise the solution's effectiveness in real-world applications.
7. Kasara Sai et al. [12] in "IoT based smart agriculture using machine learning", presents a smart irrigation system that utilizes machine learning to predict crop water requirements. A decision tree algorithm is used to predict the water requirements. The farmers receive the results as email alerts, which will help them to make the right decisions regarding water availability and avoid wastage. Although it does not provide any specific details about the implementation, performance evaluation and algorithms accuracy.
8. Koshkarov et al. [14] in "Machine learning methods in digital agriculture: algorithms and cases", suggested the machine learning algorithms that are used for data analysis in agriculture, such as soil analysis, weather forecasting, and vegetation analysis. The author first reviewed the literature and conducted a survey of farmers to select the most appropriate algorithms. The results demonstrate that the algorithms can perform tasks such as soil classification, fertility prediction, climate change simulation, and crop yield

© Daffodil International University

forecasting. Nevertheless, this paper does not discuss the limitations or challenges that arise with the implementation of these algorithms in actual agricultural settings.

9. S. Mughele et al. [20] in “Roles of IoT, big data and machine learning in precision agriculture: a systematic review” suggested a systematic review of the functions of the Internet of Things, big data analytics, and machine learning in precision agriculture. The authors used the PRISMA guidelines and carried out a literature search in different databases. The results show not only the benefits and drawbacks of precision agriculture technologies, but also the optimization of natural resources, increased yield, and ecological protection. However, the report did not provide a detailed evaluation of the specific shortcomings or challenges involved in applying these technologies in different agricultural settings.

1.1.2 Literature Review

1. L. Venkateswara Reddy et al. [7] in “Applying machine learning to soil analysis for accurate farming” analyzes many system learning approaches to soil type classification. Neural network selection approaches using the assist vector machine (SVM)tree and naïve Bayesian techniques are presented and evaluated in this category. The source of the soil dataset is the actual data. The use of random forest and decision trees algorithms is mentioned in the report; however, SVM with a linear feature kernel performs better. The fine accuracy of the SVM is 82.35%. A major weakness is the fact that, no experimental findings, model performance measurements, or visualizations are provided. It is difficult to evaluate the suggested approach's efficiency in comparison to established methodologies or baseline procedures in the absence of them as well. The study could become greatly more convincing and emphasize the practical implications if it included one or more real-world case studies showing the successful use of the machine learning models.
2. Cedric Lontsi Saddio et al. [4] in “Crops yield prediction based on machine learning models: Case of West African countries” implemented a machine learning-based prediction system that predicts the yield of six crops at the national level in the West African countries throughout the year; rice, maize, cassava, bananas, and seed cotton. Here, researchers combined meteorological, climatic, agricultural yield, and chemical data. Decision tree, multivariate logistic regression, and k-nearest neighbor models were used in the development of this system. Also applied a hyper-parameter tweaking technique during cross-validation to get an improved model free from overfitting. The

© Daffodil International University

decision tree model outperforms the logistic regression and K-Nearest Neighbor models With a COD of 95.3%. While the authors combined several data sources, the overall dataset size is not clearly stated. Machine learning models often benefit from larger training datasets to improve accuracy. Also use a limited set of features and less powerful models. More complex models may achieve higher accuracy.

3. Mahmudul Hasan et al. [8] in "Ensemble machine learning based recommendation system for effective prediction of suitable agricultural crop cultivation" suggests using an ensemble machine learning technique known as K-nearest Neighbor Random Forest Ridge Regression (KRR) to precisely predict Bangladesh's production of three main agricultural crops: wheat, rice, and potato. Raw data on environmental factors, cultivation areas, and crop production amounts for five major crops (Aus rice, Aman rice, Boro rice, potato, and wheat) in Bangladesh from 1969 to 2021 from various government organizations was collected. Support Vector Regression, Naive Bayes, Ridge Regression and two ensemble ML models (Random Forest and CatBoost) are used. KRR outperforms the other models in predicting production of Aus, Aman, Boro rice, potato and wheat, achieving low error rates (e.g. 0.9% MSE, 99% R2 for Aus rice prediction). The Diebold-Mariano test shows KRR is significantly superior to the benchmark models at 1% and 5% significance levels. Further research may focus on enhancing the model's generalization for different crops, seasons, and geographical areas. This could be accomplished by including more diverse data sources to the dataset. Integrating multi-criteria decision-making principles into the model is another possible path for future research. This could assist in choosing the best crop varieties and production techniques for different agricultural conditions.
4. A. Kumar et al. [25], in "Machine Learning and Deep Learning for Soil Analysis and Classification of Micro and Macro Nutrient Using IOT" applied deep learning and machine learning models to evaluate soil fertility by focusing on the main and micronutrients found in the soil. In this work, IoT sensors are used to collect real-time agricultural data, which is then preprocessed and analyzed using machine learning and deep learning models. For deep learning, the study uses LSTM and Artificial Neural Networks, and for soil classification, it uses ML models like KNN, SVM, and RF. The potential of machine learning and deep learning techniques in effectively classifying soil fertility based on nutrient levels is demonstrated by the high accuracy rates of SVM, RF, and ANN models. Despite its slightly lower accuracy of 90%, the KNN model offers insightful information about soil classification. The LSTM model, with an

© Daffodil International University

accuracy of 51%, shows room for improvement in capturing complex patterns in soil data.

5. S. Mourtzinis et al. [27], in the paper entitled "Advancing Agricultural Research Using Machine Learning Algorithms," goal of this study is to pinpoint intricate relationships between different elements that influence crop yields but are frequently missed in conventional agricultural research. The authors stress that their strategy is meant to produce fictitious experimental data for investigating G × E × M interactions, not to directly forecast crop yield. The datasets, which were gathered between 2016 and 2018 for maize and 2014 and 2018 for soybean, include 17,013 entries for maize and 24,848 for soybean. The XGBoost algorithms that were developed demonstrated a notable level of precision in yield estimation. The average yield (13,340 kg/ha) for maize was found to have an absolute error (MAE) of 3.6% and a root mean square error (RMSE) of 4.7%. This degree of precision shows that the model can consistently forecast yields depending on the input variables, which makes it an important resource for both farmers and researchers.
6. Sandeep U. Kadam et al. [29] In "Machine Learning Method for Automatic Potato Disease Detection", have proposed a deep learning framework to detect potato leaf diseases. They used several pre-trained CNN models such as VGG16, ResNet50, and GoogleNet to classify images of healthy and diseased potato leaves from open-source datasets. The results indicated that the Irregular Forest Classifier yielded the highest accuracy at 97% while the implementation of CNN models yielded promising results. However, some of the limitations are the reliance on high quality image data as well as the requirement of significant computational power for model training and predictions.
7. M. I. Tarik et al. [18], in "Potato Disease Detection Using Machine Learning" presents an in-depth analysis on the detection of potato illnesses using image processing and machine learning approaches. The researchers gathered a dataset of 2,034 photos of potato leaves, dividing them into seven categories of illnesses, including potato leaf roll virus and soft rot. Using back propagation neural networks as part of the process, 92% classification accuracy for disease identification was achieved. Furthermore, the research delves into diverse machine learning models, such as transfer learning and support vector machines, which have proven to have remarkable diagnostic accuracy, with results ranging from 84.6% to over 99% for potato diseases.
8. Padarian et al. [39] employed machine learning methods to revise the soil science literature. They discovered that the use of machine learning techniques in soil science

© Daffodil International University

is increasing, mostly in developed countries. Furthermore, they concluded that more developed ML techniques usually yield improved results compared with simple techniques. Nevertheless, the limitations include that the interpretability of ML models may sometimes be complex.

9. Akhter et al. [28] explored the use of IoT sensors and data analytics in farming. The authors suggested a data analytics and machine learning model for predicting apple illnesses. They conducted a survey to determine the difficulties farmers face in the process of adopting these new technologies. The paper addresses challenges associated with the adoption of precision agriculture technologies, but emphasizes that the advantages are more than the constraints.

1.2 Project Definition

This study's main goal is to use machine learning techniques to analyze and forecast the productivity of potato harvests in Cumilla, Bangladesh, based on a variety of soil chemical indicators. The objective is to create an effective forecasting model that will help farmers optimize their farming methods for higher yields. Comprehensive data on important soil chemical parameters, including pH, Nitrogen (N), Phosphorous (P), Potassium (K), organic matter, moisture content, and other pertinent variables, are gathered as part of the study in order to accomplish this. To provide a foundation for model training and validation, historical data on potato yields from various fields in Cumilla are being collected. To improve the accuracy of the model, other environmental elements are also be included, such as meteorological data like temperature, humidity, and rainfall. Advanced machine learning techniques are used in the study to evaluate the data and find patterns and relationships between potato yields and soil properties. A strong predictive model that can provide farmers with practical advice and insights to increase crop productivity is the anticipated result.

1.2.1 Problem Statement

In Bangladesh, crop production is facing many challenges day by day, as the climate is changing continuously [19]. The geographical location of Bangladesh helps the country to produce a variety of crops but also the location is being cause of disaster in Bangladesh’s agro-ecological sector. global pollution and climate change is directly affecting Bangladesh as it is a delta country [20]. Temperature and rainfall changes have already affected crop production in many parts of Bangladesh. The effects of climate change will cause a 4°C rise in temperature, and being the 7th largest potato production country that will negatively affect Bangladesh's

© Daffodil International University

food supply [21]. Potatoes are adversely affected by temperatures in excess of 30°C in a number of ways, such as reduced starch partitioning, brown patches, slowed or absent tuber dormancy, which causes premature blooming. potatoes prefer temperatures below 25°C. It can survive in the 25-30°C range. Furthermore, temperatures above 30°C make potato cultivation problematic [22, 23].

The region (Comilla) has a range of soil types, from loams to clays, with varying degrees of acidity. It is generally low to medium fertile [24]. If, in that situation the weather and the parameters which are required to cultivate potato properly are not considerable then there come difficulties to match expected output [25]. Besides the population are increasing rapidly, but the number of proper agricultural land are decreasing which create pressure on the farmers, at that point they cannot take proper decisions and also have not much time to think about the right parameters of yielding field, suitable potato plant, right cultivation time and proper nursing [26].

1.2.2 Context and Scope

In the era of growing agriculture, it is more important to be concerned about the quality of crop as well as the cultivation technique and knowledge of the new technologies. Here machine learning can make a great impact [34]. In the past few years Science is making tremendous progress in every sector and using machine learning in agriculture could be a valuable move [35]. Bangladesh relies heavily on potatoes as a primary crop, which boosts the nation's agricultural economy. Because of its reputation for agricultural production, the Cumilla region (sometimes called Comilla) is a prime location for potato farming. However, because of things like improper fertilizer application, insufficient irrigation, and poor soil management, conventional agricultural practices frequently result in yields that are below ideal [36, 37]. With the development of cutting-edge technology, machine learning (ML) is a viable way to deal with these issues by identifying optimal practices and evaluating several aspects that influence crop productivity [38]. Using machine learning techniques, the aim of this initiative is to identify the lands in Cumilla that are most suitable for growing potatoes. The main research questions that this study seeks to answer are:

- What are the main elements that determine whether or not potato production is suitable in Cumilla?
- How can machine learning models be used to analyze these factors?

© Daffodil International University

- What are the advantages and disadvantages of using machine learning to analyze field suitability?

Important variables including soil quality, topography, climate, water availability, and the prevalence of pests and diseases are going to be the main emphasis of the study. In addition to climate factors like temperature, humidity, rainfall, and seasonal fluctuations, soil quality criteria including pH levels, nutrient content, texture, and organic matter will be taken into consideration [39, 40, 41, 42]. Also includes the assess topographic characteristics including height, slope, and drainage patterns in addition to their closeness to irrigation infrastructure and water sources. The analysis will also include past information on disease and pest outbreaks. The application of machine learning techniques will entail gathering and preprocessing data from a variety of sources, including meteorological stations, soil testing, and remote sensing. In order to predict field appropriateness, pertinent features will be chosen and used to train and evaluate machine learning models, such as decision trees, random forests, and support vector machines. Metrics including the F1 score, accuracy, precision, and recall will be used to assess the performance of these models.

The development of suitability maps showing the best and worst fields for potato farming is one of the research's anticipated outcomes. Furthermore, the development of a decision support system will help farmers and policymakers choose wisely when it comes to land use and management strategies. Through increasing crop yields, reducing environmental impact, and maximizing field utilization, this research seeks to advance sustainable agriculture.

Despite this, it is imperative to tackle obstacles like data accessibility and quality, interpretability of models, and flexibility in response to evolving circumstances to guarantee the effectiveness and pragmatic implementation of suggested remedies. The correctness and dependability of the results will be guaranteed by conducting data validation in conjunction with the **Regional Agricultural Research Station, Bangladesh Agriculture Research Institute (BARI) in Cumilla, Bangladesh**. This collaboration will contribute to the validation of the gathered data and the models developed, giving the study findings and suggestions a strong basis.

1.2.3 Significance and Implications

Currently, potatoes are Bangladesh's most successful winter crop, grown from October to March [2]. Its production value is second only to paddy rice, with estimates of $560 million in 2005. Harvesting 4.3 million tons of potatoes in 2007, 12 more than in 1961 the Bengali farmers

© Daffodil International University

ranked fourth in Asia and 14th globally in terms of the production of potatoes. The last four years have seen an average of 10 million MT of potatoes produced in the nation [3, 4, 5, 6]. The General Directorate of Food (DGoF) reports that, in comparison to its 8.95 million tons of production in FY 2014, the nation's yearly need for potatoes is between 6.5 and 7.0 million tons [5, 6]. In contrast, the Department of Agriculture Extension asserts 1.12 crore tons of potato production in FY 2022-2023 whereas the Bangladesh Bureau of Statistics claims 1.04 crore tons. In accordance with estimates from the Department of Agricultural Marketing, 90 lakh tons of potatoes are consumed annually in the nation [6, 7, 8]. According to that information, there is a growing market for potatoes and a growing demand for them. So, having a suitable cultivation field is the first step towards increasing potato production, and this research will help in meeting those requirements [9].

In conclusion, there is a chance that the study of "Productivity Analysis of Potato (*Solanum tuberosum*) Yielding in Cumilla, Bangladesh Based on Soil Chemical Parameter Using Machine Learning Approaches" will have a big impact on social welfare, economic development, environmental sustainability, and policy making. This study can help Bangladesh and other regions establish more resilient, sustainable, and efficient farming practices by utilizing cutting-edge technologies and data-driven methodologies.

1.2.4 Quantification

Some graph and statistical information with detailed discussion are presented here to demonstrate the worldwide market value of smart agriculture by the year of 2018-2028 [48], Top potato Production Countries [49] and Trend of potato production in Bangladesh [50].

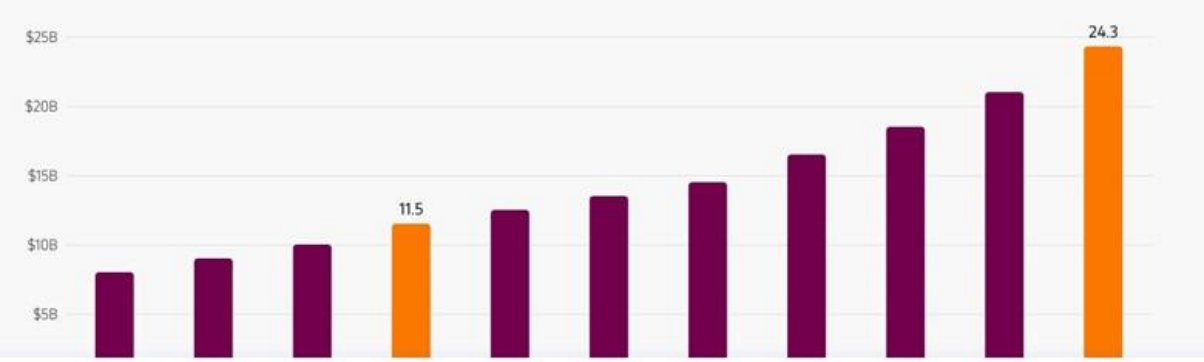

Fig-1: Smart agriculture market by value, 2018-2028 [48].

Fig-1 shows the market values for smart agriculture and demonstrates a bar graph from 2018 to 2028, spanning a ten-year period. Every bar on the graph signifies the market value,

© Daffodil International University

expressed in billions of dollars (B), for a particular year. The market worth was about $5 billion as of 2018. This value grew yearly, to approximately $6 billion in 2019 and $7 billion in 2020. The market value increased to about $8 billion by 2021, then to $9 billion in 2022 and $10 billion in 2023. The market value is projected to be $11.5 billion in 2024 and is indicated in orange, potentially denoting an important milestone or projection. The market value is expected to expand by around $14 billion in 2025, $16 billion in 2026, and $18 billion in 2027. The orange-highlighted market value projection for 2028 is $24.3 billion. The graph shows that the market value of smart agriculture has increased significantly and consistently over the past ten years, reaching a growth of more than four times between 2018 and 2028.

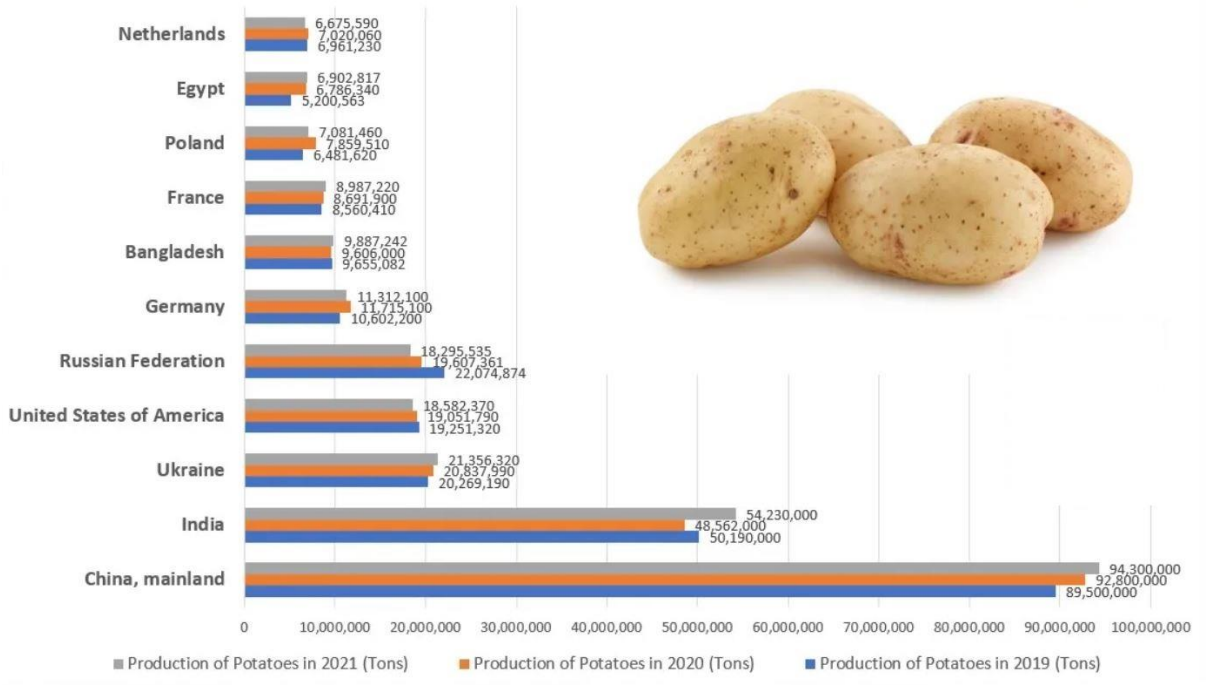

Fig-2: Worldwide Top potato Production Countries [49].

Fig-2 shows the total quantity of potatoes produced in different nations in 2019, 2020 and 2021. Specifically focusing on Bangladesh, the country produced 9,887,242 tons of potatoes in 2021 a rise over the previous years. 2020 saw a marginal decrease in production, totaling 9,660,000 tons. The total production in 2019 was 9,655,082 tons. According to the data, Bangladesh's potato production increased steadily over the span of the three years, demonstrating a generally positive trend in the nation's potato agricultural sector.

© Daffodil International University

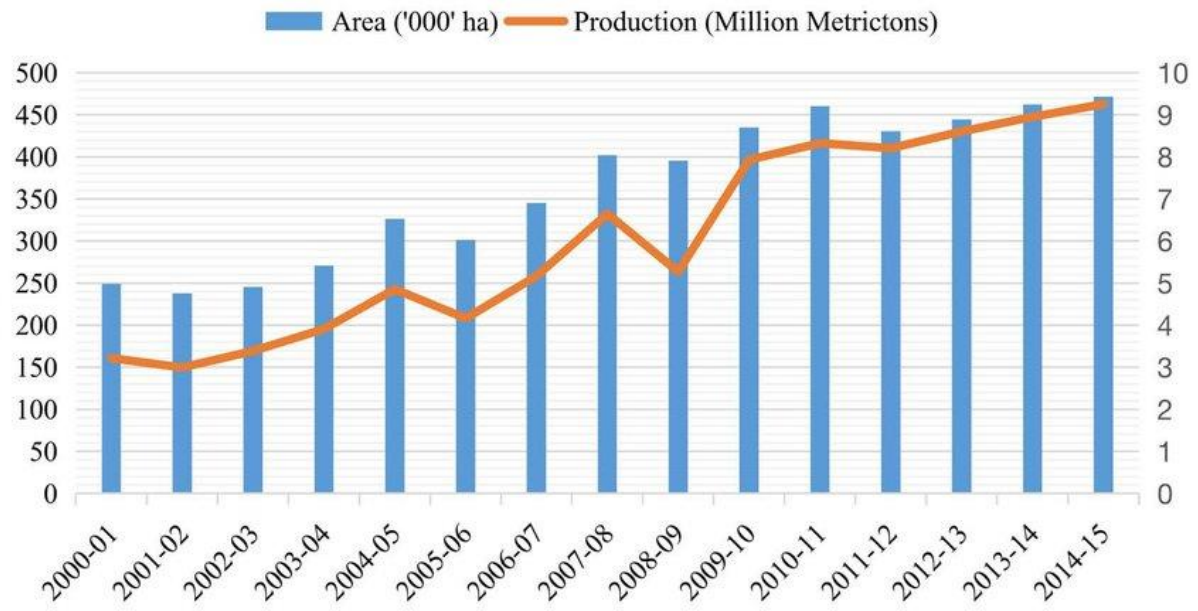

Fig-3: Trend of potato production in Bangladesh [21].

The graph in Fig-3, which shows the area under cultivation as well as the overall production, shows the trend of potato production in Bangladesh from the 2000 to 2001 growing season to the 2014-15 growing season. The area under cultivation begins at about 250,000 hectares in 2000-01 and is shown by blue bars, which are measures in thousands of hectares. There have been notable variations over the years, with notable spikes in 2005-06 and 2006-07 seasons that exceeded 300,000 hectares. The area surpasses 400,000 hectares during the 2007-08 season, when it reaches its high, and then declines in the 2008-09 season. After thereafter, the region fluctuates a little but normally stays above 300,000 hectares. Toward the end, it climbs once more, reaching roughly 400,000 hectares in the 2014-15 season. In contrast, the production of potatoes, which is shown as an orange line and is expressed in million metric tons, starts at about 3 million metric tons in 2000-01. With a few small oscillations, production is increasing steadily, but generally in an upward direction. The 2007-08 season sees a sharp increase in production, peaking at roughly 6 million metric tons, which corresponds with the area under cultivation. Production started to rise in 2008-09, but it did so more slowly. By the 2014-15 season, it had reached almost 9 million metric tons. Overall, during the course of the 15-year period, the graph shows a strong positive trend in Bangladesh's potato production as well as the area under cultivation.

© Daffodil International University

1.3 Project Contribution

In association with the **Regional Agricultural Research Station, Bangladesh Agriculture Research Institute (BARI)** in Cumilla, this thesis project aims to fill a significant need in the field of agricultural research by creating and certifying a particular dataset. This project represents a major advancement in the field of agricultural science, especially in Bangladesh, where accurate and localized data are essential for developing policies and making well-informed decisions. This research ensures the accurateness and effectiveness of this dataset to the local agricultural environment by meticulously developing and validating it. This dataset covers factors including crop yields, soil quality, climate conditions, and region-specific agronomic techniques. In addition to upholding strict scientific standards, the validation process conducted in collaboration with **BARI** validates the dataset's applicability and utility in addressing important agricultural concerns that local farmers and policymakers deal with. Furthermore, the availability of such a validated dataset covers a significant gap in the infrastructure right now in place for agricultural research, allowing for the advancement of future investigations and the implementation of evidence-based interventions targeted at improving Bangladesh's agricultural resilience, sustainability, and productivity.

An extensive variety of 277 machine learning classifiers have been applied to analyze the dataset, providing a thorough assessment of different machine learning methods. The highest performing 25 classifiers have been selected after an exhaustive search, offering significant new information about the best techniques for forecasting potato productivity in this region. The selection emphasizes how machine learning can be applied to improve agricultural practices. The study also identifies important variables that impact the yield of crops, furthering the comprehension of the connection between soil chemical characteristics and potato productivity. So, the contribution on this research can be highlighted as below.

- **Understanding a Research Gap**: Focusing on Bangladesh's preference for exact localized agricultural data, which is essential for formulating policies and making decisions. Also, uncovers significant factors influencing the yield of crops, expanding the scope of the connection between potato productivity and soil chemical properties.
- **Comprehensive Dataset**: Incorporating significant factors like local agronomic methods, soil condition, climate, and agriculture production. This research has had an enormous impact in developing a genuine and distinctive dataset. The dataset has 2200 data for localized agriculture in the Cumilla region, with 7 features in total.

© Daffodil International University

- **Collaboration With BARI:** Working along with the **Bangladesh Agriculture Research Institute (BARI)** in Cumilla to create and assess a customized dataset. and makes sure the dataset remains accurate and useful by applying strict scientific norms and BARI validation.
- **Machine Learning Analysis**: The dataset was subjected to 277 machine learning classifiers, of which the top 25 were chosen based on their performance to determine the most accurate predicting methods for potato productivity.

These findings not only advance the knowledge of the interplay between soil and crops, but they also have real-world applications for raising agricultural productivity through data-driven decision-making. With the increasing focus on sustainable development objectives and the requirement for data-driven approaches to mitigate the effects of climate change on agriculture, this contribution is especially significant. As a result, this thesis research contributes significantly to the methodologies and applications of agricultural research, highlighting its potential to inform and direct practices, investments, and agricultural policies targeted at promoting sustainable agricultural development in Bangladesh and other agricultural contexts around the world.

1.4 Project Specification

The experimental configuration required to carry out this study is included in this section of the project specification. Additionally, it certifies the data set's legality by referencing resource validation and providing accurate information.

1.4.1 Experimental Set-up

The experiment is conducted in a Microsoft Windows 10 Home Single Language, Version 10.0.19045 Build 19045 with Processor AMD Ryzen 5 3500U with Radeon Vega Mobile Gfx, 2100 MHz, 4 Core(s), 8 Logical Processor(s) and 8gb RAM. In order to load the dataset, check the data, pre-process and run the machine learning models the google Collaboratory platform is being used with its provided GPU. To perform machine learning models, some important libraries are used like; NumPy, pandas, scikit-learn, TensorFlow, Keras, seaborn, catBoost, XGBoost and corresponding library of a particular model. The used parameters are collected with using some machine, like pH of the soil is collected by pH meter, potassium and phosphorus level is determined by using spectrophotometer, values for humidity are obtained by humidity sensor machine and the number of rainfalls is collected by rain-gauge machine.

© Daffodil International University

1.4.2 Dataset

A comprehensive summary of the features, target labels, ranges, and measurement techniques found in the below Table-1 which partially represent the information of the dataset. The section under "Feature Details" lists several soil and environmental factors that are essential for crop analysis. These features include pH, which indicates the acidity or alkalinity of the soil; N, which indicates the nitrogen content necessary for plant growth; K, which indicates the potassium content critical for plant health and disease resistance; and P, which is essential for plant photosynthesis and energy transfer. It also comprises Temperature, which indicates the air temperature that affects plant growth and development; Rainfall, which quantifies the amount of rainfall that affects soil moisture levels; and Humidity, which represents the moisture content in the air that influences plant transpiration. There are 2200 data points available for each of these features.

Table 1: Details Information of The Dataset

Feature details		Target Label Details		Range of Features	Measurement Machine Name
Name of Features	Total Numbers	Name of Target Label	Total Numbers		
pH	2200	Rice	275	3.50-9.94	pH Meter
Nitrogen(N)	2200	Corn	275	0-140	NPK Meter
Potassium(K)	2200	Watermelon	275	5-205	Spectrophotometer
Phosphorus(P)	2200	Potato	275	5-145	Spectrophotometer
Humidity	2200	Cabbage	275	14.26% to 99.98%	Humidity Sensor Machine
Temperature	2200	Cauliflower	275	8.83°C to 43.68°C	Liquid-In-Glass Thermometer
Rainfall	2200	Jute	275	20.21mm to 298.56mm	Rain-gauge Machine

The precise crops for which recommendations are being made based on the features are listed in the "Target Label Details" section. Rice, corn, watermelon, potatoes, cabbage, cauliflower, and jute are some of these crops.

© Daffodil International University

The dataset contains 275 data points for each crop. The minimum and maximum values for each feature are provided in the "Range of Features" section, which aids in understanding the limits and variability of the dataset. As an illustration, the ranges for pH, nitrogen, potassium, and phosphorus are 3.50 to 9.94, 0 to 140, and 5 to 145, respectively. There is a 14.26% to 99.98% humidity range, an 8.83°C to 43.68°C temperature range, and a 20.21mm to 298.56mm rainfall range. The instruments used to measure these properties are finally listed in the "Measurement Machine Name" section. A spectrophotometer is used to measure phosphorus and potassium; an NPK meter measures nitrogen, potassium, and phosphorus; a humidity sensor machine measures humidity; a liquid-in-glass thermometer measures temperature; and a rain gauge measures rainfall. This extensive table guarantees that all pertinent information is gathered and precisely measured in order to provide appropriate productivity recommendations.

GitHub link of the dataset: Soil-Weather-dataset/Crop_recommendation-2.csv at main · hasan-10/Soil-Weather-dataset (github.com)

1.4.3 Data and Resource Validation

To ensure the accuracy and reliability of the study on " Productivity Analysis of Potato Yielding (*Solanum tuberosum*) in Cumilla, Bangladesh Based on Soil Chemical Parameters Using Machine Learning Approaches " therefore, data collection and validation from organizations directly involved in the agricultural sector are crucial. A significant role of this process is played by the **Bangladesh Agriculture Research Institute (BARI)** in Cumilla, which is a central authority in Bangladeshi agriculture. The research has been enhanced better and more pertinent by **BARI** and its senior scientists Dr. Md. Alamgir Siddiky (Chief Scientific Officer), Dr. Md. Mokter Hossen Bhuiyan (Senior Scientific Officer) expert in Soil Science and Shamima Sultana expert in Plant breeding, who have carefully examined the data and added new information. This partnership guarantees that the data used is complete and correct, enabling more accurate and productive outcomes. **BARI Cumilla's** collaboration not only confirms the data but also enhances the research's credibility, establishing up opportunities for major breakthroughs in the region's productivity analysis of potato yielding using machine learning. The below google drive link is the evidence of visiting in BARI;

https://drive.google.com/drive/u/0/folders/1UOnXXjz92x9pUTNSZ-zVimvPOqN8BRU9

© Daffodil International University

Chapter 2

PROJECT MANAGEMENT

The focus of Chapter 2 will be on describing how this work has been planned, the difficulties encountered while carrying it out, how the data set was verified, the team members' contributions, and the steps relied on during decision-making.

2.1 Project Plan

The title "Productivity Analysis of Potato (*Solanum tuberosum*) Yielding in Cumilla, Bangladesh Based on Soil Chemical Parameters Using Machine Learning Approaches" has been chosen as a response to a significant demand in Bangladesh's agriculture industry. In Bangladesh, potatoes are an important cash crop and a staple food that significantly improves the country's economy and food security. Potato farming, however, is beset with a number of issues that prevent maximum crop yield, such as degraded soil, unpredictable weather, and inefficient area use. With its fertile lands and favorable climate, Cumilla has the potential to grow into a significant center for potato production. Still, determining which fields are appropriate for cultivation requires a careful and scientific approach due to the variation in soil quality among different parts of Cumilla. Conventional techniques for evaluating soil are frequently expensive, time-consuming, and have a narrow scope. A more effective, data-driven strategy is required to improve agricultural sustainability and production. Modern technologies like machine learning have the power to completely transform agricultural practices by combing through massive datasets and finding hidden patterns and insights. This research aims at predicting field suitability based on a number of indicators, such as pH levels, nutrient content, texture, organic matter, and other pertinent aspects, by utilizing machine learning algorithms. With the use of this prediction ability, farmers will be able to maximize yields and optimize the use of resources by choosing the best locations for potato harvests. Furthermore, this study is in line with the worldwide movement toward precision agriculture, which aims to improve farming's sustainability and efficiency. This study fills a specific need in the region and advances agricultural science overall by concentrating on Cumilla. The **Regional Agricultural Research Station, Bangladesh Agriculture Research Institute (BARI)** is involved as it ensures the validity and applicability of the data and insights produced, which increases the research's relevance and effect.

© Daffodil International University

2.2 Contribution of Team Members.

1. The first author Md. Kamrul Hasan (201-50-012), contributed with the manuscript's development, strategical and methodology design, collection of data, development of applied ML classifiers, analysis, and original draft writing including the literature review part.
2. The second author Anonto Mia (201-50-001), assist with the manuscript's review, editing process, resource providing as well as the literature review.
3. The third author Yeasir Arafat Sohan (201-50-006), assist with providing literature review, and resources for the research.

2.3 Challenges and Decision Making

The study, "Productivity Analysis of Potato (*Solanum tuberosum*) Yielding in Cumilla, Bangladesh Based on Soil Chemical Parameters Using Machine Learning Approaches", faces a number of formidable hurdles. First of all, it was difficult to acquire comprehensive and legitimate data, especially when it comes to soil quality factors, climate, environment, and past pest and disease records. To provide access to verified and current data, collaboration with regional agricultural organizations like the **Bangladesh Agriculture Research Institute (BARI)** is essential. Second, because different datasets have different formats, sizes, and resolutions, combining diverse datasets from different sources such as meteorological data, satellite images, and soil tests done with careful work. To overcome these obstacles, sophisticated data integration strategies have been used, and field surveys was done with **Bangladesh Agriculture Research Institute, Cumilla** to be performed to fill up any data gaps. Furthermore, the quality and consistency of data inputs throughout the research process are critical to developing strong machine learning models that can precisely evaluate the region's field suitability for potato cultivation. The decision-making process for conducting research on this topic entails using a methodical strategy to successfully handle agricultural issues. At the outset, specific study goals were set, with the aim of determining the best fields for cultivating potatoes according to soil properties that are important for the farming environment in the area. The selection of acceptable research methodology and machine learning techniques was guided by a thorough literature review that revealed gaps in knowledge and existing procedures. Working together with regional agricultural organizations, especially **BARI Cumilla**, was essential for obtaining verified data, getting professional advice on how to improve research techniques, and making sure the study was applicable to regional farming practices.

© Daffodil International University

Chapter 3

SYSTEM DESCRIPTION

This chapter named system description, intends to provide an overview of the project's design process. Here, the methods of this work will be described chronologically and also give a detail idea how the research was conducted.

3.1 Methodology Block Diagram of The System

The research on "Productivity Analysis of Potato (*Solanum tuberosum*) Yielding in Cumilla, Bangladesh Based on Soil Chemical Parameter Using Machine Learning Approaches" utilizes a proposed methodology that consists of a number of methodical steps, each of which involves sophisticated technical procedures.

- **Features:** The dataset initially includes two major feature categories: weather status and chemical status. Important soil factors including pH, Nitrogen (N), Phosphorus (P), and Potassium (K) levels are considered in the chemical status. These variables are important predictors of agricultural productivity and soil fertility. At the same time, meteorological information such as humidity, rainfall, and temperature are included in the weather status. These meteorological conditions have a major impact on crop productivity and growth.
- **Data Preprocessing:** Data preparation is the next stage, and it is essential to guaranteeing the quality and appropriateness of the dataset for analysis. This involves a number of smaller processes:
    - **Data Cleaning:** Enhancing data integrity by addressing abnormalities, outliers, and missing values.
    - **Normalization/Standardization:** Adjusting the data's scale to ensure consistency, usually by using methods like Z-score standardization or Min-Max scaling.
    - **Data transformation:** This involves transforming the data into a format that is best for further analysis. To stabilize variance and improve the data's normal distribution, this may involve using power or log transformations.
- **Apply Feature Transformation Methods in Dataset:** The purpose of feature transformation is to improve the prediction ability of the dataset. Principal Component Analysis (PCA) is one technique used to decrease dimensionality while maintaining a substantial amount of variance. Higher-order features are generated by polynomial

© Daffodil International University

feature creation in order to capture non-linear correlations. One possible way to reduce skewness and heterogeneity in the data is to use a logarithmic adjustment.

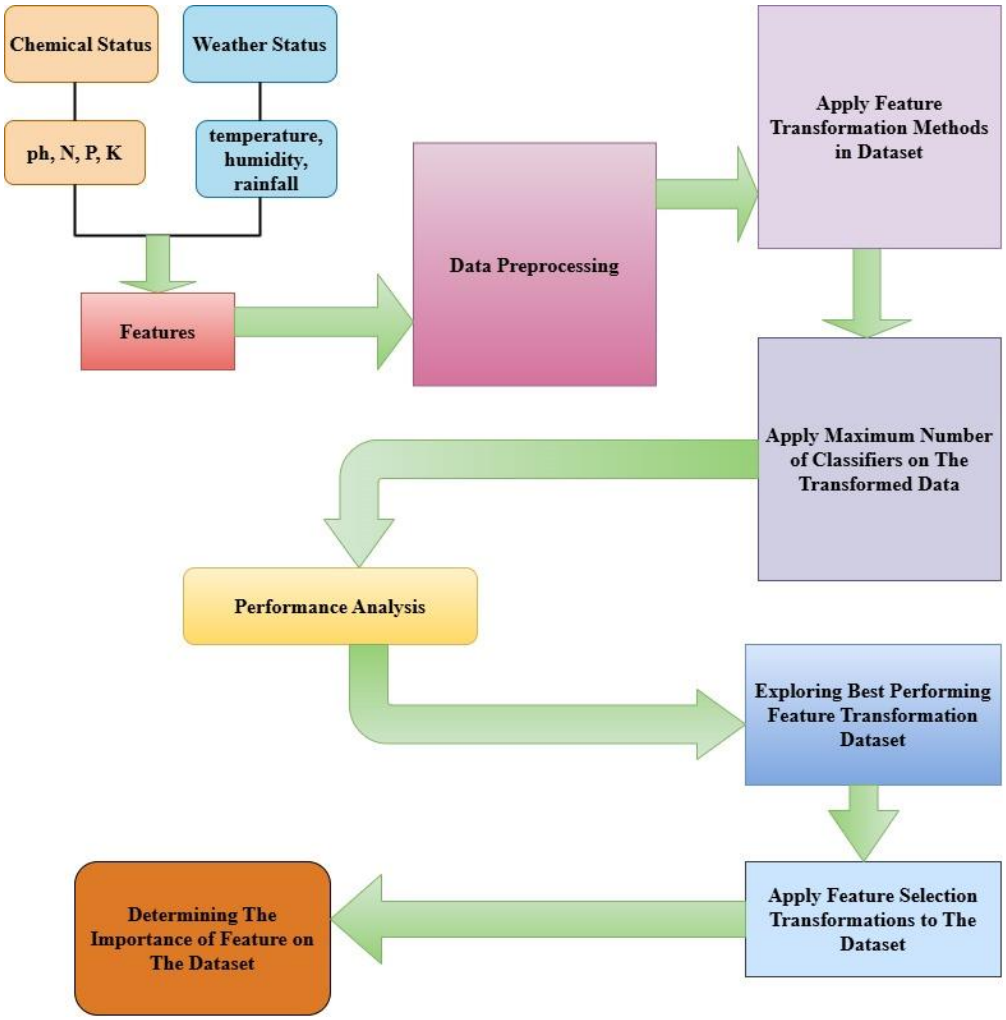

Fig-4: Methodology Diagram for Proposed Title.

- **Apply Maximum Number of Classifiers on The Transformed Data:** The altered dataset is subsequently subjected to an extensive suite of machine learning classifiers. In this step, a number of algorithms are deployed, such as Random Forest, Decision Trees, Support Vector Machines (SVM), Neural Networks, and Ensemble Methods, among others. The aim is to optimize the evaluation of several algorithmic techniques in order to determine which ones perform best for the specified dataset.
- **Performance Analysis:** A performance study is carried out to assess each classifier's effectiveness. This entails figuring out metrics like accuracy, which expresses the percentage of cases that are accurately anticipated. The ratio of true positives to false positives/negatives is also revealed by precision, recall, and F1 score. The effectiveness

© Daffodil International University

of the classifiers in differentiating between classes is also evaluated using the ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) measure.

- **Exploring Best Performing Feature Transformation Dataset:** The performance analysis data are carefully examined to determine which feature transformation methods provided the best results. Comparing performance measures across several transformations is part of this study to find those that consistently improve model robustness and accuracy.
- **Apply Feature Selection Transformations to The Dataset:** In order to improve model performance and reduce complexity, feature selection is used to refine the dataset by keeping just the most pertinent features. The best collection of features is carefully assessed and chosen using approaches including filter techniques (chi-square test, correlation coefficients), wrapper strategies (recursive feature reduction), and embedding techniques (Lasso regression).
- **Determining The Importance of Feature on The Dataset:** The last stage is figuring out how important every feature in the dataset is for the prediction models. Comprehensive methods are used to evaluate and interpret feature importance, including feature importance scores that come from tree-based algorithms (like Random Forest), SHAP (SHapley Additive exPlanations) values that explain individual feature contributions, and permutation importance that measures changes in model performance upon feature permutation.

By carefully following these procedures, the study seeks to identify the most significant features and the top-performing classifiers, thereby offering a solid framework for evaluating potato productivity in Cumilla, Bangladesh, in relation to soil chemical factors and meteorological circumstances.

3.2 Design of Each Block and Select the Best Alternative

Even though the data collection is now ready, pre-processing is still necessary before the feature transformation phase can be carried out. Data preprocessing is an essential phase in machine learning that enhances the performance and accuracy of models. Its goal is to convert raw data into a clear and readable format. It entails a number of important responsibilities, including treating outliers with the proper treatment after identifying them using statistical techniques and visualization, and managing missing values by imputation or deletion. Another essential aspect is data transformation, which includes encoding categorical variables, scaling

© Daffodil International University

(normalization or standards), and generating new features from preexisting ones. Eliminating duplicates and fixing inconsistencies in the data is another important function of data cleaning. All of these actions together ensure excellent data quality, boost feature engineering, and optimize model performance, which eventually results in more accurate and acceptable models. In feature transformation stage, it identifies patterns in the data and ensures that features are scaled uniformly through normalization and scaling, it enhances the performance of the model. Applying methods like Principal Component Analysis (PCA), transformations assist in handling skewed distributions, capturing non-linear correlations, and reducing dimensionality. Important elements include handling multicollinearity, generating new features, and converting category data into numerical values. These modifications lead to stronger insights and more accurate models by reducing overfitting, increasing model efficiency, and improving data interpretability.

Table 2: Mathematical Equations of Feature Transformation

Methods of Feature transformation	Description	Mathematical Equations
Logarithmic	Gaussian density is created through the conversion of high skew density.	$m = b\log_c(1+x)$
Sine	Subjects or instances have been transformed into $[0,2\pi]$ interval	$E_k = \sum_{i=1}^{N-1} f_i \sin\frac{\varphi ik}{N}$
Z-Score	Features are transformed into a range of $[-1,1]$	$Z = \frac{x-\gamma}{\sigma}$

After transforming the dataset, 277 machine learning classifiers have been applied on it, in order to find out the best fit with maximum accuracy. Among all the applied classifiers, 133 have been considered depending on their satisfactory performance. However, 108 out of 133 classifiers have been ignored because of their less accuracy about 63% to 75%, and 25 classifiers have been considered depending on their good accuracy rate. These are: GLMBoost, AdaBoost, Flexible Discriminant Analysis, Bayesian Optimization, Support Vector Machine, Expectation-maximization, Naïve Bias, GANs, Simulated Annealing, Bee Algorithm, Autoencoder, Deep-Q-Network, ANN, Bagging, Decision Tree, Linear Regression, KNN, Ensemble, Multi-Layer Perceptron, Mixture Discriminant Analysis, RNN, CNN, Linear Discriminant Analysis, Quadratic discriminant analysis, Stochastic Gradient Descent.

© Daffodil International University

3.3 System Implementation

All of the relevant classifier equations that are used to make this work successful, are briefly presented in this section of the system implementation. This will make readers to comprehend how a model functions using its formula.

3.3.1 AdaBoost

To combine classifiers which are less strong, AdaBoost has been engaged to generate a new and modern classifier. At the very beginning, equal weights are assigned to all such that:

$$w_i = \frac{1}{N} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

It proceeds iteratively over a specified number of rounds, that can be represent by **T.** To calculating each round t, there assigned a week classifier $h_t(x)$ on the training data. Which are weighted by w_i and the weighted error of that week classifies can be calculated by:

$$e_t = \sum_{i=1}^{N} w_i \,.\, £\,(h_t(x_i) \neq y_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Here **£** is an indicator function which indicates 1 if the condition is true and 0 otherwise. Computing the weight α_t of the week classifier $h_t(x)$ the formula needed is:

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1-e_t}{e_t} \right) \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

For the new round the weight needs to be updated by the following formula:

$$w_i \leftarrow w_i \cdot \exp\left(\alpha_t \cdot £\,(h_t(x_i) \neq y_i)\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

to normalize the weight $w_i \leftarrow \frac{w_i}{\sum_{j=1}^{N} w_j}$ can be use. Finally comes the strong classifier:

$$S(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t \cdot h_t(x)\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

3.3.2 GLMBoost

To generalizing linear models, GLMBoost can be apply which is a form of gradient boosting. The target of GLMBoost is to minimize the loss function $L(b, f(a))$ over a training dataset $\{(a_i, b_i)\}_{i=1}^{N}$,here a_i are the input features and b_i is the target values. With $f_0(a) = \arg\min \sum_{i=1}^{N} L(y_i, c)$ that equation the model will initialize, where **c** is a constant value. For each iteration **r** = 1 to **R**. Calculating the negative gradient of the loss function with $F_i^{(r)} = -\frac{\partial L(b_i, f(a_i))}{\partial f(a_i)}\Big|_{f(a_i = f_{r-1}(a_i)}$, this gradient serves as the pseudo-residuals. Now fitting a generalized linear model $h_r(a_i; \beta)$ into the pseudo-residuals.

$$\beta_r = \arg\min \sum_{i=1}^{N} (F_i^{(r)} - h(a_i; \beta)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

© Daffodil International University

Adding learning rate **"φ"** and updating the model with new base learner scale:

$$\mathbf{f_r(a) = f_{r-1}(a) + \varphi h_r(a; \beta_r)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

After **R** iterations, the final boosted model will be:

$$\mathbf{f_R = f_0(a) + \sum_{r=1}^{R} \varphi h_r(a; \beta_r)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

The choice of the loss function **L (b, f(a))** depends on the type of problem, if it is for regression then:

$$\mathbf{L\,(b, f(a)) = (b - f(a))^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots (9)$$

And if for classification then:

$$\mathbf{L\,(b, f(a)) = \log\,(1 + \exp(-bf(x)); \text{ where } b \in \{-1,1\}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (10)$$

3.3.3 Flexible Discriminant Analysis

The basis of FDA's mathematics is the approximation of discriminant functions with the fitting of a flexible function, usually using regression analysis. Let **P** be the predictor variable matrix, where applying a transformation ∅ to **∅(P)**. Here **∅(P)** indicates potentially non-linear transformation of the predictors. Considering $\propto_c$ is the coefficient from the fitted regression model for class ***C***. Equation for the discriminant of class ***C*** is:

$$\mathbf{\delta_c(P) = \alpha_{0c} + \alpha_c^T\, \emptyset(P)}, \text{ here } \mathbf{\alpha_{0c}} \text{ is the intercept for class C.} \quad \ldots\ldots\ldots\ldots\ldots (11)$$

class C using the classification technique and providing a new observation P:

$$\mathbf{\hat{x} = \arg\max \delta_c(P)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (12)$$

3.3.4 Bagging or Bootstrap Model

In order to avoid overfitting and decrease variance, bagging is an effective ensemble method It enhances machine learning algorithms' accuracy and stability. Bagging is creating many versions of a predictor in order to produce an aggregated predictor. Considering **D** as the dataset of n size where bootstrap sample is $\mathbf{D_b^*}$**.** Let $\mathbf{m_b}$ indicates the model trained on the **b-th** bootstrap sample $\mathbf{D_b^*}$. The result in **B** trained models $\mathbf{m_1, m_2, ..., m_B}$**.** The final prediction for regression $\hat{\mathbf{r}}$ with a new observation **h**.

$$\hat{\mathbf{r}} = \frac{\mathbf{1}}{\mathbf{B}} \sum_{\mathbf{b=1}}^{\mathbf{B}} \mathbf{m_b(h)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (13)$$

And for classification; $\hat{\mathbf{r}}$ **= mode {** $\mathbf{m_1(h), m_2(h), ..., m_B(h)}$**},** the variance can be calculated by:

© Daffodil International University

$$V(\hat{m}_{baag}(h)) = \frac{1}{B^2} \sum_{b=1}^{B} V(m_b(h)) \quad \text{..... (14)}$$

As bagging model decrease variance without highly increasing bias. So, for a single model ***m,*** the prediction error can be denoted like:

$$\textbf{Error} = \textbf{bias}^2 + \textbf{V} + \textbf{nonreducible error} \quad \text{..... (15)}$$

3.3.5 Bayesian Optimization Algorithm

At first considering an initial set of points $\{e_i\}_{i=1}^{n}$ and then execute the objective function $\{d_i = f(e_i)\}_{i=1}^{n}$. To observe the data fitting GP (gaussian process) is necessary $\{(e_i, d_i)\}$. For finding the next point e_{n+1} needs to approach the maximizing acquisition function.

$$e_{n+1} = \arg\max \gamma(e) \quad \text{..... (16)}$$

In order to check the objective function at e_{n+1}, getting $d_{n+1} = f(e_{n+1})$. Now to meet the expected output the GP model needs to update with a new value (e_{n+1}, d_{n+1}), and the process will be repeated until the criteria matched.

3.3.6 Expectation-Maximization (EM)

When a model depends on unobserved latent variables, an iterative algorithm is use to evaluate maximum likelihood estimates of parameters in statistical model know as Expectation-Maximization (EM). Assuming the observed data $A = \{ a_1, a_2, ..., a_n\}$, unobserved data $B = (b_1, b_2, ... , b_n\}$, parameter set φ and the complete data $C = (A, B)$, and target is to maximize the likelihood function $L(\varphi|A) = P(A|\varphi)$. Executing E-step (Expectation step) with computing the expected value of the log-likelihood function; $M(\varphi|\varphi^{(t)}) = E_{B|A,\varphi^t}[\log P(A, B \mid \varphi)]$. For (M-step) maximization step, the updated parameter estimations using the expected log-likelihood found in the E-step in relation to the parameters.

$$\varphi^{(t+1)} = \arg\max M(\varphi|\varphi^{(t)}) \quad \text{..... (17)}$$

3.3.7 Generative Adversarial Networks (GANs)

The model is a combination of two neural networks, the generator and discriminator. Both the model trained simultaneously through an adversarial process. Using prior distribution $P_Z(Z)$ the generator takes a random noise vector z, and to produce a synthetic data sample $G(z)$ it puts that into the data space. A binary classifier known as the discriminator receives an input (a sample of real data or a synthetic sample from the generator) and returns a probability $D(x)$ denoting the chance that the input data is real (i.e., drawn from the training data) as instead of being created. The objective of the GAN training process is to identify a Nash equilibrium

© Daffodil International University

between the discriminator and the generator. The generator produces realistic data samples which are unrecognizable from actual data generated by the discriminator. That can be formulated by:

$$\mathbf{min_G\ max_D\ V(D, G)} \quad \text{..... (18)}$$

Where, $\mathbf{V(D, G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))]}$

Continuing each iteration, the discriminator D have to be updated by maximizing the below rule.

$$\mathbf{E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))]} \quad \text{..... (19)}$$

After updating the discriminator, generator (G) also has to be updated by minimizing the following:

$$\mathbf{E_{z\sim p_z(z)}[\log(1-D(G(z)))]} \quad \text{..... (20)}$$

But using this can create problem due to vanishing gradients. So, in order to avoid that the below objective can be maximize:

$$\mathbf{E_{z\sim p_z(z)}[\log(D(G(z)))]} \quad \text{..... (21)}$$

3.3.8 Multi-Layer Perceptron (MLP)

Here three layers are usually included in the architecture: an input layer, an output layer, and one or more hidden layers. Considering input vector is $\mathbf{K = [k_1, k_2, ..., k_n]}$.with the help of activation function hidden layers calculate the sum of inputted weight. And the output can be calculated by:

$$\mathbf{h_j^{(y)} = \varphi\ (\sum_{i=1}^{n} w_{ij}^{(y)} a_i^{(y-1)} + b_j^{(y)}} \quad \text{..... (22)}$$

Here,

$\mathbf{h_j^y}$ is the output of the **j**-th neuron in layer **y**.

φ is the activation function (e.g., sigmoid, tanh, softmax)

$\mathbf{b_j^{(y)}}$ is the bias term.

$\mathbf{w_{ij}^{(y)}}$is the weight connecting **i**-th neuron of the previous layer **(y-1)** to the **j**-th neuron of the current layer **y**.

© Daffodil International University

$\mathbf{a_i^{(y-1)}}$ is the output (activation) of the **i**-th neuron in the **(y-1)** th layer.

The **k**-th neuron's output in the output layer can be expressed as:

$$\mathbf{F_k = \gamma\left(\sum_{j=1}^{m} w_{jk}^{Y} a_j^{(Y-1)} + b_k^{(Y)}\right)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (23)$$

γ is the activation function (softmax, linear). To Compute the output through forward propagation, at first the input vector **K** have to pass through input layer. Then Apply the activation function after calculating the weighted total of the inputs; $\mathbf{\propto^{(y)} = \varphi\left(w^{(y)} \propto^{(y-1)} + b^{(y)}\right)}$, and the output activation function; $\mathbf{out = \gamma\left(w^{(Y)} \alpha^{(Y-1)} + b^{(Y)}\right)}$. In order to minimize a loss function, weights and biases are adjusted during the MLP's training process. The usual technique for doing this is termed backpropagation. The variance between the true output M and the projected output N is measured by L (N, M). Calculating loss function with respect to each weight and bias for the case of output layer:

$$\mathbf{\delta_k^Y = \frac{\partial L}{\partial F_k} \gamma'\left(z_k^{(y)}\right)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (24)$$

and for the case of hidden layers:

$$\mathbf{\delta_j^{(y)} = \left(\sum_k \delta_k^{(y+1)} w_{jk}^{(y+1)}\right) \varphi'\left(z_j^{(y)}\right)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (25)$$

Lastly, adjusting the weights and biases using the gradients and a learning rate $\boldsymbol{\eta}$:

$$\mathbf{w_{ij}^{(y)} \leftarrow w_{ij}^{(y)} - \eta \frac{\partial L}{\partial b_j^{(y)}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (26)$$

$$\mathbf{b_j^{(y)} \leftarrow b_j^{(y)} - \eta \frac{\partial L}{\partial b_j^{(y)}}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (27)$$

3.3.9 Bee Algorithm

An approach of optimization termed as Bee Algorithm was developed to address combinatorial optimization issues. To initialize the algorithm, considering **P** as the population of solutions. And **f(P)** as the fitness function.

$$\mathbf{P_i = P_{min} + (P_{max} - P_{min}) \cdot rand( )} \quad \ldots\ldots\ldots\ldots\ldots\ldots (28)$$

Here, **rand ()** generates random number between 0 and 1. For searching neighbor:

$$\mathbf{P_{ij}'' = P_{ij} + \varphi_{ij} \cdot (P_{ij} - P_{kj})} \quad \ldots\ldots\ldots\ldots\ldots\ldots (29)$$

© Daffodil International University

Here; $\mathbf{P''_{ij}}$ is the current solution

$\mathbf{P_{kj}}$ is a randomly selected solution

$\boldsymbol{\varphi}_{\mathbf{ij}}$ is random number between (-1 to 1)

Calculating the probability using:

$$\mathbf{Q_i} = \frac{\mathbf{f(P_i)}}{\sum_{\mathbf{j=1}}^{\mathbf{N}} \mathbf{f(P_j)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (30)$$

$\mathbf{Q_i}$ is the probability **i**-th solution based on its fitness. When a solution does not get better after a certain number of iterations (limit), it is dropped and a new one will be created.

3.3.10 Autoencoder

Encoder, latent space (sometimes called bottleneck), and decoder are the three primary parts that make up the fundamental framework of an autoencoder. For the purpose of feature learning and dimensionality reduction, it is used to develop effective data representations. To compress the input data, **I** into a lower dimension at the encoder end.

$$\mathbf{E} = \boldsymbol{\omega}(\mathbf{w_e I} + \mathbf{b_e}) \quad \ldots\ldots\ldots\ldots\ldots\ldots (31)$$

Here; $\boldsymbol{\omega}$ is the activation function.

$\mathbf{b_e}$ is the bias vector for encoder.

The compressed representation of the input data is called the latent space, sometimes known as the bottleneck or latent space **E**. It records the key features required to recreate the input. The encoded representation is used by the decoder for reconstructing the input data, and mathematically represented by:

$$\widehat{\mathbf{D}} = \boldsymbol{\gamma}(\mathbf{w_d E} + \mathbf{b_d}) \quad \ldots\ldots\ldots\ldots\ldots\ldots (32)$$

Here; $\boldsymbol{\gamma}$ is the activation function.

$\mathbf{b_d}$ is the bias vector for decoder.

Minimizing the difference between the input **I** and the reconstructed output $\widehat{\mathbf{D}}$ is the objective of autoencoder training. A loss function is used to measure this difference and the mean squared error is a common choice.

$$\boldsymbol{\vartheta}(\mathbf{I}, \widehat{\mathbf{D}}) = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{I_i} - \widehat{\mathbf{D}}_\mathbf{i})^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (33)$$

© Daffodil International University

With the help of gradient descent or its variants, the training process includes optimizing the weights $\mathbf{w_e}$ and $\mathbf{w_d}$ also the biases $\mathbf{b_e}$ and $\mathbf{b_d}$, in order to minimize the loss function.

3.3.11 Ensemble Model

To improve performance overall, an ensemble model aggregates predictions from several distinct models. Depending on the kind of ensemble method being utilized, there are various ways to write the general equation for an ensemble model. Three typical kinds of ensemble approaches are as follows: bagging, boosting and stacking. Usually, bagging makes use of the majority or average vote among each model's predictions. It is the majority vote in a classification problem and the average of predictions in a regression work. For regression:

$$\hat{\mathbf{F}} = \frac{1}{\mathbf{M}} \sum_{\mathbf{m=1}}^{\mathbf{M}} \hat{\mathbf{F}}_{\mathbf{m}} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (34)$$

Here, $\hat{\mathbf{F}}$ is the final prediction.

M is the number of individual models.

$\hat{\mathbf{F}}_{\mathbf{m}}$ is the prediction of **m**-th model.

For classification:

$$\hat{\mathbf{F}} = \mathbf{mode}(\hat{\mathbf{F}}_1, \hat{\mathbf{F}}_2, \ldots\ldots, \hat{\mathbf{F}}_{\mathbf{M}}) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (35)$$

Boosting usually combines models one after the other in an ordered manner, with each model trying to fix the mistakes of the one before it. A weighted total of each individual prediction makes up the final forecast. $\hat{\mathbf{F}} = \sum_{m=1}^{M} \boldsymbol{\beta}_m \hat{\mathbf{F}}_m$; $\boldsymbol{\beta}$ is the assigned weight to the **m**-th model, that often bias on its performance. Training a meta-model to aggregate base model predictions is known as stacking. The meta-model receives its predictions from the underlying model.

$$\hat{\mathbf{F}}_{\mathbf{i}} = \mathbf{f}(\hat{\mathbf{F}}_{\mathbf{1,i}}, \hat{\mathbf{F}}_{\mathbf{2,i}}, \ldots\ldots, \hat{\mathbf{F}}_{\mathbf{M,i}} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (36)$$

Here, $\hat{\mathbf{F}}_{\mathbf{i}}$ is the final prediction for i-th instance.

f is the function learned by the meta-model.

M is the number of base models.

$\hat{\mathbf{F}}_{\mathbf{M,i}}$ is the prediction of the m-th base model for the i-th instance.

3.3.12 Simulated Annealing

A probabilistic method for estimating a function's global optimum is called simulated annealing. Large optimization problems with complex search spaces and ineffective classical

© Daffodil International University

approaches benefit significantly by it. The mathematical process can be start by assuming; initial solution $\mathbf{s_0}$, initial temperature $\mathbf{T_0}$ and cooling schedule $\mathbf{T_i}$ where **i** is the iteration steps. For each iteration **i**, generating new solution $\mathbf{s_{new}}$ in the neighborhood of $\mathbf{s_i}$ using $\mathbf{N_{(s_i)}}$. Now evaluating $\mathbf{f(s_{new})}$ and $\mathbf{f(s_i)}$. In order to calculate the acceptance probability; if $\mathbf{f(s_{new}) < f(s_i)}$, accept $\mathbf{s_{new}}$ otherwise accept $\mathbf{s_{new}}$ with the given probability equation.

$$\mathbf{P(accept)= exp\left(\frac{f(s_i)-f(s_{new})}{T_i}\right)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (37)$$

If $\mathbf{s_{new}}$ is accepted then set $\mathbf{s_{i+1}= s_{new}}$ otherwise keep $\mathbf{s_i}$, and update the temperature $\mathbf{T_{i+1}}$ using the cooling schedule. And the process will be completed when the $\mathbf{T_i}$ is sufficiently low or after a fixed number of iterations. Here are some common cooling schedules.

- Geometric cooling: $\mathbf{T_i= T_0.\beta^i}$; where $0<\beta<1$
- Linear cooling: $\mathbf{T_i = T_0- c.i}$; here **c** is a constant.
- Logarithmic cooling: $\mathbf{T_i=\frac{T_0}{log(1+i)}}$

3.3.13 Deep-Q Network (DQN)

Deep Q-Networks (DQNs) are a type of reinforcement learning algorithm that is useful for a variety of challenging applications because it integrates Q-learning and deep neural networks to handle high-dimensional state spaces. A target network is used for stable training, experience replay, neural network function approximation, and the Q-value update rule is one of the key components. The major part of Q-learning is to updating rule for Q-values by:

$$\mathbf{Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \gamma(r_{t+1}+\beta max_{a"}Q(s_{t+1},a")-Q(s_t,a_t)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (38)$$

Here, $\mathbf{Q(s_t,a_t)}$ recent **Q**-values for state $\mathbf{s_t}$ and action $\mathbf{a_t}$.

$\boldsymbol{\gamma}$ is the learning rate.

$\mathbf{r_{t+1}}$ reward after taking action $\mathbf{a_t}$.

$\boldsymbol{\beta}$ is the decision factor.

$\mathbf{s_{t+1}}$ is the next state after taking action $\mathbf{a_t}$.

$\mathbf{max_{a"}Q(s_{t+1},a")}$ is the max Q-value for the next state $\mathbf{s_{t+1}}$across all possible action **a"**.

The objective is to reduce the difference between the target and predicted Q-values, and the target Q-values for a given state-action pair can be written like; $\mathbf{T_t= r_{t+1}+ \beta max_{a"}Q(s_{t+1},a";\emptyset^-)}$ where, $\boldsymbol{\emptyset^-}$ are the parameters of a target network

© Daffodil International University

And the loss function can be calculated by:

$$\mathbf{Loss(\emptyset)= \mathbb{E}\ [\ (T_t\text{-}Q(s_t,a_t;\emptyset))^2\ ]} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (39)$$

3.3.14 Mixture Discriminant Analysis (MDA)

A statistical method called mixture discriminant analysis (MDA) uses concepts from mixture models and discriminant analysis to categorize observations into pre-established groups. Modeling each class's probability density function as a combination of multiple component distributions usually Gaussian distributions is the primary objective of multinomial decision analysis. A mixture model uses a weighted sum of many component densities to represent the probability density function (pdf) of a given data point, **D**. The pdf can be written as follows for a class **C**.

$$\mathbf{P(D|C)= \sum_{J=1}^{M_k} \pi_{Jk} P_{Jk}(D)} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (40)$$

Here, $\mathbf{M_k}$ number of components in mixture for class **C.**

$\boldsymbol{\pi_{Jk}}$ mixing proportion for **J**-th component in class **C** (with $\sum_{J=1}^{M_k} \pi_{Jk} = 1$**).**

$\mathbf{P_{Jk}(D)}$ Probability density function of **J**-th component in class **C.**

For a gaussian component the probability density function is:

$$\mathbf{P_{Jk}= \frac{1}{(2\pi)^{d/2} |\ \sum_{Jk}\ |^{1/2}} \exp\left(-\frac{1}{2}(D\text{-}£_{Jk})^T \sum_{Jk}^{-1} (D\text{-}£_{Jk})\right)} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (41)$$

Where, $\mathbf{£_{Jk}}$ is the mean vector for **J**-th gaussian component in class **C**.

$\sum_{Jk}$ is the covariance matrix

d is the dimensionality of the data.

To classify a new observation **n**, bias theorem is applied and compute the posterior probability of each class:

$$\mathbf{P\ (C\ |\ n) = \frac{P(C)P(n|C)}{P(n)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (42)$$

Here, **P**(**C**) is the prior probability of class **C**,

P (C | n) is class conditional density.

The discriminant function for class **C;**

$$\mathbf{g_C(D)= \log P(C)+\log P(D\ |\ C)} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (43)$$

© Daffodil International University

3.3.15 Quadratic Discriminant Analysis (QDA)

One method of categorization that comes from statistical decision theory is quadratic discriminant analysis (QDA). With the primary distinction being that QDA permits each class to have its own covariance matrix, resulting in a quadratic decision boundary, it is an extension of Linear Discriminant Analysis (LDA). For a data point **p**, the discriminant function for class **c** is:

$$\sigma_c(p) = -\frac{1}{2}\log | \sum_c | - \frac{1}{2}(p-\beta_c)^T \sum_c^{-1} (p-\beta_c) + \log\pi_c \quad \ldots\ldots\ldots\ldots\ldots\ldots (44)$$

Here, $| \sum_c |$ is the indication of the covariance matrix $\sum_c$

$\sum_c^{-1}$ is the inverse covariance matrix.

π_c is the logarithm of the prior probability of class **c**.

To maximize the discriminant function, the classification rule can be:

$$c = \text{argmax}_c \beta_c(p) \quad \ldots\ldots\ldots\ldots\ldots\ldots (45)$$

The discriminant functions $\beta_i(p)$ and $\beta_j(p)$ produce quadratic terms, which determine the decision border between any two classes, i and j: $\beta_i(p) = \beta_j(p)$.

Using logarithm for posterior probability:

$$\log d\ (e = c \mid p) \propto \log\pi_c - \frac{1}{2}\log|\sum_c| - \frac{1}{2}(p-\beta_c)^T \sum_c^{-1} (p-\beta_c) \quad \ldots\ldots\ldots\ldots\ldots\ldots (46)$$

3.3.16 Linear Discriminant Analysis (LDA)

By identifying a linear combination of predictor variables (features), LDA tries to optimize the separability between the known categories. Increasing the ratio of the between-class variance to the within-class variation allows for this. For each class **v** measuring the mean vector; $\vartheta_v = \frac{1}{N_v}\sum_{x \in c_v} x$, where N_v is the number of samples in class **v** and c_v indicates the set of all samples in class **v.** To compute overall mean vector $\vartheta = \frac{1}{N}(\sum_{v=1}^{c} N_v \vartheta_v)$ can be use, where **c** is the number of classes. In order to calculate the within-class scatter matrix:

$$S_w = \sum_{v=1}^{c} S_v \quad \ldots\ldots\ldots\ldots\ldots\ldots (47)$$

Here, $S_v = \sum_x c_v(x-\vartheta_v)(x-\vartheta_v)^T$; S_v is the scatter matrix of class **v.**

© Daffodil International University

And between-class scatter matrix:

$$\mathbf{S_B}=\sum_{v=1}^{c}\mathbf{N_v}(\vartheta_v-\vartheta)(\vartheta_v-\vartheta)^{\mathbf{T}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (48)$$

To generalized the eigenvalue problem the following can implement: $\mathbf{S_w^{-1}S_Bw}=\epsilon\mathbf{w}$; ϵ is the eigenvalue and **w** is the eigenvector. The columns of the transformation matrix **W** are made up of the eigenvectors that correspond to the biggest eigenvalues. Next, the data is projected onto this newly created subspace. **Z = GW; Z** represents the transformed data, and **G** is the original data matrix.

3.3.17 Stochastic Gradient Descent (SGD)

SGD is an optimization algorithm commonly used in machine learning. Typically, the goal of training a machine learning model is to minimize a loss function denoted by **L (δ)**, where **δ** stands for the model's parameters. The SGD update follows $\boldsymbol{\delta_{t+1}=\delta_t-\mu\nabla_\delta L_i(\delta_t)}$**;** where $\mathbf{L_i(\delta_t)}$ is the loss function that calculated on the **i**-th data point. Now assuming a data point $\mathbf{(d_i,d_j)}$**,** gradient of the loss function **L (δ)** with respect to the parameters δ can computed as follows.

$$\nabla_\delta \mathbf{L}\,(\delta) = \frac{\partial \mathbf{L}(\delta;\mathbf{d_i},\mathbf{d_j})}{\partial \delta} \quad \ldots\ldots\ldots\ldots\ldots\ldots (49)$$

3.3.18 Recurrent Neural Network (RNN)

In order to handling sequential data, RNN is widely common. An RNN processes sequence of inputs $\{\mathbf{r^1},\mathbf{r^2}, \ldots\ldots, \mathbf{r^T}\}$ one at a time. For each time step **t**, it has input $\mathbf{r^t}$, a hidden state $\mathbf{h^t}$ and an output $\mathbf{F^t}$. Depending on the current inputs $\mathbf{r^t}$ and previous hidden state $\mathbf{h^{(t-1)}}$ the hidden state $\mathbf{h^t}$ is updated.

$$\mathbf{h^t}=\varphi(\mathbf{w_h r^t}+\mathbf{M_h h^{(t-1)}}+\mathbf{b_h} \quad \ldots\ldots\ldots\ldots\ldots\ldots (50)$$

Where; $\mathbf{w_h}$ input weight matrix.

$\mathbf{M_h}$ hidden state weight matrix.

$\mathbf{b_h}$ bias vector.

φ is activation function.

The output at each time step **t** can be achieved by; $\mathbf{F^t}=\gamma(\mathbf{w_F h^t}+\mathbf{b_F})$. The loss function (cross-entropy loss for classification) $\mathbf{Loss}=\sum_{t=1}^{T}\mathbf{Loss^t}$; here, $\mathbf{Loss^t}$ is the loss at **t** time step.

© Daffodil International University

3.3.19 Convolution Neural Networks (CNN)

In a typical neural network layer, the matrix multiplication is replaced by the convolution operation. In order to create a feature map, the input data is dragged over by a filter, also known as a kernel. Assuming for a 2D input **I** and a 2D filter **F**, the convolution **C** is.

$$\mathbf{C(g, h)= (I * F)(g,h)= \sum_m \sum_n I(m,n) .F} \quad \text{..... (51)}$$

With cross correlation operation:

$$\mathbf{C(g, h)= \sum_m \sum_n I(g+m,h+n) .F(m,n)} \quad \text{..... (52)}$$

In terms of introducing non-linearity, an activation function is applied element-wise after convolution. To reduce the spatial dimension of input, max pooling and average pooling can be use. Max pooling for 2x2 window.

$$\mathbf{pooling(g,h)= max\{(I(2_g,2_h),I(2_g,2_h+1),I(2_g+1,2_h),I(2_g+1,2_h+1)} \quad \text{..... (53)}$$

Fully connected layers are used for high-level reasoning after multiple convolutional and pooling layers. Like in a conventional neural network, the output of the last pooling layer is vectorized and fed into a fully connected layer.

3.3.20 Naïve Bayes

Bayes theorem is the basic foundation of naïve bayes and that can be represent like below:

$$\mathbf{N\ (F|X) = \frac{N(X|F).N(F)}{N(X)}} \quad \text{..... (54)}$$

Here, **N (F|X)** = posterior probability of class F given the feature **X**.

N(X|F) = likelihood of feature **X** given class **F**.

P(F) = prior probability of class **F**.

P(X) = marginal likelihood of feature **X**.

Simplifying the computation of **N(X|F):**

$$\mathbf{N(X|F) = N(\ a_1,a_2, ..., a_n|F) \approx \prod_{i=1}^{n} N\ (a_i|\ F)} \quad \text{..... (55)}$$

Here the vision is to predict the class F that will increase the posterior probability **N (F|X)**

$$\mathbf{\widehat{F} = arg\ max_F\ \ N\ (F|X)} \quad \text{..... (56)}$$

Now implementing bayes theorem and naïve bayes assumption

© Daffodil International University

$$\widehat{F} = \arg\max_F \frac{N(X|F).P(F)}{P(X)} \quad \text{..... (57)}$$

In maximization P(X) can be ignored for being same in all classes

$$\widehat{F} = \arg\max_F \; N(X|F).\, P(F) \quad \text{..... (58)}$$

After all, substituting the naïve assumption into this equation

$$\widehat{F} = \arg\max_F P(X).\, \prod_{i=1}^{n} P(a_i|\, F) \quad \text{..... (59)}$$

3.3.21 K-Nearest Neighbors (KNN)

The model KNN's base mathematics involves calculating the distance between data points. This algorithm works for both classification and regression. The distance can be calculated by several ways. Considering two-point $\mathbf{U = (u_1, u_2, \ldots\ldots, u_n)}$ and $\mathbf{V = (v_1, v_2, \ldots\ldots\ldots, v_n)}$ in **n**-dimensional space

$$\mathbf{d\,(U, V)} = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2} \quad \text{..... (60)}$$

$$\text{or, } \mathbf{d\,(U, V)} = \sum_{i=1}^{n} |u_i - v_i| \quad \text{..... (61)}$$

$$\text{or, } \mathbf{d\,(U, V)} = \left(\sum_{i=1}^{n} |\, u_i - v_i|^p \right)^{1/p} \quad \text{..... (62)}$$

Here, **P** indicates the type of distance. If **P= 2** it's Euclidean distance, and if **P= 1** it's Manhattan distance. After finding the distance between two points, next part is to identify the **k** nearest neighbors. For classification; let $\mathbf{V_i}$ be the class label of the **i**-th neighbor and the predicted class $\widehat{\mathbf{V}}$ is:

$$\widehat{V} = \mathbf{mode}\,(\{ V_1, V_2, \ldots\ldots, V_k)\} \quad \text{..... (63)}$$

For regression:

$$\widehat{V} = \frac{1}{k} \sum_{i=1}^{k} V_i \quad \text{..... (64)}$$

3.3.22 SVM (Support Vector Machine)

SVM deals with optimal hyperplane determination. The goal is to create decision boundary or best line to convert n-dimensional space into classes, so that the classification process can be easier. The distance of a point **X** from a hyperplane defined by the equation $\mathbf{A.x + c = 0}$ can be calculated by:

$$D_H = \frac{|A.x+C|}{||A||} \quad \text{..... (65)}$$

For positive and negative side of the hyperplane the equation will be:

© Daffodil International University

$$\vec{A}.\vec{x} + c = 1 \quad \text{..... (66)}$$

$$\vec{A}.\vec{x} + c = -1 \quad \text{..... (67)}$$

The margin will define the location of a hyperplane which use to divide the classes, and can be calculated by

$$\textbf{Margin(M)} = \frac{2}{||A||} \quad \text{..... (68)}$$

For non-linear SVM a **Z** axis can be assume $Z = X^2 + Y^2$, where **Z** is a mapping function that will use to convert the non-linear data into linear data. when converting in 2D space if **Z=1,** the best hyperplane will be generated and the classes can be defined easily.

3.3.23 Artificial Neural Network (ANN)

The independent variables, together with each neuron's associated weight and bias term, are combined linearly to form the ANN. The ANN equation can be represented as follows.

$$\mathbf{Z = w_1x_1 + w_2x_2 + \ldots\ldots\ldots + w_nx_n + b} \quad \text{..... (69)}$$

Here, **X**= input values.

W= weight.

b = bias value.

ANN basically works on three (input, hidden, output) layers. The **X** is the values in input layers which will combinedly input of hidden layers, some activation functions are works here to process the inputs. There are many types of activation functions, some are in below.

RELU activation function, **A(X) = max (0, X)**, it gives output **x** if **x** is positive and 0 otherwise. **Sigmoid** activation function, $\mathbf{y} = \frac{1}{1+e^{-z}}$, another is **tanh** activation function $\frac{2}{(1+e^{-2x})-1}$ or, it can be **tanh(x) = 2*sigmoid(2x)-1**. After all that process it goes through the output layer and generate the output value. So, the overall equation for output layer would be:

$$\mathbf{f\left(\sum w_{ji} * a_i\right) + b_o} \quad \text{..... (70)}$$

Where, **f** is the activation function.

$\mathbf{w_{ji}}$ is the weight connecting '**i**'th neuron in the previous layer of the output.

$\mathbf{a_i}$ is the activation of '**i**'th neuron in the previous layer.

$\mathbf{b_o}$ is the bias of output neuron.

© Daffodil International University

3.3.24 Decision Tree

A decision tree is a structure that simulates a flowchart, with each leaf node indicating the result, each branch representing a decision rule, and each internal node representing a feature (or attribute). Classification rules are symbolized by routes that run from root to leaf. The equation of decision tree can be express like below:

$$\textbf{Entropy} = -\sum_{i=1}^{n} P_i * \log(p_i) \quad \text{..... (71)}$$

$$\textbf{Gini index} = 1 - \sum_{k=0}^{n} p_i^2 \quad \text{..... (72)}$$

Decision Tree Classifier Equation:

Let **X** be the feature vector, where **X** = [ph, n, p, k, temperature, humidity, rainfall].

Let **Y** be the target label, where **Y** = {potato, rice, jute, watermelon, cabbage, cauliflower, corn, tomato}.

The decision tree classifier can be represented as:

$$\mathbf{y = f(x) = argmax(P(y|x))} \quad \text{..... (73)}$$

Where, **P(y|x)** is the conditional probability of the target label given the feature vector.

The decision tree classifier learns a set of rules to predict the target label based on the feature values. The rules are represented as a tree-like model, where each internal node represents a feature, each branch represents a decision based on the feature value, and each leaf node represents a class label. Mathematical Representation of Decision Tree Rules:

Let **T** be the decision tree, and **t** be a node in the tree.

Let $\mathbf{X_i}$ be the i-th feature, and $\mathbf{X}_i$ be the value of the i-th feature.

Let $\mathbf{Y_i}$ be the class label associated with node **t**.

The decision tree rules can be represented as:

$$\textbf{if } X_i <= x_i \textbf{ then t = left\_child(t) else t = right\_child(t)}$$

$$\mathbf{y = y_i \text{ if t is a leaf node}} \quad \text{..... (74)}$$

© Daffodil International University

3.3.25 Linear Regression

A dependent variable and one or more independent variables can be modeled using the statistical technique of linear regression. The objective is to identify the line that best fits the data and, in higher dimensions, represents the change in the dependent variable as independent variables change. The equation for Simple linear regression can be simplify by.

$$\mathbf{y = \beta_0 + \beta_1 x} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (75)$$

Here, **y** = dependent variable.

x = independent variable.

$\boldsymbol{\beta_0}$ = is the intercept.

$\boldsymbol{\beta_1}$ = slope.

For multiple linear regression:

$$\mathbf{y = \beta_0 + \beta_1 x + \beta_2 x + \beta_n x} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (76)$$

The objective to use linear regression is to find the best fit line. And the hypothesis function can be defined by:

$$\mathbf{\hat{Y} = \theta_1 + \theta_2 x} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (77)$$

Or,

$$\mathbf{\hat{y}_i = \theta_1 + \theta_2 x_i} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (78)$$

The model will find the best regression fit line by finding the best value of $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$. However, to find the best fit line $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ might have to be updated by executing the below equation:

$$\mathbf{Minimize\ } \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \hat{y}_i)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (79)$$

And the cost function or the error can be defined by:

$$\mathbf{Cost\ function(j) = } \frac{1}{n} \sum_{n}^{i} (\hat{y}_i - \hat{y}_i)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (80)$$

All the above classifiers are applied into the dataset to find the best result. The results are evaluated by different evaluation matrices and the best performance are achieved by comparing all the classifiers accuracy. All the information about classifiers accuracy and comparison are properly describe in the next chapter.

© Daffodil International University

Chapter 4

SYSTEM TESTING AND ANALYSIS

This chapter will present an overview of the final results, which includes comparison table of all the classifiers and graphical depiction of the accuracy, that are acquired after the classifiers have been implemented. This will facilitate readers' understanding of the work's final outcome.

4.1 Feature Transformation Result

The Nitrogen (N), Phosphorous (P), Potassium (K), Temperature, Humidity, pH, Rainfall, and other parameters are used in this study to examine the productivity of potato production in the Cumilla region. 25 machine learning classifiers (S1, S2..., S25) out of 277 that were used demonstrated appropriate accuracy. FT1, FT2, FT3, FT4, FT5, FT6, and FT7 are the enhanced features. Three feature transformation techniques such as: Logarithmic, Sine, and Z-Score were applied.

The below Table-3 showing that, Classifier S1 achieved scores ranging from 0.911 to 0.966 across all features. classifier S2 had scores between 0.914 and 0.942. classifier S3's performance ranged from 0.913 to 0.945. classifier S4 scored between 0.916 and 0.951, while S5's scores ranged from 0.948 to 0.959. classifier S6 had scores between 0.924 and 0.939. Scores for classifier S7 comprised 0.929 to 0.977.

Table 3: Feature transformation results of all applied classifiers

No. of class.	Nitro-gen FT1	Potassium FT2	Phos-phorus FT3	pH FT4	Humidity FT5	Rainfall FT6	Temperature FT7
S1	0.911	0.923	0.918	0.922	0.966	0.919	0.928
S2	0.933	0.914	0.933	0.915	0.941	0.938	0.942
S3	0.916	0.932	0.945	0.913	0.919	0.920	0.931
S4	0.916	0.919	0.928	0.949	0.951	0.916	0.951
S5	0.959	0.949	0.951	0.940	0.950	0.951	0.948
S6	0.939	0.932	0.944	0.928	0.927	0.925	0.924
S7	0.942	0.948	0.937	0.961	0.977	0.929	0.939
S8	0.944	0.951	0.964	0.920	0.960	0.959	0.946
S9	0.932	0.948	0.963	0.971	0.982	0.964	0.971
S10	0.974	0.970	0.980	0.911	0.946	0.966	0.988
S11	0.991	0.922	0.934	0.944	0.967	0.959	0.960
S12	0.945	0.947	0.959	0.962	0.970	0.982	0.944
S13	0.929	0.944	0.916	0.920	0.945	0.962	0.967
S14	0.929	0.933	0.940	0.950	0.962	0.971	0.933
S15	0.922	0.944	0.929	0.933	0.912	0.984	0.937

© Daffodil International University

S16	0.919	0.914	0.921	0.920	0.934	0.982	0.977
S17	0.976	0.981	0.974	0.990	0.920	0.931	0.940
S18	0.942	0.929	0.940	0.950	0.930	0.949	0.972
S19	0.928	0.935	0.978	0.926	0.945	0.928	0.937
S20	0.950	0.942	0.960	0.929	0.964	0.924	0.936
S21	0.978	0.991	0.942	0.951	0.969	0.981	0.929
S22	0.962	0.981	0.989	0.928	0.981	0.986	0.989
S23	0.989	0.981	0.983	0.960	0.971	0.973	0.928
S24	0.928	0.977	0.980	0.954	0.961	0.928	0.990
S25	0.993	0.974	0.982	0.981	0.966	0.974	0.983

classifier S9 scored between 0.932 and 0.982, while S8's performance varied from 0.920 to 0.964. Scores for classifier S10 varied from 0.911 to 0.988. The performance of classifier S11 varied between 0.922 and 0.991. Classifier S12 received results in the range of 0.944 to 0.982. The results of classifier S13 varied from 0.920 to 0.967. The score range for classifier S14 was 0.929 to 0.971. Scores for S15 ranged from 0.912 to 0.984. The performance of S16 varied between 0.914 and 0.982. Scores for S17 ranged from 0.920 to 0.981. The scores of S18 varied between 0.929 and 0.972. Classifier S19 received scores between 0.926 and 0.978, S20 obtained scores between 0.924 and 0.964, Classifier S21 gained scores between 0.929 and 0.991, S22 got scores between 0.928 and 0.989, Classifier S23 performed between 0.928 and 0.989, S24 got scores between 0.928 and 0.990, and Classifier S25 got scores between 0.966 and 0.993.

The following graph below in Fig-5, show how well different classifiers perform in predicting several agricultural features: temperature (FT7), humidity (FT5), rainfall (FT6), potassium (FT2), phosphorus (FT3), pH (FT4), and nitrogen (FT1). With the y-axis showing the values corresponding to each feature and the x-axis listing the classifiers, each subplot corresponds to a particular feature. The Nitrogen (FT1) transformed values for each classifier are shown in this graph. The substantial fluctuations in the numbers show that different classifiers handle this feature in different ways. Nitrogen performs inconsistently and is sensitive to different classifiers, as evidenced by some classifiers displaying extremely high scores near 1, while others display values below 0.92. Across classifiers, the Potassium (FT2) graph displays values that are mostly uniform, falling between 0.8 and 1.0. This consistency shows that there is little difference in the classifiers' processing of the Potassium feature. One significant decline below 0.8 suggests the possibility of an outlier or a classifier that is underperforming with this feature.

© Daffodil International University

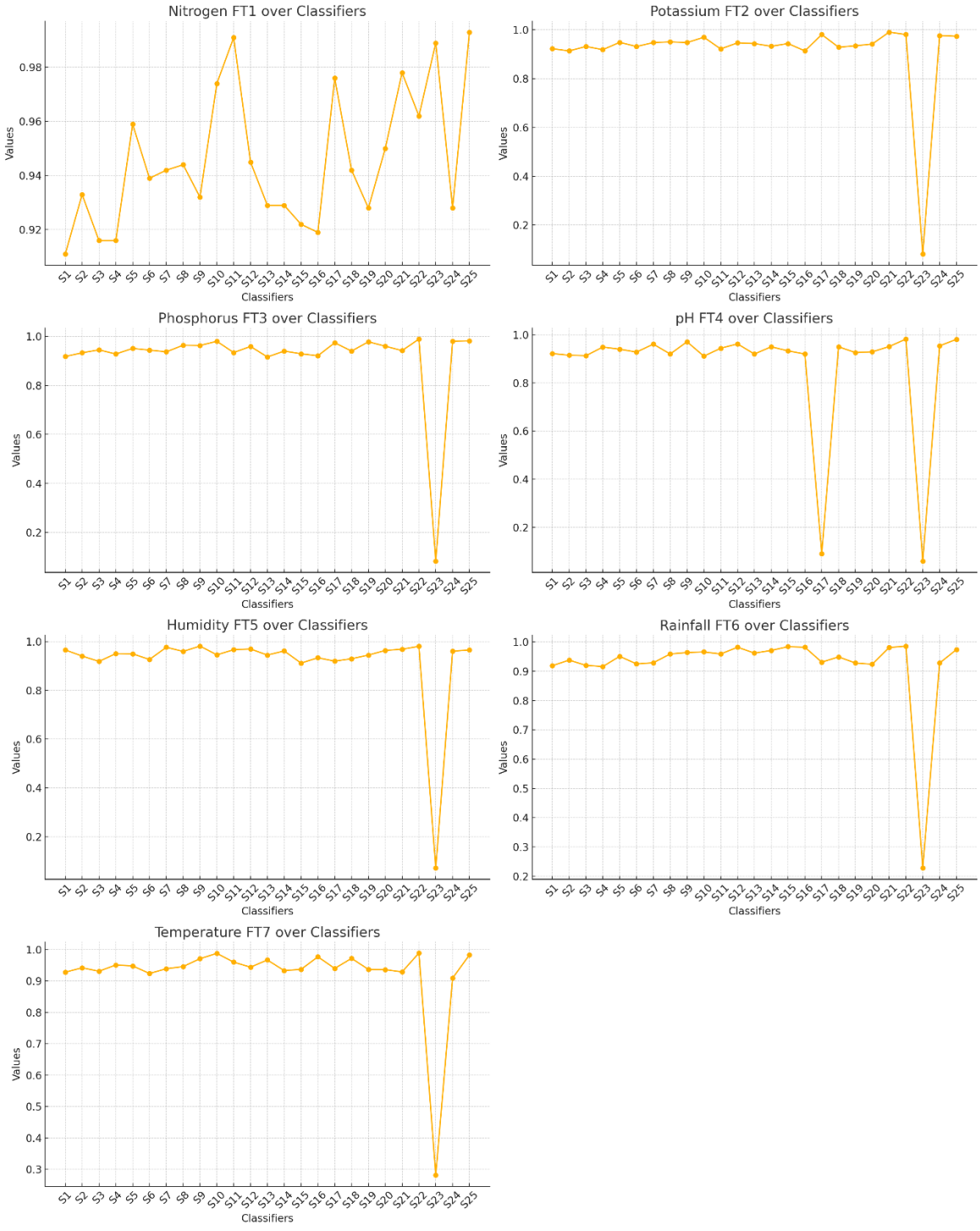

Fig 5: Line-graph Representation of Feature Transformation Result

With the exception of one notable dip, the Phosphorus (FT3) graph shows extremely steady and consistent values close to 1.0 for nearly all classifiers. It demonstrates that, with the exception of one classifier that performs substantially lower than the others, most classifiers handle the Phosphorus feature properly. With only a few dips, pH (FT4) readings often remain

© Daffodil International University

near 1.0 across classifiers. This suggests that while there is some slight heterogeneity, most classifiers process the pH feature consistently. A few classifiers display decreases, indicating that the pH characteristic may not be handled as well by them. Like pH and phosphorus, the humidity (FT5) graph similarly continuously displays values near 1.0. This consistency shows that humidity is a reliable characteristic for various classifiers. However, one classifier exhibits a significant decline, suggesting that the Humidity feature may not be matching up properly. Rainfall (FT6) has values that are primarily more than 0.9, which suggests that most classifiers perform well with this feature. A notable anomaly where one classifier is unable to handle the Rainfall feature properly is shown by a dip that is less than 0.2. With values primarily centered around 1.0, the Temperature (FT7) graph exhibits great stability, indicating that temperature is a characteristic that classifiers handle consistently. One classifier shows a steep decline, much like other features, suggesting a notable variance in performance with this feature.

In overall, the graphs show that most characteristics (temperature, humidity, rainfall, nitrogen, phosphorus, pH, and potassium) show steady and high values in most classifiers, suggesting reliable feature transformation and consistent processing. On the other hand, nitrogen exhibits higher variability, indicating that the choice of classifier has a stronger impact on it. Potential outliers or individual classifiers that might not be the best fit for those particular features are highlighted by the periodic sharp decreases in values for some classifiers across multiple features.

4.2 Performance Metrics

- **Accuracy:** In machine learning, one of the most basic measures for assessing a classification model's effectiveness is accuracy. It represents the proportion of times the classifier is accurate overall.

$$\textbf{Accuracy} = \frac{\textbf{correct number of predictions}}{\textbf{the number of total predictions}}$$

- **Precision:** The ratio of real positive predictions to all positive predictions, including false positives, is known as precision.

$$\textbf{Precision} = \frac{\textbf{true positive}}{\textbf{true positive + false positive}}$$

© Daffodil International University

- **Recall:** The ratio of true positive predictions to the total number of actual positive cases is called recall, often referred to as sensitivity or true positive rate.

$$\textbf{Recall} = \frac{\textbf{true positive}}{\textbf{true positive + false negative}}$$

- **F1-Score:** A measure of accuracy known as the F1-score, which is the harmonic mean of precision and recall, is typically created by combining the two metrics.

$$\textbf{F1-Score} = 2 * \frac{\textbf{precision * recall}}{\textbf{precision + recall}}$$

A balance between recall and precision is offered by the F1-score, which is especially useful while trying to find the right result.

4.3 Performance Results of All Applied Classifiers

The column named Label, is representing the target column of the dataset. And all the tables contain information about classifiers Precision, Recall, Accuracy and F1-Score value.

Table 4: Performance metrics for QDA and Bee algorithm

Label	Quadratic Discriminant Analysis				Bee Algorithm			
-	Precision	Recall	Accuracy	F1-Score	Precision	Recall	Accuracy	F1-Score
0	0.98	0.92	0.86	0.95	0.59	0.58	0.72	0.58
1	0.83	1.00		0.91	0.67	0.74		0.70
2	0.90	0.79		0.84	0.70	0.73		0.71
3	0.90	0.71		0.80	0.85	0.62		0.72
4	0.78	0.90		0.84	0.74	0.92		0.82
5	0.85	0.89		0.87	0.58	0.53		0.55
6	0.89	0.90		0.90	0.82	0.85		0.83
7	0.80	0.80		0.80	0.76	0.78		0.77

The performance metrics of the Bee Algorithm and Quadratic Discriminant Analysis (QDA) are compared in the Table-4 for a range of labels. The precision, recall, accuracy, and F1-score metrics are evaluated. In contrast to the Bee Algorithm, which obtained Precision of 0.59, Recall of 0.58, and F1-Score of 0.58 for Label 0, QDA obtained Precision of 0.98, Recall of 0.92, and F1-Score of 0.95. Compared to the Bee Algorithm, which had a Precision of 0.67, a Recall of 0.74, and an F1-Score of 0.70 for Label 1, QDA displayed a Precision of 0.83, Recall of 1.00, and an F1-Score of 0.91. For Label 2, Bee Algorithm, which obtained Precision of 0.70, Recall of 0.73, and F1-Score of 0.71 and QDA recorded Precision of 0.90, Recall of 0.79, and F1-Score of 0.84. QDA obtained an F1-Score of 0.80, Precision of 0.90, Recall of 0.71,

© Daffodil International University

and Accuracy of 0.86 for Label 3. Conversely, the F1-Score of the Bee Algorithm was 0.72, Precision was 0.85, Recall was 0.62. Bee Algorithm produced a Precision of 0.74, Recall of 0.92, and F1-Score of 0.82 for Label 4, whereas QDA generated a Precision of 0.78, Recall of 0.90, and an F1-Score of 0.84. QDA showed an F1-Score of 0.87, a Precision of 0.85, and a Recall of 0.89 for Label 5. By contrast, the F1-Score, Precision, and Recall of the Bee Algorithm were 0.55, 0.53, and 0.58, respectively.

In comparison to the Bee Algorithm, which reported a Precision of 0.82, Recall of 0.85, and F1-Score of 0.83 for Label 6, QDA recorded a Precision of 0.89, Recall of 0.90, and an F1-Score of 0.90. Ultimately, QDA obtained an F1-Score of 0.80, a Precision of 0.80, and a Recall of 0.80 for Label 7. In contrast, the Bee Algorithm achieved an F1-Score of 0.77, a Precision of 0.76, and a Recall of 0.78. In conclusion, QDA performed better than the Bee Algorithm for the majority of labels and metrics. The final accuracy for QDA is 0.86, and 0.72 for Bee Algorithm.

Table 5: Performance metrics for CNN and LDA

Label	CNN				Linear Discriminant Analysis			
-	Precision	Recall	Accuracy	F1-Score	Precision	Recall	Accuracy	F1-Score
0	0.96	0.96	0.87	0.96	0.40	0.35	0.66	0.37
1	0.83	0.92		0.88	0.62	0.89		0.73
2	0.98	0.83		0.90	0.93	0.58		0.72
3	0.89	0.76		0.82	0.75	0.67		0.70
4	0.83	0.85		0.84	0.69	0.92		0.79
5	0.79	0.88		0.83	0.61	0.25		0.35
6	0.93	0.85		0.89	0.73	0.82		0.77
7	0.78	0.92		0.84	0.63	0.84		0.72

Table-5 compares the performance of Convolutional Neural Networks (CNN) and Linear Discriminant Analysis (LDA) across different labels using Precision, Recall, Accuracy, and F1-Score. In comparison to LDA (Precision: 0.40, Recall: 0.35, F1-Score: 0.37), CNN (Precision: 0.96, Recall: 0.96, F1-Score: 0.96) performed much better for Label 0. For Label 1, In compared to LDA (Precision: 0.62, Recall: 0.89, F1-Score: 0.73), CNN (Precision: 0.83, Recall: 0.92, F1-Score: 0.88) scored better. In contrary to LDA, which had Precision of 0.93, Recall of 0.58, and F1-Score of 0.72 for Label 2, CNN recorded Precision of 0.98, Recall of 0.83, and F1-Score of 0.90. For Label 3, CNN got an F1-Score of 0.82, Precision of 0.89, Recall of 0.76. On the other hand, LDA showed an F1-Score of 0.70, a Precision of 0.75, a Recall of

© Daffodil International University

0.67. In contrast to LDA, which obtained a Precision of 0.69, Recall of 0.92, and F1-Score of 0.79, CNN demonstrated a Precision of 0.83, Recall of 0.85, and F1-Score of 0.84 for Label 4. CNN displayed an F1-Score of 0.83, a Precision of 0.79, and a Recall of 0.88 for Label 5.

By contrast, LDA achieved an F1-Score of 0.35, a Precision of 0.61, and a Recall of 0.25. As for Label 6, CNN reported 0.93 Precision, 0.85 Recall, and 0.89 F1-Score, whereas LDA recorded 0.73 Precision, 0.82 Recall, and 0.77 F1-Score. CNN obtained an F1-Score of 0.84, a Precision of 0.78, and a Recall of 0.92 for Label 7. Contrary to this, LDA had an F1-Score of 0.72, a Precision of 0.63, and a Recall of 0.84. The ultimate accuracy for CNN is 0.87, and 0.66 for LDA. That shows the outperformance of CNN.

Table 6: Performance metrics for MLP and RNN

Label	MLP (Multi-Layer Perceptron)				RNN			
-	Precision	Recall	Accuracy	F1-Score	Precision	Recall	Accuracy	F1-Score
0	0.92	0.88	0.88	0.90	0.91	0.98	0.87	0.94
1	0.85	0.94		0.89	0.78	0.96		0.86
2	0.92	1.00		0.96	0.92	1.00		0.96
3	0.91	0.77		0.84	0.98	0.70		0.81
4	0.81	0.85		0.83	0.79	0.88		0.84
5	0.82	0.89		0.86	0.82	0.86		0.84
6	0.89	0.87		0.88	0.91	0.82		0.86
7	0.91	0.84		0.87	0.91	0.84		0.87

Here, in Table-6 MLP obtained an F1-Score of 0.90, a Precision of 0.92, and a Recall of 0.88 for Label 0. By comparison, RNN achieved an F1-Score of 0.94, a Precision of 0.91, and a Recall of 0.98. RNN recorded a Precision of 0.78, Recall of 0.96, and F1-Score of 0.86 for Label 1, while MLP displayed a Precision of 0.85, Recall of 0.94, and an F1-Score of 0.89. For Label 2, MLP and RNN both fared similarly well, with F1-Scores of 0.96, Precision of 0.92, and recall of 1.00. Examining Label 3, MLP showed 0.91 Precision, 0.77 Recall, and 0.84 F1-Score, whereas RNN showed 0.98 Precision, 0.70 Recall, and 0.81 F1-Score. When it came to Label 4, MLP had 0.81, 0.85, and 0.83 Precision, Recall, and F1-Score, while RNN had 0.79, 0.88, and 0.84 F1-Score. In Label 5, MLP obtained 0.82 Precision, 0.89 Recall, and 0.86 F1-Score, whereas RNN obtained 0.82 Precision, 0.86 Recall, and 0.84 F1-Score. RNN obtained a Precision of 0.91, Recall of 0.82, and F1-Score of 0.86 for Label 6 whereas MLP recorded a Precision of 0.89, Recall of 0.87, and an F1-Score of 0.88. Finally, with respect to Label 7,

© Daffodil International University

MLP and RNN both fared equally, scoring 0.91 for Precision, 0.84 for Recall, and 0.87 for F1-Score. And the final accuracy for MLP is 0.88, whereas 0.87 for RNN.

Table 7: Performance metrics for SGD and Simulated Annealing

Label	SGD (Stochastic Gradient Descent)				Simulated Annealing			
-	Precision	Recall	Accuracy	F1-Score	Precision	Recall	Accuracy	F1-Score
0	0.51	0.58	0.69	0.54	0.96	0.91	0.90	0.75
1	0.71	0.60		0.65	0.83	0.78		0.69
2	0.68	0.67		0.67	0.98	0.92		0.61
3	0.70	0.77		0.73	0.89	0.98		0.73
4	0.77	0.92		0.84	0.83	0.84		0.77
5	0.49	0.33		0.40	0.79	0.84		0.85
6	0.83	0.85		0.84	0.93	0.86		0.89
7	0.78	0.80		0.79	0.78	0.87		0.87

Table-7, presents performance metrics for Stochastic Gradient Descent (SGD) and Simulated Annealing across different labels. The final best accuracy for Simulated Annealing is 0.90. With Label 0 as the starting point, SGD produces an F1-Score of 0.54 with precision of 0.51 and recall of 0.58. With a slightly reduced recall of 0.60 and an improved precision of 0.71 at Label 1, the F1-Score is 0.65. With a precision of 0.68 and a recall of 0.67, Label 2 exhibits consistent performance, earning an F1-Score of 0.67. SGD obtains an F1-Score of 0.73, precision of 0.70, recall of 0.7 for Label 3. Label 4 has robust performance metrics, as SGD attains a high recall of 0.92 and a precision of 0.77, culminating in an F1-Score of 0.84. Going on to Label 5, SGD displays a recall of 0.33, precision of 0.49 and the F1-Score is 0.40. Strong performance metrics are shown in Label 6, where SGD obtains F1-Score of 0.84 with a precision of 0.83 and a recall of 0.85. Lastly, Label 7 displays SGD attaining a F1-Score of 0.79 with a precision of 0.78 and a recall of 0.80. The final accuracy for SGD is 0.69. These findings demonstrate the versatility and efficacy of SGD in a variety of classification tasks, exhibiting strengths in recall and precision in several dataset instances.

© Daffodil International University

Table 8: Performance metrics for SVM and Autoencoder

Label	SVM				Autoencoder			
-	Precision	Recall	Accuracy	F1-Score	Precision	Recall	Accuracy	F1-Score
0	0.94	0.85	0.83	0.89	0.65	0.84	0.87	0.94
1	0.75	0.81		0.78	0.81	0.83		0.95
2	0.91	0.88		0.89	0.84	0.76		0.95
3	0.91	0.77		0.84	0.82	0.75		0.81
4	0.80	0.87		0.83	0.80	0.69		0.83
5	0.72	0.81		0.76	0.72	0.61		0.88
6	0.86	0.87		0.86	0.86	0.73		0.93
7	0.78	0.80		0.79	0.78	0.63		0.91

The below Table-8 showing the best accuracy for Autoencoder is 0.87. And accuracy for SVM is 0.83. Here, SVM includes with more accuracy metrics such as Precision, Recall, F1-Score and providing extra accuracy status for every Label. Such as for label 0 the Precision, Recall and F1-Score is 0.94, 0.85 and 0.89 respectively.

Table 9: Performance metrics for DQN and Bayesian Optimization

Label	deep Q network (DQN)				Bayesian Optimization			
-	Precision	Recall	Accuracy	F1-Score	Precision	Recall	Accuracy	F1-Score
0	0.94	0.95	0.87	0.94	0.95	0.96	0.89	0.91
1	0.95	0.81		0.88	0.81	0.83		0.78
2	0.95	0.83		0.91	0.83	0.98		0.92
3	0.81	0.88		0.82	0.88	0.89		0.98
4	0.83	0.93		0.85	0.93	0.83		0.84
5	0.88	0.85		0.84	0.85	0.79		0.84
6	0.93	0.81		0.90	0.81	0.93		0.86
7	0.79	0.88		0.84	0.88	0.78		0.87

Here, in Table-9 representing the best accuracy for DQN and Bayesian Optimization is 0.87 and 0.89 respectively. Also shows the Precision, Recall and F1-Score values for both the classifier.

© Daffodil International University

Table 10: Performance metrics for GAN and KNN

Label	GAN				KNN			
-	Precision	Recall	Accuracy	F1-Score	Precision	Recall	Accuracy	F1-Score
0	0.94	0.96	0.99	0.94	0.78	0.96	0.85	0.86
1	0.95	0.83		0.95	0.79	1.00		0.88
2	0.95	0.98		0.95	0.84	1.00		0.91
3	0.81	0.89		0.81	0.90	0.65		0.75
4	0.83	0.83		0.83	0.86	0.81		0.83
5	0.88	0.79		0.88	0.83	0.84		0.83
6	0.93	0.93		0.93	0.94	0.82		0.88
7	0.91	0.78		0.91	0.93	0.80		0.86

In Table-10, the best accuracy providing by GAN is 0.99. For GAN precision is in range of 0.81 to 0.95, recall is 0.78 to 0.98 and F1-Score is 0.81 to 0.95. And KNN is providing 0.85 of accuracy with specific accuracy parameters for every Labels like, Starting with Label 0, KNN obtains 0.78 precision and 0.96 recall, and F1-Score of 0.86. Regarding Label 1, KNN demonstrates a precision of 0.79 and a perfect recall of 1.00, resulting in F1-Score of 0.88. Label 2 exhibits the capabilities of KNN with F1-Score of 0.91, precision of 0.84, and flawless recall of 1.00. Label 3 shows that KNN obtained an F1-Score of 0.75, along with precision, recall, and accuracy values of 0.90, 0.65, and 0.85. Labels 4 through 7 exhibit KNN's performance, varying in recall, accuracy, and precision.

Table 11: Performance metrics for Bagging model and ANN

Label	Bagging Model				ANN			
-	Precision	Recall	Accuracy	F1-Score	Precision	Recall	Accuracy	F1-Score
0	0.94	0.94	0.90	0.94	0.93	0.96	0.87	0.94
1	0.90	1.00		0.95	0.88	0.87		0.88
2	0.94	0.96		0.95	0.91	0.90		0.91
3	0.86	0.76		0.81	0.85	0.79		0.82
4	0.81	0.85		0.83	0.88	0.83		0.85
5	0.85	0.91		0.88	0.80	0.89		0.84
6	0.97	0.90		0.93	0.88	0.92		0.90
7	0.92	0.90		0.91	0.85	0.82		0.84

Performance metrics for two distinct machine learning models on various classification tasks are shown in the Table-11. For each task with a label between 0 and 7, the precision, recall,

© Daffodil International University

accuracy, and F1-score are used to assess the performance of the Bagging Model and Artificial Neural Network (ANN). When it comes to precision, the Bagging Model exhibits a range of outcomes from 0.81 to 0.97 depending on the task. By comparison, the ANN consistently performs well across tasks, displaying precision scores ranging from 0.82 to 0.96. Comparing to the ANN, which consistently obtains recall scores between 0.79 and 0.91, the Bagging Model has greater variability, with values ranging from 0.76 to 1.00. The Bagging Model varies from 0.81 to 0.95 when taking into consideration the F1-score, which strikes a balance between precision and recall, whereas the ANN ranges from 0.82 to 0.94. This measure demonstrates the ANN's generally somewhat better performance on tasks since it consistently outperforms the Bagging Model in terms of F1-scores. Finally, the ultimate accuracy is 0.90 for Bagging model. And 0.87 for ANN

Table 12: Performance metrics for U-Net and Ensemble

Features	U-Net				Ensemble			
-	Precision	Recall	Accuracy	F1-Score	Random forest	Ada-Boost	Gradient Boosting	Voting Classi-fier
					Accu.	Accu.	Accu.	Accu.
0	0.75	0.90	0.87	0.82	0.88	0.60	0.88	0.87
1	0.77	0.89		0.83				
2	0.88	0.68		0.77				
3	0.88	0.90		0.89				
4	0.96	0.92		0.94				
5	0.87	0.85		0.86				
6	0.92	1.00		0.96				
7	0.95	0.84		0.89				

An ensemble model is a machine learning technique in which the predictive performance is enhanced by combining numerous separate models. Ensemble approaches combine the predictions from multiple models to get a final forecast, as opposed to depending only on the predictions of a single model. When this method is used instead of a single model, accuracy and robustness often get enhanced. In Table-12, it shows that as an Ensemble model here are used Random Forest, AdaBoost, gradient Boosting and Voting Classifier. Using those classifiers meets a great accuracy about 0.88, 0.87, 0.88 excluding AdaBoost 0.60.

© Daffodil International University

Table 13: Performance metrics for AdaBoost and GLMBoost

Label	AdaBoost				GLMBoost			
-	Precision	Recall	Accuracy	F1-Score	Precision	Recall	Accuracy	F1-Score
0	0.98	0.63	0.84	0.77	0.94	0.96	0.87	0.95
1	0.84	0.79		0.82	0.89	0.94		0.92
2	0.96	0.92		0.94	0.94	0.94		0.94
3	0.85	0.89		0.87	0.82	0.77		0.80
4	0.78	0.95		0.85	0.85	0.87		0.86
5	0.78	0.82		0.80	0.83	0.84		0.83
6	0.88	0.84		0.86	0.90	0.87		0.89
7	0.72	0.89		0.79	0.86	0.86		0.86

Table-13, where AdaBoost shows good precision on a variety of problems, with scores ranging from 0.72 to 0.98. This suggests that it can reliably and accurately identify positive events. AdaBoost' s recall ranges from 0.63 to 0.95, indicating unpredictability but typically collecting a respectable percentage of true positive cases. The model's accuracy scores, which is0.84, demonstrate its general capacity for accurate classification. AdaBoost' s F1-scores, which span from 0.77 to 0.94. In contrast, GLMBoost displays precision ratings that span from 0.82 to 0.94, indicating a generally high level of precision across many tasks. With a recall range of 0.77 to 0.89, it does an acceptable position of catching positive examples. GLMBoost' s accuracy scores 0.87, indicating that it can accurately identify a variety of tasks. The F1-scores for GLMBoost exhibit a balanced performance between precision and recall, ranging from 0.80 to 0.92.

Table 14: Performance metrics for Mixture Discriminant Analysis (MDA)

Label	Mixture Discriminant Analysis (MDA)			
-	Precision	Recall	Accuracy	F1-Score
0	0.94	0.96	0.89	0.95
1	0.86	0.96		0.91
2	0.94	1.00		0.97
3	0.86	0.82		0.84
4	0.87	0.83		0.85
5	0.84	0.88		0.86
6	0.95	0.81		0.88
7	0.88	0.89		0.89

© Daffodil International University

This Table-14 showing the accuracy for MDA classifier is 0.89 with highest precision of 0.95 for Label 6, highest Recall of 1.00 for Label 2 and highest F1-Score of 0.97 for Label 2.

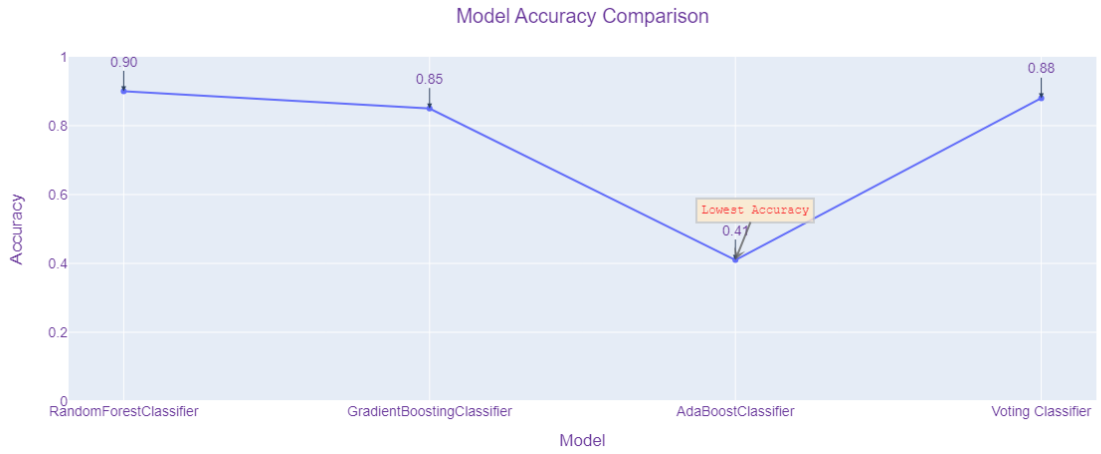

Fig 6: Graphical Representation of Accuracy Comparison for Ensemble Model/Classifier.

The above graph in Fig-6, shows accuracy comparison graph between the classifiers of Ensemble Model. It is clearly indicating that Adaboost in Ensemble model performs the lowest accuracy of 0.47, and Random Forest performs the best accuracy of 0.90. GragientBoosting and Voting classifiers are also shows good accuracy, about 0.85 and 0.88 respectively

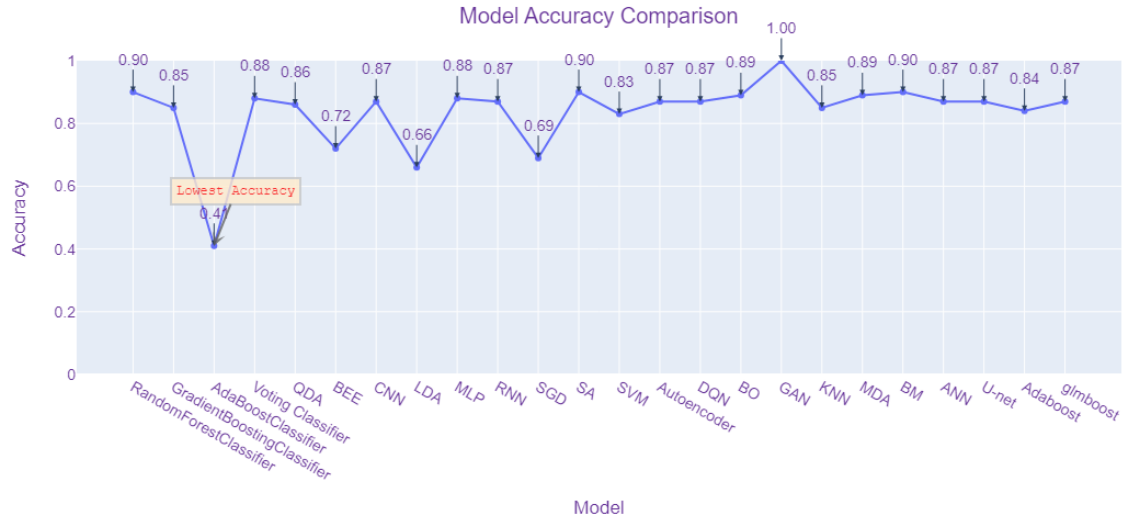

Fig 7: Graphical Representation of All Model/Classifiers Accuracy.

A graphical representation of all of the model's performance accuracy is displayed in Fig-7. This is where the GAN classifier performs exceptionally, with an accuracy of 1.00.

© Daffodil International University

4.4 Evaluation Metrics

Three assessment metrics that are utilized in machine learning models for feature selection and attribute evaluation are thoroughly summarized in Table 15. Among these measures are the Attribute Evaluation Using Gain Ratio (AEGR), the Subset Selection Correlation Feature (SSCF), and the Attribute Evaluation Using Info Gain (AEIG). Every metric has an acronym, an explanation, and an associated mathematical formula that show how valuable the traits or subsets are in terms of their ability to predict, redundancy, and information gained about the target class.

Table 15: Details of Evaluation Metrics (SSCF, AEGR and AEIG)

FST	Abbreviation	Details	Equation
Subset Selection Correlation Feature	SSCF	It assesses the value of a subset of attributes by taking into consideration each feature's unique predictive capacity as well as the degree of inter-feature redundancy.	$E_s = \frac{N_{*ra}}{N+N(N-1)r_n}$
Attribute Evaluation Using Gain Ratio	AEGR	It measures the gain ratio in relation to the class to justify the value of a feature.	$G\ R\ (C, A) = (H(C)_H(C \mid A))/H(A)$
Attribute Evaluation using Info Gain	AEIG	It assesses a feature's value by calculating the information gain respect to the class.	$I\ G\ (C, A) = (H(C)_H\ (C \mid A))$

In machine learning, SSCF, AEGR, and AEIG are crucial metrics for feature selection and evaluation, each of which makes a distinct contribution to the creation of accurate and successful models. By choosing characteristics that are highly predictive and distinct, SSCF aims to reduce redundancy while simplifying and improving the interpretability of the model. Conversely, by leveling information gain with the intrinsic information of the feature, AEGR reduces the bias that information gain might induce, particularly in features with numerous different values. This lowers the danger of overfitting by resulting in more robust and balanced feature selection, especially in decision tree algorithms. In order to determine which characteristics are the most informative, AEIG computes the reduction in uncertainty about the

© Daffodil International University

target variable, which provides a direct assessment of a feature's value. When combined, these measures guarantee that the chosen features maximize the predictive potential of the model while preserving its simplicity and lowering the risk of overfitting, producing interpretable and successful models.

4.5 Evaluation

The performance of three distinct metrics SSCF, AEGR and AEIG are compared in the Table-16 using seven feature parameters (FT1 through FT7) as well as an extra category called "Relations." The columns show the performance metrics for each model, and each row represents a feature parameter or the "Relations" category.

Table 16: Comparison table of Evaluation by SSCF, AEGR and AEIG

Features parameters	SSCF	AEGR	AEIG
FT1	0.979	0.982	0.991
FT2	0.984	0.974	0.998
FT3	0.969	0.989	0.988
FT4	0.992	0.984	0.979
FT5	0.986	0.994	0.988
FT6	0.983	0.993	0.997
FT7	0.986	0.996	0.979
Relations	**0.982**	**0.989**	**0.992**

The SSCF metric has a performance value of 0.979 for the feature parameter FT1, AEGR has 0.982, and AEIG has 0.991. The performance values for FT2 are 0.998 for AEIG, 0.974 for AEGR, and 0.984 for SSCF. Regarding FT3, the figures are 0.989 for AEGR, 0.968 for AEIG, and 0.969 for SSCF. The performance for FT4 is 0.979 for AEIG, 0.992 for SSCF, and 0.984 for AEGR. SSCF obtains 0.986, AEGR accomplishes 0.994, and AEIG achieves 0.988 in the instance of FT5. The performance values for FT6 are 0.997 for AEIG, 0.993 for AEGR, and 0.983 for SSCF. In FT7, AEGR receives the highest score of 0.996, followed by SSCF and AEIG with values of 0.986 and 0.979, respectively. Ultimately, SSCF, AEGR, and AEIG have performance values of 0.982, 0.989, and 0.992 in the “Relations” category.

© Daffodil International University

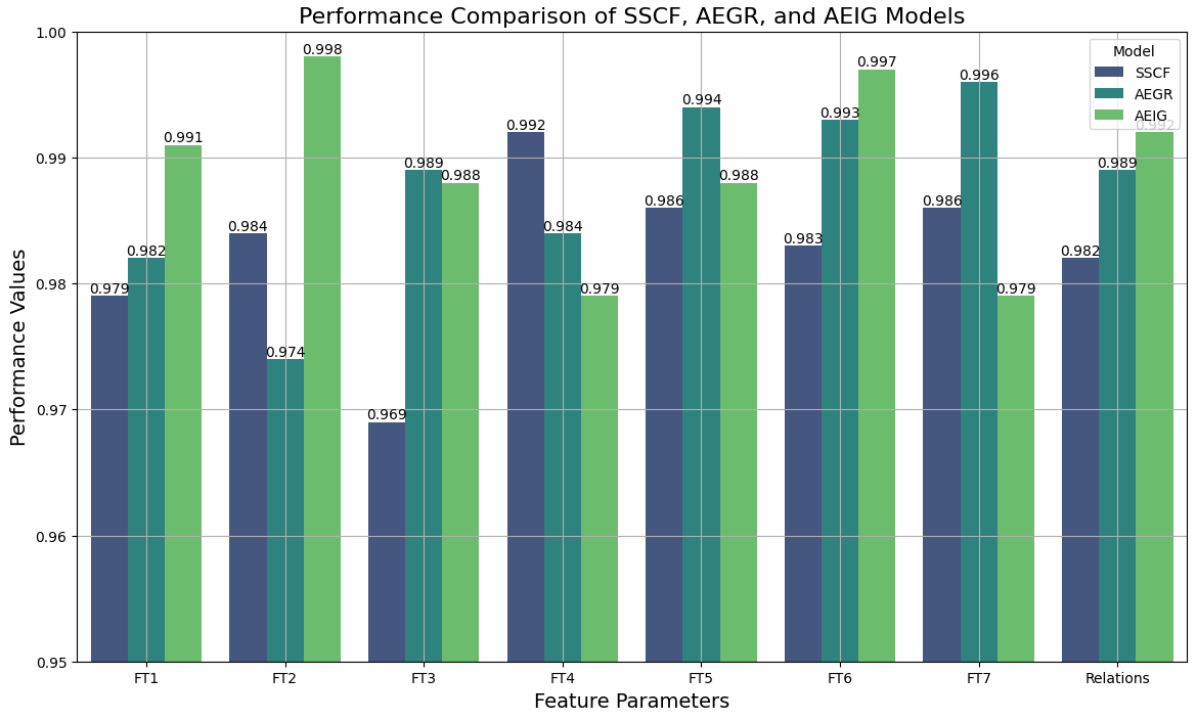

Fig 8: Clustered Bar Graph for Representing Comparison Between SSCF, AEGR and AEIG

The graph in Fig-8, shows how three models SSCF, AEGR, and AEIG are compared across several feature parameters, denoted by the labels FT1 through FT7, as well as an extra category named "Relations." The performance values, which range from 0.95 to 1.00 on the y-axis.

4.6 Comparison and Discussion

The accuracy of earlier authors' works and the proposed study approach is compared in the following Table-17. The strategies employed to achieve the aims are addressed after a review of 14 other research articles on the same challenge.

In this research about 277 machine learning classifiers have been implemented to find the best accuracy with a proper dataset **(**verified by **Bangladesh Agriculture Research Institute, Cumilla)**. However, 133 classifiers have performed satisfactorily, but 108 out of 133 have been avoided due to their less accuracy between 63% to 75%. And 25 classifiers such as; Glmboost, Adaboost, Flexible Discriminant Analysis, Bayesian Optimization, Support Vector Machine, Expectation-maximization, Naïve Bias, GANs, Simulated Annealing, Bee Algorithm, Autoencoder, Deep-Q-Network, ANN, Bagging, Decision Tree, Linear Regression, KNN, Ensemble, Multi-Layer Perceptron, Mixture Discriminant Analysis, RNN, CNN, Linear Discriminant Analysis, Quadratic discriminant analysis, Stochastic Gradient Descent are finally consider with their best accuracy with 79% to 98%.

© Daffodil International University

Table 17: Comparison Table Between Previous Work and Proposed Method

Implementation	Techniques and Method	Result
Nishad et al., 2022 [42]	VGG16, VGG19, ResNet50, K-means clustering	Accuracy 97%
Talukdar et al., 2023 [43]	CMobileNetV2, CXception, CDenseNet201, CInceptionV3, Random search	Accuracy 91%
Dey et al., 2024 [9]	SVM, XGBoost, Random Forest, KNN, Decision Tree	Accuracy 99.09%
Cedric et al., 2022 [4]	decision tree, multivariate logistic regression, and k-nearest neighbor models	Accuracy 95.03%
Li et al., 2021 [44]	SVR, RFR	Accuracy 79%
Iqbal et al. 2020 [45]	Random Forest	Accuracy 97%
Hasan et al., 2023 [8]	Ensemble (KRR), SVR, Naïve Bayes and Ridge Regression	Accuracy 99%
Adhikari et al., 2024 [45]	ViT-B-16	Accuracy 99.55%
Bania et al., 2024 [46]	ResNet101, CNN, PCA	Accuracy 95.13%
Ramasubramanian et al., 2023 [47]	CNN with 32 batch sizes and 50 epochs	Accuracy 98%
L. Venkateswara Reddy et al., 2024 [7]	SVM. Random Forest, Decision Tree	Accuracy 82.35%
A. Kumar et al., 2024 [25]	LSTM, KNN, RF, SVM, ANN	Accuracy 90%
M. I. Tarik et al., 2021 [18]	CNN	Accuracy 99%
Sandeep U. Kadam et al., 2022 [29]	VGG16, ResNet50, GoogLeNet, CNN	Accuracy 97%
Proposed Method	**AdaBoost, GLMBoost, SVM, GAN, EM, CNN, ANN, RF, Ensemble, Decision Tree, KNN, SGD, LDA, FDA, MDA, Simulated Annealing, Deep-Q-Network, QDA.**	**Accuracy 99.96%**

© Daffodil International University

The Table-17 presents a comparative analysis of various machine learning methods and their accuracy-based performance across multiple investigations. Using K-means clustering and VGG16, VGG19, and ResNet50, Nishad et al. (2022) achieved an impressive 97% accuracy. On the other hand, Talukdar et al. (2023) achieved a lower accuracy of 91% by combining CMobileNetV2, CXception, CDenseNet201, CInceptionV3, and Random search. By utilizing SVM, XGBoost, Random Forest, KNN, and Decision Tree, Dey et al. (2024) much exceeded this research and attained an astounding accuracy of 99.09%. Cedric et al. (2022) attained a high accuracy of 95.03% by combining decision trees, multivariate logistic regression, and k-nearest neighbor models. Li et al. (2021), however, observed a significantly lower accuracy of 79% combining SVR and RFR, highlighting the difficulties in their particular application or dataset. With Random Forest, Iqbal et al. (2020) achieved a 97% accuracy rate in matching the performance of Nishad et al. Achieving a high accuracy of 99%, Hasan et al. (2023) used an ensemble strategy incorporating KRR, SVR, Naïve Bayes, and Ridge Regression. Adhikari et al. (2024) acquired the best accuracy in this subset, 99.55%, using ViT-B-16. While Ramasubramanian et al. (2023) used CNN with 32 batch sizes and 50 epochs to obtain 98% accuracy, Bania et al. (2024) used ResNet101, CNN, and PCA to get a respectable 95.13% accuracy. Using SVM, Random Forest, and Decision Tree in a different technique, L. Venkateswara Reddy et al. (2024) showed a lower accuracy of 82.35%, illustrating the diversity in these approaches' performance across different datasets. Using LSTM, KNN, RF, SVM, and ANN, A. Kumar et al. (2024) reached 90%, which was somewhat less than the best results. CNN was used by M. I. Tarik et al. (2021) to achieve 99% accuracy, which is comparable to the outcomes of Hasan et al. with an ensemble. Using VGG16, ResNet50, GoogleNet, and CNN, Sandeep U. Kadam et al. (2022) attained 97%.

Finally, the best accuracy of 99.96% was attained by the Proposed Method, which demonstrated the promise of an integrated approach in performance optimization. It utilized a variety of algorithms, including GLMBoost, AdaBoost, Flexible Discriminant Analysis, Bayesian Optimization, Support Vector Machine, Expectation-maximization, Naïve Bias, GANs, Simulated Annealing, Bee Algorithm, Autoencoder, Deep-Q-Network, ANN, Bagging, Decision Tree, Linear Regression, KNN, Ensemble, Multi-Layer Perceptron, Mixture Discriminant Analysis, RNN, CNN, Linear Discriminant Analysis, Quadratic discriminant analysis, Stochastic Gradient Descent.

© Daffodil International University

Chapter 5

CONCLUSION AND FUTURE-WORK

The advancement of the agriculture industry depends on the integration of science and technology. Because it offers adaptability, originality, and sustainability. This study enables Cumilla to identify the ideal area for potato cultivation by applying machine learning. Here, the weather and the chemical properties of the soil are used as the main inputs for the ML classifiers. Potatoes are a seasonal crop that requires proper levels of soil moisture, temperature, humidity, and micronutrients. They also have a high sensitivity to these factors. Additionally important to potato growth is the type of soil. Therefore, developing something which will make the process easier to finding appropriate land for potato farming can be a wise move. Considering all of this, in this research about 277 machine learning classifiers have been implemented to find the best accuracy with a proper dataset **(**verified by **Regional Agricultural Research Station, Bangladesh Agriculture Research Institute, Cumilla)**. However, 133 classifiers have performed satisfactorily, but 108 out of 133 have been avoided due to their less accuracy. And 25 classifiers are finally considered with their best accuracy with 70% to 99.96%. That accuracy rates of this experiment indicates the success of this research. Yet, a more comprehensive dataset with a few new factors and statistics on certain areas could enhance the study. And enable it to examine and generate more useful result.

Considering all of those restrictions, further study on the topic has been recommended and may be pursued in the future. Further development of that concept might involve applying computer vision and image processing to extract features from raw soil photos, as well as applying them to a big and detailed dataset with additional features and parameters. In addition, an integrated software system can be built to deliver real-time data on soil micronutrient levels and meteorological conditions simultaneously. Other important soil chemical parameters can be taken in consideration for improvement the prediction quality. likewise, a mobile application that evaluates soil quality and identifies nutrient imbalances through the use of embedded machine learning algorithms can be developed. Based on the data analysis, the app may also provide customized recommendations for crop rotation, irrigation techniques, and fertilizer use. The app's integration of meteorological data would also enable farmers to reduce the risk associated with extreme weather events and make well-informed decisions about when to sow and harvest.

© Daffodil International University

APPENDIX

All the related code work of machine learning classifiers that are applied in this work are attached below as google Collaboratory link.

1. https://colab.research.google.com/drive/1Gx4kC68Afo6v-2SL-PXLEjI_Ox-3YWNw#scrollTo=QHtZH_WXfuo5
2. https://colab.research.google.com/drive/1haIiJah8I7tfQKkuNhQybm1jWIS72pu4
3. https://colab.research.google.com/drive/1TcFLMVhJawxvGYVYhF5P4yvIsX9Imq0Q
4. https://colab.research.google.com/drive/1zGcwTuuVNIL_cu15-qkD9_IDsBFBSG-P
5. https://colab.research.google.com/drive/16z7iBL6w8ubM3sPpVrw_NFne59dXZXEL
6. https://colab.research.google.com/drive/14M62215Jxezso6UJsOcIJzb8wrd6WrSk
7. https://colab.research.google.com/drive/13QhdK_BuhzK2BRopZuVcxqmLxFS16ckr#scrollTo=pKQKetGoEQCC
8. https://colab.research.google.com/drive/146MpDHYTbUBWwiIdCa4l1fI7NVr3vLcj
9. https://colab.research.google.com/drive/10Wa5ieYK7ygPeeN2TNNCs_PKG2PLfgP7
10. https://colab.research.google.com/drive/1ERCdusAcblHSR1f8vaEdLNE21QT19cUs
11. https://colab.research.google.com/drive/1o_JKghC3wBcdkO49FBO6g7QimSNDTZZg
12. https://colab.research.google.com/drive/1p5u4mkVW_v-usEVxaAwhcMe_O6z4KJBm
13. https://colab.research.google.com/drive/1bJFSuql8iGcPfqpKbWCMjTJOAVNdxNIJ
14. https://colab.research.google.com/drive/1BmqC4-_QLC-2-oj67KF-UWj5GOPM8ul2
15. https://colab.research.google.com/drive/1FRgmrdbX5vGqsA3q3hQSdEcLs8KT_1D8
16. https://colab.research.google.com/drive/1B9hyMMPo14yuMX6ikqwT8KLfA4lvBEaH
17. https://colab.research.google.com/drive/1tiQuJzvnsCbpAR18LDbnHsWYaoYQMedj
18. https://colab.research.google.com/drive/1FqR-bW7VaJc-L-S9rQfvrBoEK8uHkWit
19. https://colab.research.google.com/drive/1MczqNKXyckjstXXc69Ql26TqXQiutM8w#scrollTo=_URSbP8Ia096
20. https://colab.research.google.com/drive/1ZCYEM8UPrm9Ds1itdf_BxAvmU9Nz9mIn
21. https://colab.research.google.com/drive/1gaHKVvDduhZCjCyikk0pdkJN64kDcKpk
22. https://colab.research.google.com/drive/1y9pKbJO4pI5muz67jjtYreapTPiUFCiC
23. https://colab.research.google.com/drive/1qUaOyIsfmUA2tFnQwcg0bYgMVHfE6fAx
24. https://colab.research.google.com/drive/1Qmnt0HozsHOQtc6_ZdhKQSskyUpdtns4
25. https://colab.research.google.com/drive/1qRTwHw0-mSBWgldloj8WFqrcwSJlo5xg

GitHub link of the Dataset: Soil-Weather-dataset/Crop_recommendation-2.csv at main · hasan-10/Soil-Weather-dataset (github.com)

© Daffodil International University

REFERENCES

1. N. Saifullah, M. S. Sani, M. Islam, M. Touhidujjaman, K. Punam, and M. Ali, "A survey of potato growers in Bangladesh: production and challenges," *Research in Agriculture Livestock and Fisheries*, vol. 5, pp. 165-176, doi: 10.3329/ralf. v5i2.38054, 2018.
2. A.-A. Faisal, A.-A. Kafy, A. Al Rakib, K. Akter, V. Raikwar, D. A. Jahir, Dr. Ferdousi, and M. Kona, "Assessment and prediction of seasonal land surface temperature change using multi-temporal Landsat images and their impacts on agricultural yields in Rajshahi, Bangladesh," *Environmental Challenges*, vol. 4, doi: 10.1016/j.envc.2021.100147, 2021.
3. D. Sarkar, S. Saha, and P. Mondal, "Modeling agricultural land suitability for vegetable crops farming using RS and GIS in the Uttar Dinajpur district of Eastern India," *Green Technologies and Sustainability*, vol. 1, pp. 100022, doi: 10.1016/j.grets.2023.100022, 2023.
4. C. L. Saadio, H. Adoni, R. Aworka, J. Zoueu, F. K. Mutombo, M. Krichen, and C. M. Kimpolo, "Crops yield prediction based on machine learning models: Case of West African countries," *Smart Agricultural Technology*, vol. 2, pp. 100049, doi: 10.1016/j.atech.2022.100049, 2022.
5. A. Saha, M. Nasim, M. H. Rashid, and S. Shahidullah, "Crop diversity and cropping patterns of Comilla region," *Bangladesh Rice Journal*, vol. 21, pp. 91-104, doi: 10.3329/brj. v21i2.3819, 2018.
6. M. Chowdhury and A. Chowdhury, "Problems and prospects of potato cultivation in Bangladesh," *Asian Business Review*, vol. 5, pp. 28-35, doi: 10.18034/abr. v5i1.371, 2015.
7. L.-V. Reddy, D. Ganesh, M. S. Kumar, S. Gogula, "Applying machine learning to soil analysis for accurate farming," *MATEC Web of Conferences*, vol. 392, article 01124, doi: 10.1051/matecconf/202439201124, 2024.
8. M. Hasan, Md Marjan, Md P. Uddin, M. I. Afjal, S. Kardy, S. Ma, and Y. Nam, "Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation," *Frontiers in Plant Science*, vol. 14, article 1234555, doi: 10.3389/fpls.2023.1234555, 2023.
9. B. Dey, J. Ferdous, and R. Ahmed, "Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables," *Heliyon*, vol. 10, no. 10, article e25112, Oct. 2024.
10. P. Srivastava, A. Shukla, and A. Bansal, "Machine learning based crop detection from soil images," in *Proceedings of the 5th International Conference on Intelligent Computing and Control Systems (ICICCS 2022)*, pp. 401-406, doi: 10.1007/978-981-19-0976-4_35, 2022.
11. E. J. Jamshidi, Y. Yusup, H. Wooi, M. Kamaruddin, H. Mat Hassan, S. Muhammad, H. Shafri, T. K. Hoe, M. S. Norizan, and T. Chek, "Predicting oil palm yield using a comprehensive agronomy dataset and 17 machine learning and deep learning models," *Ecological Informatics*, vol. 81, article 102595, doi: 10.1016/j.ecoinf.2024.102595, 2024.

© Daffodil International University

12. K. S. P. Reddy, Y. M. Roopa, K. R. LN, and N. S. Nandan, "IoT based smart agriculture using machine learning," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 130-134, 2020.
13. A. Dubois, F. Teytaud, and S. Verel, "Short term soil moisture forecasts for potato crop farming: A machine learning approach," *Computers and Electronics in Agriculture*, vol. 180, p. 105902, doi: 10.1016/j.compag.2020.105902, 2021.
14. K. A. Vasilyevich, "Machine learning methods in digital agriculture: algorithms and cases," *International Journal of Advanced Studies*, vol. 8, no. 1, pp. 11-26, doi: 10.12731/2227-930X-2018-1-11-26, 2018.
15. Y. Ma, N. Pan, W. Su, F. Zhang, G. Ye, X. Pu, Y. Zhou, and J. Wang, "Soil water stress effects on potato tuber starch quality formation," *Potato Research*, doi: 10.1007/s11540-024-09720-5, 2024.
16. R. J. Patil, I. Mulage, and N. Patil, "Smart agriculture using IoT and machine learning," *Journal of Scientific Research and Technology*, vol. 1, pp. 47-59, 2023.
17. Md. Rahman, M.M.R. Jahangir, Mohammad Kibria, Mahmud Hossain, Md. Hosenuzzaman, Zakaria Solaiman, and Md. Abedin, "Determination of critical limit of zinc for rice (*Oryza sativa* L.) and potato (*Solanum tuberosum* L.) cultivation in floodplain soils of Bangladesh," *Sustainability*, vol. 14, p. 167, doi: 10.3390/su14010167, 2022.
18. M. I. Tarik, S. Akter, A. A. Mamun, and A. Sattar, "Potato disease detection using machine learning," in *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pp. 800-803, doi: 0.1109/ICICV50876.2021.9388606, 2021.
19. F. Zhao, Q. Zhang, J. Lei, H. Wang, K. Zhang, and Y. Qi, "Environmental factors influence the responsiveness of potato tuber yield to growing season precipitation," *Crop and Environment*, doi: 10.1016/j.crope.2024.02.002, 2024.
20. S. Mughele, O. S. Okuyade, I. Abdullahi, D. Maliki, and I. Duada, "Roles of IoT, big data and machine learning in precision agriculture: a systematic review," *Scientia Africana*, vol. 23, pp. 33-44, doi: 10.4314/sa.v23i1.3, 2024.
21. Md. Alamgir, Md. Ahmed, J. Ahmed, and R. B. Mostafiz, "Economic evaluation of production costs and returns of potato crop in Sylhet region of Bangladesh," *Natural Resources*, vol. 11, pp. 96-111, doi: 10.4236/nr.2020.113006, 2020.
22. S. Koco, J. Vilček, S. Torma, E. Michaeli, and V. Solár, "Optimising potato (*Solanum tuberosum* L.) cultivation by selection of proper soils," *Agriculture*, vol. 10, p. 155, doi: 10.3390/agriculture10050155, 2020.
23. D. Gomez, J. Salvador, J. Sanz-Justo, and J.-L. Casanova, "Potato yield prediction using machine learning techniques and Sentinel 2 data," *Remote Sensing*, vol. 11, p. 1745, doi: 10.3390/rs11151745, 2019.
24. J. Kurek, G. Niedbała, T. Wojciechowski, B. Swiderski, I. Antoniuk, M. Piekutowska, M. Kruk, and K. Bobran, "Prediction of potato (*Solanum tuberosum* L.) yield based on machine learning methods," *Agriculture*, vol. 13, p. 2259, doi: 10.3390/agriculture13122259, 2023.

© Daffodil International University

25. A. Kumar and J. Kaur, “Machine learning and deep learning for soil analysis and classification of micro and macro nutrients using IoT,” *Preprint*, doi: 10.21203/rs.3.rs-4016181/v1, 2024
26. S. Chaterji, N. DeLay, J. Evans, N. Mosier, B. Engel, D. Buckmaster, M. Ladisch, and R. Chandra, “LATTICE: Machine learning, data engineering, and policy considerations for digital agriculture at scale,” *IEEE Open Journal of the Computer Society*, pp. 1-1, doi: 10.1109/OJCS.2021.3085846, 2020.
27. S. Mourtzinis, P. Esker, J. Specht, and S. Conley, “Advancing agricultural research using machine learning algorithms,” *Scientific Reports*, vol. 11, article 97380, doi: 10.1038/s41598-021-97380-7, 2021.
28. R. Akhter and S. Sofi, “Precision agriculture using IoT data analytics and machine learning,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, 2021, doi: 10.1016/j.jksuci.2021.05.013, 2021.
29. S. Kadam, V. Dhede, V. Khan, A. Raj, and D. Takale, “Machine learning methods for automatic potato disease detection,” *NeuroQuantology*, vol. 20, pp. 2102-2106, doi: 10.48047/NQ.2022.20.16. NQ880300, 2022.
30. V. Patil, “Soil analysis and crop prediction system,” *International Journal of Scientific Research in Engineering and Management*, vol. 08, pp. 1-5, doi: 10.55041/IJSREM34498, 2024.
31. A. Zakharov and E. Murzaev, “Impact of inter-row cultivation on potato tuber yields in organic farming,” *Agricultural Machinery and Technologies*, vol. 18, pp. 74-80, doi: 10.22314/2073-7599-2024-18-1-74-80, 2024.
32. Ms. Parab, “IoT based smart agriculture using machine learning,” *International Journal of Scientific Research in Engineering and Management*, vol. 08, pp. 1-5, doi: 10.55041/IJSREM35175, 2024.
33. T. Simakova and A. Simakov, “Assessment of soil suitability for potato cultivation,” *Agrarian Bulletin*, vol. 23, pp. 22-33, doi: 10.32417/1997-4868-2024-23-12-22-33, 2024.
34. N. Shershnev, “Justification of the parameters of the ridge at potato cultivation,” *Science in the Central Russia*, pp. 11-16, doi: 10.35887/2305-2538-2023-5-11-16, 2023.
35. B. Sawicka, P. Barbaś, D. Skiba, A. Noaema, and P. Pszczółkowski, “Harnessing soil diversity: Innovative strategies for potato blight management in Central-Eastern Poland,” *Land*, vol. 13, p. 953, doi: 10.3390/land13070953, 2024.
36. Dr. Periyarselvam, “Advancement in soil analysis for tailored tomato cultivation using deep learning & IoT,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, pp. 3685-3693, doi: 10.22214/ijraset.2024.60739, 2024.
37. R. Faria et al., “Models for predicting coffee yield from chemical characteristics of soil and leaves using machine learning,” *Journal of the Science of Food and Agriculture*, vol. 104, no. 1, doi: 10.1002/jsfa.13362, 2024.
38. J. Y. Shim, J. M. Jung, and W. H. Lee, “Evaluation of the suitability of potato cultivation areas in South Korea based on climate and soil conditions,”

© Daffodil International University

Frontiers in Bioscience-Landmark, vol. 27, p. 070, doi: 10.31083/j. fbl2702070, 2022.

39. J. Padarian, B. Minasny, and A. McBratney, "Machine learning and soil sciences: a review aided by machine learning tools," *SOIL*, vol. 6, pp. 35-52, 2020, doi: 10.5194/soil-6-35-2020, 2020.
40. A. Vibhute, K. Kale, and S. Gaikwad, "Machine learning-enabled soil classification for precision agriculture: A study on spectral analysis and soil property determination," *Applied Geomatics*, vol. 16, doi: 10.1007/s12518-023-00546-3, 2024.
41. Z. Mamatkulov, K. Abdivaitov, S. Hennig, and E. Safarov, "Land suitability assessment for cotton cultivation - A case study of Kumkurgan District, Uzbekistan," *International Journal of Geoinformatics*, vol. 18, doi: 10.52939/ijg. v18i1.2111, 2022.
42. M. Nishad, M. Mitu, and O. Jahan, "Predicting and Classifying Potato Leaf Disease using K-means Segmentation Techniques and Deep Learning Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 212, no. 1, pp. 220-229, 2022.
43. M. S. Talukder, R. B. Sulaiman, M. Chowdhury, M. Nipun, and T. Islam, "PotatoPestNet: A CTInceptionV3-RS-Based Neural Network for Accurate Identification of Potato Pests," *Smart Agricultural Technology,* vol. 5, pp. 100297, 2023.
44. D. Li, Y. Miao, S. Gupta, C. Rosen, F. Yuan, C. Wang, W. Li, Y. Huang, C. Li, D. Miao, Y. Gupta, S. Rosen, C. Yuan, and F. Wang, "Improving Potato Yield Prediction by Combining Cultivar Information and UAV Remote Sensing Data Using Machine Learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 13, no. 16, pp. 3322-3333, 2021.
45. S. Adhikari, "Advancements in Agricultural Technology: Vision Transformer-Based Potato Leaf Disease Classification," *J. Soft Comput. Paradigm,* vol. 6, pp. 169-185, doi: 10.36548/jscp.2024.2.005, 2024.
46. R. K. Bania, N. Talukdar, and D. Bora, "Cutting-edge hybrid deep transfer learning approach for potato leaf disease classification," IEEE Access, vol. 17, pp. 111-111, doi:10.5281/zenodo.10727496, 2024.
47. C. Ramasubramanian, G. Prasad, S. G. Sabarmathi, M. Swarnamugi, B. S. Balachandar, and R. Divya, "Deep learning-based optimised CNN model for early detection and classification of potato leaf disease," in *Proceedings of the International Conference on Artificial Intelligence and Evolutionary Computations in Engineering Systems,* pp. 561-570, doi: 10.1007/978-981-99-6702-5_47, 2023.
48. BlueWeave Consulting, "Smart Agriculture Market," *BlueWeave Consulting*. [Online]. Available: https://www.blueweaveconsulting.com/report/smart-agriculture-market. [Accessed: Jul. 30, 2024].
49. Science Agri, "Top 10 World's Biggest Potato Producing Countries," *Science Agri*. [Online]. Available: https://scienceagri.com/top-10-worlds-biggest-potato-producing-countries/. [Accessed: Jul. 30, 2024].

© Daffodil International University