https://doi.org/10.48047/AFJBS.6.12.2024.394-403

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Enhancing Agricultural Productivity: Predicting Crop Yields from Soil Properties with Machine Learning

Manju G[1], Sania Thomas[2] and Binson V A[3*]

[1]Department of Computer Science, Government College, Ambalapuzha, Kerala, India
[2] Department of Computer Science and Engineering, Saintgits College of Engineering, Kerala, India
[3] Department of Electronics Engineering, Saintgits College of Engineering, Kerala, India

Author for Correspondence

[*]E- mail: binsonvsabraham@gmail.com

Article History

Volume 6 Issue 12, 2024
Received: 25 May 2024
Accepted : 25 June 2024
doi:
10.48047/AFJBS.6.12.2024.394-403

Abstract— This research paper introduces a machine learning-based approach for predicting crop yields using soil properties. The study focuses on four major crops: coconut, plantain, ginger, and tapioca, and examines key soil characteristics including pH, organic carbon, potassium, phosphorus, zinc, sulfur, boron, iron, manganese, and copper as classification features. The soil data used for analysis consists of 500 samples collected from the Agricultural Department of Kulashekharapuram Panchayat, Kollam District, Kerala, India. Three machine learning models, namely k-NN, SVM, and neural network, are implemented and compared in terms of accuracy and performance. The results demonstrate that the k-NN model outperforms the other two models, achieving an impressive accuracy of 88.8% using 10-fold cross-validation. This finding emphasizes the significant role of soil characteristics in predicting crop yields and highlights the efficacy of machine learning techniques in leveraging this information to enhance agricultural productivity. The proposed approach holds practical implications for farmers as it provides valuable insights for informed decision-making regarding crop selection and management practices. By utilizing soil variables, farmers can make optimized choices and improve their crop yield by maximizing the use of soil nutrients. This research contributes to the growing body of literature on the application of machine learning in agriculture, specifically in the domain of crop yield prediction. The findings offer valuable guidance for farmers, facilitating improved crop selection and ultimately contributing to enhanced food security.

Keywords -Crop prediction; Machine learning; Agriculture; Soil characteristics; k-NN

1. Introduction

One of the most crucial aspects of a country's development is its capacity to produce food. Agriculture and agricultural output for essential food crops have long been interdependent. However, the accelerating rate of population growth is a significant concern for our society. This has greatly reduced agriculture's potential, especially in terms of land utilization and fertility. In this age of urbanization and globalization, it is unlikely that any more land will be under cultivation, so the focus must be on utilizing what is already available. Accurate prediction of crop cultivars is critical to agriculture. Despite recent advances in agricultural statistics, few studies have investigated crop prediction using historical data [1, 2]. Furthermore, predicting crop cultivation is challenging due to the unrestricted use of fertilizers containing nitrogen, potassium, and micronutrients. Since agricultural input factors vary from region to region, it is difficult to gather this information across vast geographic areas. The enormous datasets that have been obtained can be used for large-scale crop forecasting. Due to the complexity of the issues involved, new machine learning algorithms for farming arable land and exploiting narrow land resources are required [3].

Agriculture is one of Kerala's primary economic sectors, contributing almost 50% of the state's overall income [4]. During the last few years, the state's total area under cultivation has increased by approximately 100,000 hectares. Coconut, plantain, ginger, and tapioca are significant crops in the Indian state of Kerala, both culturally and economically. The following are some of the ways in which these crops are important. Please let me know if you need further corrections or additions.

Coconut: Kerala is one of the leading coconut-producing states in India, and coconut plays an important role in the state's economy. Coconut palms are used for a variety of purposes, including the production of coconut oil, coconut milk, and coconut water. Coconut oil is used in cooking and as a base for many cosmetics and personal care products. Coconut milk is used in curries and other dishes, and coconut water is a popular drink. The coconut husk is also used for making coir, which is used for a variety of purposes, including making ropes, mats, and other products.

Plantain: Plantains are a type of banana that are larger and firmer than the sweet bananas that are more commonly eaten in Western countries. Plantains are an important staple food in Kerala, and are used in a variety of dishes, including curries and snacks. They are also used in traditional Ayurveda medicine for their nutritional and medicinal properties.

Ginger: Ginger is a popular spice in Kerala, and is used in a variety of dishes for its flavor and health benefits. Ginger is believed to have anti-inflammatory and digestive properties, and is used in traditional Ayurveda medicine to treat a variety of ailments. Kerala is one of the leading ginger-producing states in India, and ginger is an important export crop for the state.

Tapioca: Tapioca, also known as cassava, is a root vegetable that is widely cultivated in Kerala. Tapioca is an important source of carbohydrates in Kerala, and is used in a variety of dishes, including stews, curries, and snacks. Overall, these crops are important for both their cultural significance and economic value in Kerala, and are widely used in the state's cuisine and traditional medicine practices.

Machine learning has proven to be a powerful tool in the agricultural field, particularly in crop prediction. The goal of crop prediction is to estimate crop yields based on environmental factors such as climate, soil type, and pest prevalence. With the help of machine learning algorithms, it is possible to make more accurate predictions about crop yields, which can aid in decision-making for farmers and other agricultural stakeholders. One common approach to crop prediction using machine learning is to use historical data to train a model. The model is then used to predict future crop yields based on current environmental conditions. The accuracy of the predictions is dependent on the quality and quantity of data used to train the model. Another approach is to use sensor data to continuously monitor crops in real-time. Sensors can be used to collect data on factors such as temperature, humidity, soil moisture, and nutrient levels [5]. This data can be used to train machine learning models that can make predictions about crop yields in real-time. This approach can enable farmers to respond quickly to changes in environmental conditions and adjust their management practices accordingly. There are several machine learning algorithms that can be used for crop prediction, including decision trees, support vector machines, neural networks, and random forests. Each algorithm has its strengths and weaknesses, and the choice of algorithm depends on the specific problem at hand. Overall, machine learning has the potential to revolutionize agriculture by enabling more accurate crop predictions and better decision-making for farmers and other stakeholders [6, 7]. With the increasing availability of data and the development of new algorithms, machine learning is likely to play an increasingly important role in the agricultural industry in the years to come.

Crop yield prediction is a critical aspect of agriculture, as it enables farmers to make informed decisions regarding crop selection and management practices. Soil properties, such as pH, organic carbon, potassium, phosphorus, zinc, sulfur, boron, iron, manganese, and copper, play a crucial role in determining crop yield. While there have been numerous studies on predicting crop yields using soil properties, there is still a need for accurate and reliable methods that can account for the complexity and variability of soil and environmental conditions. In this study, we aim to develop a machine learning-based approach for predicting crop yields from soil properties using a self-taken dataset. The novelty of this research lies in the use of a self-taken dataset, which allows for more accurate and reliable predictions based on specific soil and environmental conditions.

The use of a self-taken dataset is a novel approach to crop yield prediction. This approach enables us to collect and analyze data that is specific to our target region and crops, which can improve the accuracy and reliability of our predictions. Additionally, the use of machine learning algorithms can effectively leverage this information to enhance agricultural productivity. The proposed approach has practical implications for farmers, as it can assist them in making informed decisions regarding crop selection and management practices, which can ultimately improve crop yields and food security. Overall, this research contributes to the growing body of literature on the application of machine learning in agriculture and highlights its potential to improve crop yields and food security.

This study focuses on using supervised learning classification algorithms to classify the crops based on soil properties such as potential for hydrogen, organic carbon, potassium, phosphorus, zinc, sulfur, boron, iron, manganese, and copper. The main contribution of this work is to identify the best classification technique that can be used to predict the best crop to cultivate based on soil variables. This research can help farmers make informed decisions on crop selection and improve their crop yield by optimizing the use of soil nutrients.

2. Materials and Methods

A. *Data Collection*

This work utilized an agricultural dataset that mainly included soil characteristics collected from the Agricultural Department of Kulashekharapuram Panchayat, Kollam District, Kerala, India. Agriculture is an important sector of its economy. The region is known for its fertile land, favorable climate, and abundant rainfall, which make it suitable for cultivating a wide range of crops. The dataset contains 500 instances (given in Table I), 10 attributes that defines soil characteristics. The multiclass representation with 4 classes and 10 characteristics is the intended class. The wards surrounding Kulashekarapuram panchayat are where the soil for the study was collected. The soil characteristics used to forecast crops are shown in Table II. The fundamental characteristics that determine the soil's yield are those that are taken into account here.

The primary factor in farming is pH. The breakdown of plant and animal remains, root exudates, living and dead microbes, and soil biota releases organic carbon (OC) into the soil. Phosphorus promotes early root and plant growth, speeds up maturity, and aids in the transport of solar energy to plants. Potassium helps plants create and move carbohydrates, sugars, and oils as well as promote plant vigor and disease resistance. Fruit quality can also be improved by it. Sulfur is a component of the amino acids that make up plant proteins and is used by plants to produce energy. A hormone produced by plants that promotes leaf and stem growth is aided by zinc. In rapidly expanding tissue, boron aids in cell wall development. A deficiency limits calcium uptake and prevents plants from utilizing it. Several substances that control and encourage growth contain iron as a component. Photosynthesis benefits from manganese. Plant enzymes require copper as a key component [8].

Soil parameters play a crucial role in determining crop yield. Different crops have different requirements in terms of soil characteristics such as pH, nutrient availability, water retention, and organic matter content. Understanding the soil properties and their impact on plant growth can help farmers make informed decisions about crop selection and management practices. By analyzing soil parameters and their impact on crop yield, farmers can adjust their management practices accordingly, such as adjusting irrigation schedules, adding fertilizers, or choosing crop varieties that are better suited to the specific soil conditions. This can lead to higher crop yields, increased profitability, and sustainable farming practices.

Table 1: Data instance

Sl.No	Crop	Total data
1	coconut	140
2	plantain	136
3	ginger	114
4	ginger	108

Table 1: Soil Attributes

Sl. No	Attributes
1	pH (potential of Hydrogen)
2	OC (Organic Carbon)
3	P (Phosphorus)
4	K(Potassium)
5	S (Sulphur)
6	Z (Zinc)
7	B (Boron)
8	Fe (Iron)
9	Mn (Manganese)
10	Cu (Copper)

B. *Machine Learning Algorithms*

Machine learning has potential to revolutionize the agriculture industry by enabling farmers to make better decisions based on real-time data. It can be trained to predict crop yields based on factors based on nutrient level of soil. It can help farmers plan their planting and harvesting schedules and optimize their use of resources [9-11].

K-Nearest Neighbors (KNN) is a simple and popular machine learning algorithm used for classification and regression tasks. In KNN, the output is classified based on the k-nearest data points to the input sample. In the case of classification, the class of the input sample is determined by the most frequently occurring class among its k nearest neighbors. In regression, the output value is determined by the average of the k nearest neighbors' output values [12]. KNN is a non-parametric algorithm, meaning that it doesn't make any assumptions about the underlying data distribution. The algorithm is flexible, easy to implement, and suitable for small to medium-sized datasets [13]. However, KNN can be computationally expensive and may not perform well on high-dimensional data. It is also sensitive to irrelevant features and outliers in the data. To address these issues, techniques like feature selection and data normalization can be used.

KNN is a popular machine learning algorithm used in various applications, including agriculture. KNN is a non-parametric and lazy learning algorithm that can be used for classification and regression tasks. In agriculture, KNN is mainly used for crop yield prediction by analyzing soil properties. KNN is used in agriculture for predicting crop yield based on soil parameters such as pH, organic matter, moisture content, and nutrient levels. KNN algorithm takes in a training set of soil samples with corresponding crop yield data and classifies the new soil samples based on the K nearest neighbors to the input sample. The K value refers to the number of nearest neighbors that the algorithm should consider. KNN algorithm can be particularly useful in agriculture because it does not make any assumptions about the distribution of the data and can capture non-linear relationships between soil parameters and crop yield. Additionally, KNN is a simple algorithm and easy to implement, making it accessible to farmers who may not have a technical background in machine learning.

Support Vector Machines (SVM) is a popular machine learning algorithm used for classification and regression problems. In SVM, the algorithm tries to find a hyper plane that separates the classes in the feature space with the largest margin possible. The hyper plane is defined as a decision boundary that separates the data into different classes [14]. SVM is particularly useful when the data has a clear separation between classes, but it can also be used for non-linear classification tasks by using kernel functions that transform the input space into a higher-dimensional space.

SVM has several advantages over other machine learning algorithms, such as being less prone to over fitting, having a strong theoretical foundation, and being effective even in high-dimensional spaces. However, it can be computationally intensive, and selecting the appropriate kernel function and hyper parameters can be challenging [15].

SVMs are widely used in agricultural applications, including crop yield prediction. In crop yield prediction, SVMs can be used to build a model that can predict crop yield based on various factors such as soil parameters, weather conditions, and agricultural practices. SVMs work by finding the best boundary that separates data into different classes, or in regression tasks, by finding the best fit line that can predict the outcome variable. SVMs have been used successfully in crop yield prediction studies, such as predicting soybean yield based on soil properties and weather variables. SVMs have also been used to predict wheat yield based on weather and soil variables. In both cases, SVMs were found to be effective in predicting crop yield. One of the advantages of using SVMs in crop yield prediction is their ability to handle high-dimensional data, making them suitable for datasets that contain a large number of variables.

A neural network is a type of machine learning model that is inspired by the structure and function of the human brain. It consists of a large number of interconnected nodes, or artificial neurons, that process information by performing mathematical operations on input data. Each neuron takes in input from multiple other neurons, processes it using a weighted function, and produces an output that is transmitted to other neurons in the network [16]. The weights on each connection between neurons are adjusted during training, allowing the network to learn and make predictions on new data. Neural networks have been successfully applied in a wide range of applications, including image and speech recognition, natural language processing, and predictive modeling in various fields including agriculture.

Neural networks have been successfully applied to agricultural applications, including crop yield prediction. In this context, neural networks are used as a supervised learning method to build a model that can predict crop yield based on a set of input features such as weather data, soil characteristics, and farming practices. In crop yield prediction, a neural network model takes in a set of input variables such as soil parameters, weather conditions, and agricultural practices, and produces an output variable that represents the predicted yield of a particular crop. The model is trained on a dataset that includes historical data on crop yields and the corresponding input variables.

The neural network model consists of layers of interconnected nodes (neurons) that process and transform the input data through a series of mathematical operations. The output of each neuron is calculated by applying a non-linear function to the sum of the weighted inputs. These weights are learned during the training process using an optimization algorithm that minimizes the difference between the predicted output and the actual output in the training data. The advantages of using neural networks for crop yield prediction include their ability to handle complex, non-linear relationships between the input variables and the output variable. They are also capable of handling high-dimensional input data, which is often the case in agricultural applications. However, the training process of neural networks can be computationally intensive, and the resulting model may be difficult to interpret. Therefore, it is important to carefully select the input variables and design the network architecture to achieve the best performance. All these algorithms have extensively used in agricultural applications and have shown better performances [17-20].

C. *Matrices*

In machine learning classification, it's important to evaluate the performance of the model. Accuracy, recall, specificity, and precision are the commonly used evaluation metrics in classification models. When evaluating a classification model, it's important to consider all of these metrics to get a better understanding of its performance.

Accuracy

The accuracy is calculated as the number of right guesses divided by the total number of forecasts.

Accuracy =$(TP + TN)/(TP + TN + FP + FN)$

where, TP- True Positive
TN- True Negative
FP- False Positive, and
FN- False Negative

Precision

The projected positive presence that actually is positive is precision . The term "positive predicted value" is sometimes used (PPV).

Precision = $TP/(TP + FP)$

Recall

The proportion of positive results out of the number of samples which were actually positive. It is also known as Sensitivity.

Recall = $TP/(TP + FN)$

Specificity

The proportion of negative results out of the number of samples which were actually negative.

Specificity = $TN/(TN + FP)$

3. Results and Discussion

The right attributes are chosen to discover precise soil factors for projecting a suitable crop for more effective farming. Machine learning algorithms are used to determine which crop would be best for a specific area of land. Following that, the procedures are assessed using criteria including accuracy, precision, recall, and specificity.

D. *Performance comparison for different crops in k fold cross validation*

Table III shows a performance evaluation of ML algorithms like KNN, SVM and neural network based on soil characteristics such as the Ph, OC, N, P, B, Fe, Mn , Mg, S and K during cross validation using MATLAB. Figure 1 show the graphical representation of accuracy level of different models. In K fold cross validation the dataset is split into K number of folds and is used to evaluate the models ability when given a new data. Here we compare the data set with 10 fold, 5 fold, 3 fold cross validation. From the results of the performance matrices like accuracy, precision, recall and specificity it is evident that the all models working best with 10 fold cross validation. Among them, KNN shows maximum accuracy of 88.8 %, precision of 89.1%, recall of 88.8%, and specificity of 96.2%.

The figure 2 shows the confusion matrix of KNN in 10 fold cross validation with 500 data set of 4 crops with 10 different attributes. This model outperformed other models during the validation. The confusion matrix is a table that depicts the performance of the model. It summarizes the overall execution of the machine learning algorithms used. The confusion matrix can be used to calculate various performance metrics of the classification model, such as accuracy, precision, recall, and specificity. A confusion matrix in a 4-class classification problem is similar to the one in a binary classification problem, but with additional categories for each class as shown in figure 2.

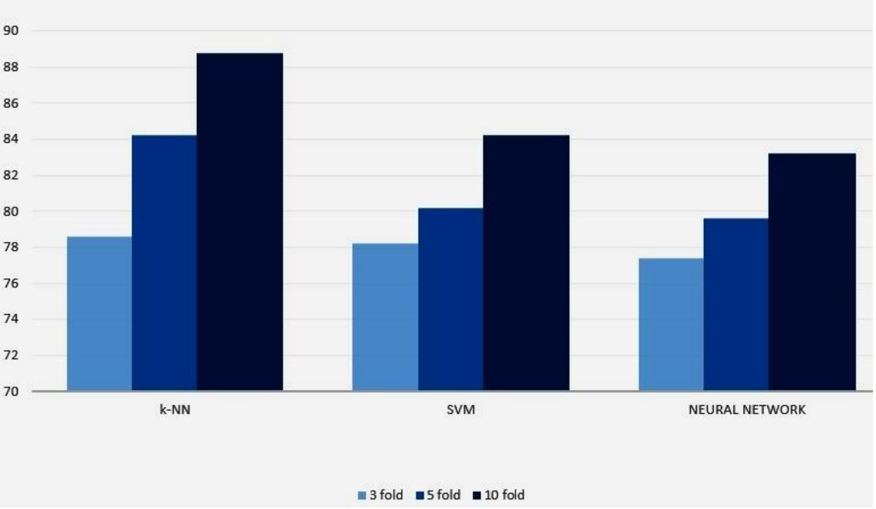

Fig.1. Performance analysis graphical representation

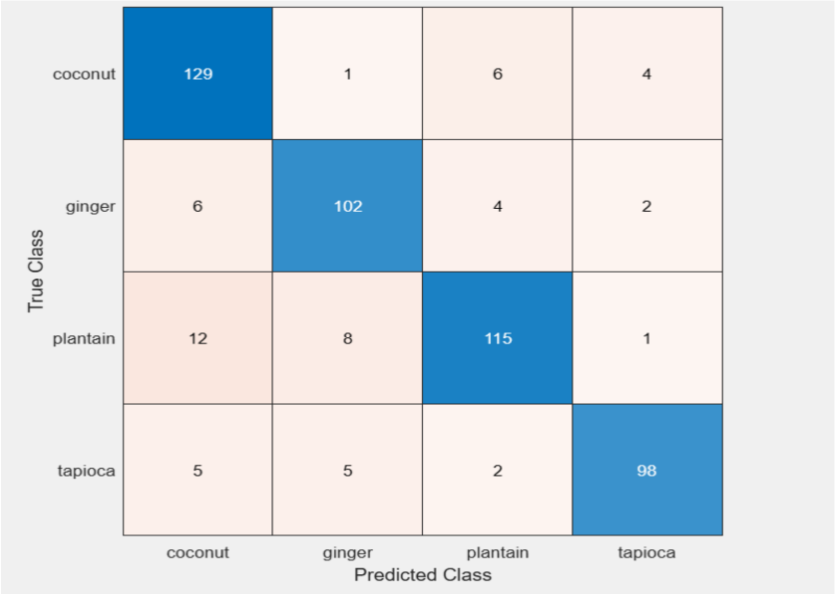

Fig.2. Confusion matrix for KNN model in 10 fold cross validation

Table 3: Performance comparison for different crops in k fold cross validation

K-Fold	**3-Fold**			**5- Fold**			**10- Fold**		
Model	k-NN	SVM	Neural Network	k-NN	SVM	Neural Network	k-NN	SVM	Neural Network
Accuracy	78.6	78.2	77.4	84.2	80.2	79.6	88.8	84.2	83.2
Precision	78.8	82.4	77.4	84	82.9	79.4	89.1	86.9	83.5
Recall	78.6	77.6	77.2	84.3	79.8	79.5	88.8	83.8	81.4
Specificity	92.6	92.4	92.6	95.8	93.2	93.2	96.2	94.5	94.3

Figure 3 depicts the K-NN model's ROC (Receiver Operating Characteristic) curve. An illustration of a binary classifier system's performance is called a 1ROC curve. For various threshold values, it plots the True Positive Rate (TPR) versus the False Positive Rate (FPR).

The ratio of actual positives that are correctly classified as such is known as the TPR, whereas the ratio of actual negatives that are wrongly classified as positives is known as the FPR. The ROC curve can be used to assess the trade-off between a classifier's TPR and FPR and to see how effectively it can distinguish between positive and negative classes. A random classifier will have a ROC curve that is a diagonal line from the lower left corner to the upper right corner, whereas a perfect classifier will have a ROC curve that passes through the upper left corner (TPR=1 and FPR=0).

A typical statistic for assessing a classifier's performance is the area under the ROC curve (AUC). The classifier operates perfectly when the AUC is 1.0, whereas an AUC of 0.5 shows that it performs no better than random guessing. Here, the K-NN model's AUC is more than.95 across all classes.

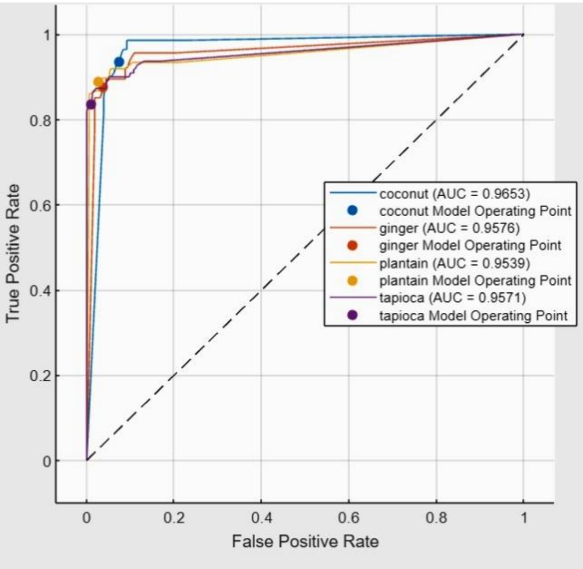

Fig.3. K-NN model's ROC curve

4. Conclusions

Crop cultivation used to be carried out based on farmers' actual experience. However, alterations in the soil's fertility have already begun to negatively impact agricultural yield. Effective productivity of the crop is influenced by the soil fertility and environmental conditions. Agriculture must include the forecast of the best crop to grow, and machine learning algorithms are crucial to this prediction. In this paper, it analyses 3 different machine learning models for 4 different crops. The k-NN model performs the best with 10 fold cross validation and it shows 88.8 % accuracy.

References

[1] Pudumalar, S., Ramanujam, E., Rajashree, R. H., Kavya, C., Kiruthika, T., & Nisha, J. (2017, January). Crop recommendation system for precision agriculture. In 2016 eighth international conference on advanced computing (ICoAC) (pp. 32-36). IEEE.

[2] Suruliandi, A., Mariammal, G., & Raja, S. P. (2021). Crop prediction based on soil and environmental characteristics using feature selection techniques. Mathematical and Computer Modelling of Dynamical Systems, 27(1), 117-140.

[3] Prathibha, S. R., Hongal, A., & Jyothi, M. P. (2017, March). IoT based monitoring system in smart agriculture. In 2017 international conference on recent advances in electronics and communication technology (ICRAECT) (pp. 81-84). IEEE.

[4] Sethi, A., Lin, C. Y., Madhavan, I., Davis, M., Alexander, P., Eddleston, M., & Chang, S. S. (2022). Impact of regional bans of highly hazardous pesticides on agricultural yields: the case of Kerala. Agriculture & Food Security, 11(1), 9.

[5] Thomas, S., & Thomas, J. (2022). Non-destructive silkworm pupa gender classification with X-ray images using ensemble learning. Artificial Intelligence in Agriculture, 6, 100-110.

[6] Gandhi, N., Armstrong, L. J., Petkar, O., & Tripathy, A. K. (2016, July). Rice crop yield prediction in India using support vector machines. In 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-5). IEEE.

[7] Binson, V. A., & Thomas, S. (2023). The Development of a Mobile E-Nose System for Real-Time Beef Quality Monitoring and Spoilage Detection. Engineering Proceedings, 56(1), 256.

[8] Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. Computers and electronics in agriculture, 121, 57-65.

[9] Chaganti, S. Y., Ainapur, P., Singh, M., & Oktaviana, S. (2019, September). Prediction based smart farming. In 2019 2nd international conference of computer and informatics engineering (IC2IE) (pp. 204-209). IEEE.

[10] Binson, V. A., Subramoniam, M., & Mathew, L. (2024). Prediction of lung cancer with a sensor array based e-nose system using machine learning methods. Microsystem Technologies, 1-14.

[11] Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on knowledge and data engineering, 17(4), 491-502.

[12] VA, B., Mathew, P., Thomas, S., & Mathew, L. (2024). Detection of lung cancer and stages via breath analysis using a self-made electronic nose device. Expert Review of Molecular Diagnostics, 1-13.

[13] Binson, V. A., & Thomas, S. (2024). Unveiling the Smell of Health: E-Nose-Based Volatile Organic Compound Analysis of Exhaled Breath in Early Lung Cancer Detection. In Proceedings (Vol. 100, No. 1, p. 23). MDPI.

[14] Binson, V. A., Thomas, S., Philip, P. C., Thomas, A., & Pillai, P. (2023, November). Detection of Early Lung Cancer Cases in Patients with COPD Using eNose Technology: A Promising Non-Invasive Approach. In 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE) (pp. 1-4). IEEE.

[15] Binson, V. A., Thomas, S., Subramoniam, M., Arun, J., Naveen, S., & Madhu, S. (2024). A Review of Machine Learning Algorithms for Biomedical Applications. Annals of Biomedical Engineering, 1-25.

[16] Zala, D. H., & Chaudhri, M. B. (2018). Review on use of BAGGING technique in agriculture crop yield prediction. International Journal for Scientific Research & Development, 6(8), 675-7.

[17] Binson, V. A., George, M. M., Sibichan, M. A., Raj, M., & Prasad, K. (2023, January). Freshness evaluation of beef using MOS based E-Nose. In 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) (pp. 792-797). IEEE.

[18] Thomas, S., & Thomas, J. (2024). Nondestructive and cost-effective silkworm, Bombyx mori (Lepidoptera: Bombycidae) cocoon sex classification using machine learning. International Journal of Tropical Insect Science, 1-13.

[19] Raman, R., Kantari, H., Gokhale, A. A., Elangovan, K., Meenakshi, B., & Srinivasan, S. (2024). Agriculture Yield Estimation Using Machine Learning Algorithms. In 2024 International Conference on Automation and Computation (AUTOCOM) (pp. 187-191). IEEE.

[20] Thomas, S., & Thomas, J. (2024). An optimized method for mulberry silkworm, Bombyx mori (Bombycidae:Lepidoptera) sex classification using TLBPSGA-RFEXGBoost. Biology Open, 1-11.