RESEARCH ARTICLE OPEN ACCESS

Soil Properties Monitoring and Crop Type Prediction Using IOT and Machine Learning Case Study: Rwanda-Kicukiro Sector

Gasana james madson [1], Habumugisha fulgence [2], Uwimana Josephine [3], Munyaneza midas adolphe [4], Uwantege stellah prossy [5], Niyonsaba maximilien [6]

[1] African Center of Excellence in Internet of Things_College of Science and Technology,University of Rwanda..
[2] University of Rwanda-College of Science and Technology. [3],[4],[5],[6] Rwanda Polytechnic-Integrated Polytechnic Regional College(IPRC) Kigali - Rwanda.

ABSTRACT

In Rwanda and all around the world agriculture plays an important role for living of human beings for survival and economic growth. Different research shows that as population increases the area of cultivation that is used for habitation and need of food is increasing while farming land is reducing.Smart farming is needed to maintain and secure food availability for this generation and future by using IoT.The overall objective of this paper is to monitor soil properties by providing real time data of a given land to farmers and decision makers like RAB for proper use and productivity of land and predicting the appropriate croptype in that land measured.Monitoring soil properties helps to monitor degradation of land and know when and how much fertilizer needed in that land based to the croptype you want to plant meanwhile system will keep showing the croptype based to whatever fertilizers available. In this paper,the Descriptive quantitative research approach were used by establishing the relationship between independent variables(NPK,temperature,humidity and PH) and dependent variable(croptype like cereals,legumes,vegetables) where the croptype will depend to the soil properties available in the land.The study provides the way of knowing the current soil properties and appropriate croptype.With this system a farmer can monitor the land nutrients in the given land and can be able to add only needed fertilizers.Again in this system there is model to predict the croptype of a given land based to its soil properties.The data was collected using sensors NPK sensor,PH sensor,temperature sensor,humidity sensor.Data from sensors are sent to microcontroller through direct connection and then sent from microcontroller to the database hosted on the cloud via gprs module and also the realtime data visible on LCD. Data was analysed by using machine learning algorithm known as classification.Among three common types of classification logistic regression,decision tree classifier and knearest neighbor(KNN),we chosed to use decision tree classifier because it provided higher accuracy(99%) than others. The main contribution of this work is that, a farmer can know the status soil properties of land by using IoT sensors instead of tradition means of using laboratory which is expensive,time consuming and tedious,database hosted on the cloud helps the decision makers to visualize and monitor changing rate of soil properties of a given which can help to mitigate the land use in future and finally we can say using machine learning in prediction of best fitting croptype in any given land is a great contribution.

Keywords: - IOT, NPK sensors, soil properties, monitoring, Machine Learning, Algorithm

I. INTRODUCTION

Crop production is the main focus nowadays in the whole world and due to rapid growth of population worldwide. Fertile land to cultivate is not increasing instead it is reducing due to the fact that as people are increasing they need land for shelter and other facilities. With the mentioned above needs there is also issue of land degradation and climate change, therefore all these put pressure on governments and researchers to find solution of how to keep feeding the population, since the farming land remain constant, thus after sometime it loses its natural fertility and due to those chemicals/organic manure is needed to restore soil fertility and these chemicals needs to be monitored and regulated and added adequately. This paper will contribute to the proper soil properties data storage and monitoring.

The Internet of Things (IoT) and machine learning has become increasingly attractive to industry and emerges to the public for daily life. Its popularity is largely derived from the impression/concept of widely available connection and easily accessible interaction network. In the IoT world, better decisions can be made automatically without people's involvement or less if any. Nowadays, people can be able to obtain information about anything anytime anywhere through the identification of "Things" regardless of its nature. In this field of IoT, it is anticipated that since network will be capable of collecting and transmitting data. Thereafter, data is automatically analyzed and it brings optimization of the solutions to different circumstances experienced in any field of life [1].

IoT in Agriculture, based on the report of Food and Agriculture Organization of the United Nations(FAO) predicts that the global population will reach 8 billion people by 2025 and 9.6 billion by 2050 [2].This means that there is need of modernization in agriculture to be able to feed this population. It is in this regard that precision agriculture matters most [3].

Since Precision Agriculture (PA) aims to optimize and improve agricultural processes to ensure maximum product and to achieve this, there are many ways and we will focus on soil and its properties to the contribution of yield needed as mentioned above. Soil is said to be fundamental of all as long as you are talking about Agriculture Therefore, IoT is the best way of monitoring soil environment (here referred as soil properties) and alerting the users (agriculturalists) on real-time to take necessary actions. It is in this regard that this paper will focus on monitoring soil properties using IoT and we know the quality and quantity of those selected properties towards predicting the most appropriate plant on that soil. To reach the expected goal, we need sensors, embedded system, database hosted on cloud for storage of data for future use, algorithms where machine learning is preferred.

II. LITERATURE REVIEW

In literature, we discuss paper related works and gaps they highlighted while relating the contribution of this paper.

A. Related Works

Smart system to support agricultural parameters using IoT [4].This system is explaining the best use of water in irrigation system based to the soil parameters like PH, Temperature, soil humidity and appropriate crops in a specific field. In the future work of this paper they suggested to use some more sensors like humidity sensor, carbon dioxide concentration sensors with more data analysis.

System mechanism to restore soil quality using biochar and brown coal waste. This article is about how to apply soil organic matter and brown coal waste to restore soil quality; they proposed future studies to focus on the long-term application these products and locally sourced alternatives [5].

Agriculture conservation and integrate soil fertility management system for the maize production in the sub-humid and semi-arid regions of Kenya [6]. This paper is about soil and water management with integration of soil fertility management to enhance maize crop production, in their work they recommended to study the effects of both management practices on soil humidity during the long rain periods in the two regions.

Optimization of monitoring farming conditions using IoT [7].It is about better use of technology in farming to save resources (water and fertilizers) and increase yield, in their recommendation they emphases on IT infrastructures which must be in rural areas and its still challenge in under developing countries.

Soil biodiversity, fertility and function in a heathland restoration is affected by Long-term acidification of pH neutral grasslands [8].
In their paper they explained how field scale experimental with Sulphur affected soil biodiversity, fertility, and function over a period of time (around 17 years). In this paper they recommended further work which is needed on application rates and timing as well as integration with grazing management.

Managing soil fertility in wheat-maize yield [9]. This article is talking about the use of organic fertilizers and recycling of crop residues as very important and they compared different chemicals. They recommended to keep regulating the application of chemicals in the soil for environmental protection also.

Current research on IOT role in Agriculture are being conducted and it is shown that agriculture is a big field and needs more attention in order to maximize land productivity. Among many research including the given ones in literature review, there is a general gap to bridge like database to host data for future analytics, creating model for prediction using machine because most of related work are talking about a situation for seeing individual situation like irrigation, monitoring one plant type like maize or tomatoes and applications are based on real-time data. Again RAB (Rwanda Agriculture Board) was visited during data collection and we met with staff in soil department to facilitate us to get reference so that we can base our work on real facts, they provided reference properties with some crops but no database, no automated system to collect data and store data. The main contribution of this work is that it has database hosted on the cloud where by a farmer can put a prototype (device) in garden with a certain crop and keep monitoring soil properties changes and this changes helps to know the rate of reduction of fertilizers and be able to add only needed fertilizer and if the farmer is not able to add fertilizer the system will suggest him the best appropriate crop can match with already available soil properties based to the prediction model.

B. SYSTEM COMPONENTS

PH sensor. This sensor is inserted in 50 cm deep the soil to capture PH of soil.
Temperature sensor. This sensor is inserted in 50 cm deep the soil to capture temperature of soil.
Humidity sensor. This sensor is inserted in 50 cm deep the soil to capture humidity of soil.
NPK sensor. This sensor captures three (3) parameters namely Nitrogen, Phosphorus and Potassium.
Arduino uno. This microcontroller is the heart of the system, it is programmed using Arduino IDE. The data from sensors are sent to it and it processes them due to the codes and send data to the cloud based database for storage and analytics.
SIM 800L.This is GSM module which provides internet connection through cellular network.it has simcard slot. This gsm uses gprs capability and it is chosen because of its wide use country wide than any network infrastructure you can think, it is also cheap in terms of internet bundles and it is best fitting in agriculture field where my work belongs.
LCD. This displays real time data on the field
Rechargeable battery. This helps to supply the system in the field
Arduino IDE. This software is used in Arduino uno by receiving live data from sensors and due to codes as programmed there is data cleaning to remove errors or nulls before sending to the cloud.
Database management system. This database is hosted on cloud to receive data from sensors by using gprs module for internet connectivity and the data stored is used for

training model and testing and finally is used for further analysis in case they are needed.

Cloud application: Due to IOT era, we will use cloud for storage and future use of data for further analysis

Machine learning algorithm. This is the most part needed to analyze captured and collected data for predictions. In the next chapter we provide details of chosen and fitting algorithm due to the type of data and expected output.

III. SYSTEM DESCRIPTION

a. system Architecture

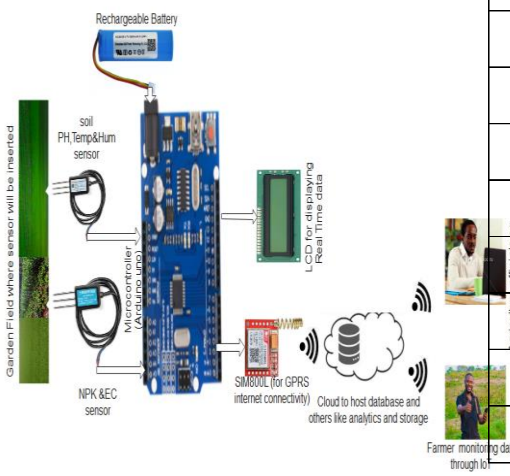

Fig. 1 Architectural Design of the proposed system

In this system, sensors are in the field (garden to use) and they collect the specified data in explanations given that sensors are connected directly to Arduino and SIM 800L and the data are processed and sent to database which is hosted online.

For Arduino and SIM 800L to send data to database, there is a need of internet connection. The data captured by sensor is displayed on LCD at the same time, the data is sent to the database and stored for prediction and for further use, For other explanations on how data from sensors are trained by model using classification regression that are used to generate the nearest crop type is detailed in the next chapter.

ID	N	P	K	Temperature	Humidity	PH	Croptype
1	90	42	43	20.88	82	6.5	cereal
2	85	58	41	21.77	80.32	7.04	cereal
3	60	55	44	23	82.32	7.84	cereal
4	74	35	40	26.49	80.16	6.98	cereal
5	40	72	77	17.02	16.99	7.49	legume
6	23	72	84	19.02	17.13	6.92	legume
7	39	58	85	17.89	15.41	6	legume
8	22	72	85	18.87	15.66	6.39	legume
9	105	95	50	27.33	83.68	5.85	fruits
10	108	92	53	27.4	82.96	6.28	fruits
11	86	76	54	29.32	80.12	5.93	fruits
12	80	77	49	26.05	79.4	5.52	fruits
13	61	68	50	35.21	91.5	6.79	vegetables
14	58	46	45	42.39	90.79	6.58	vegetables
15	45	47	55	38.42	91.14	6.75	vegetables
16	39	65	53	35.33	92.12	6.56	vegetables

Table I. Sensor Data as Captured from the Field by Taking First 4 Values on Each Crop Type

In the table.1 above, only sixteen (16) data were given but more than 1200 was collected.

b. Algorithm to use

An algorithm is said to be a procedure or formula used to solve a problem. Therefore, we need codes to perform manipulations to achieve the results. More explanations and flow chart will be given in as we progress. Due to the type of output to predict which is categorical instead of numerical values, we found that Decision tree classifier model which is under classification algorithm is best fitting among others we trained like logistic regression and knn.

Classification is the type of supervised machine learning which is mostly fitting in a predictive modelling problem where class or labels are categorical instead of numerical values. I chose classification among others because it best fits in prediction where dependent variable (output) is a label (e.g. cereals).

Machine learning is a method of data analysis that automates analytical model building based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention [10].

Supervised learning is one of the most basic types of machine learning among others like unsupervised learning, in this work we chose supervised learning because we the model is trained using known labelled datasets while in unsupervised learning inputs are provided to the model [11].

c. system Flowchart

This flowchart shows the flow of information from step one to the last in summary. Each step is detailed in next chapter

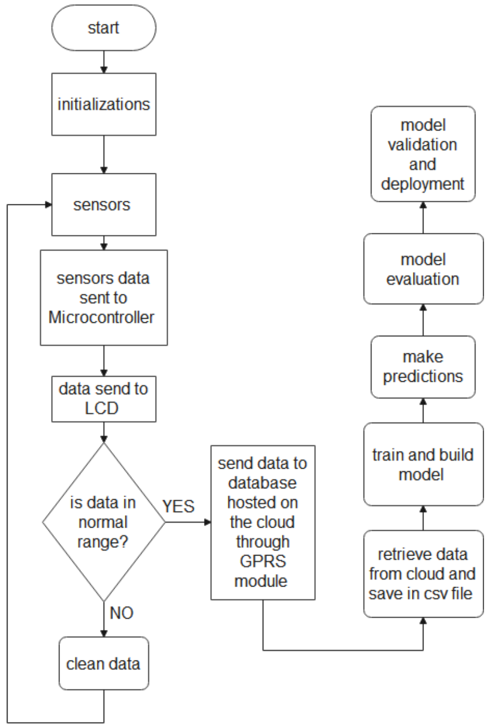

Fig.2 Flowchart of the proposed system algorithm

d. Dataset Cleaning/Pre-Processing

There are different ways of python editor and chose to use jupyter note book which is user friendly and has many features.

```
# This scattering graph shows how features/independent variables are scatterd.here I used one but
#you can check on all features in any combination

sns.FacetGrid(plants, hue="croptype", height=10).map(plt.scatter, "temperature", "ph").add_legend()
```

<seaborn.axisgrid.FacetGrid at 0x1dd8a5021f0>

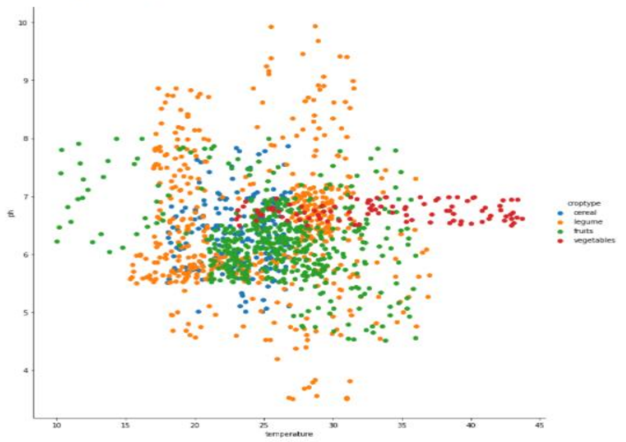

Fig.3 Dataset scattering distribution based to dependent variables

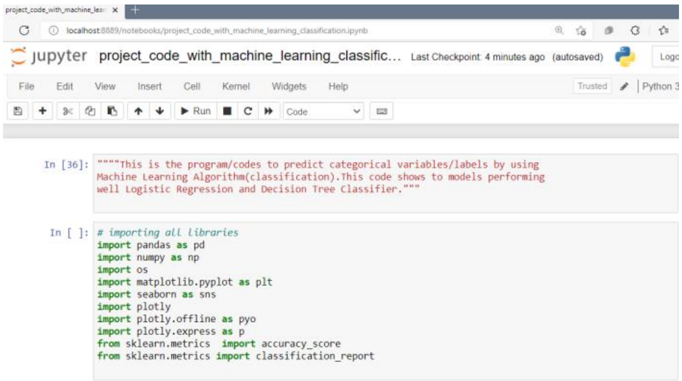

Fig.4 Main functions needed to visualize and manipulate our dataset

```
#Load the data to use for training and testing,df means Dataframe
df=pd.read_csv('testingdata.csv')
df
```

	ID	N	P	K	temperature	humidity	ph	croptype
0	1	90	42	43	20.88	82.00	6.50	cereal
1	2	85	58	41	21.77	80.32	7.04	cereal
2	3	60	55	44	23.00	82.32	7.84	cereal
3	4	74	35	40	26.49	80.16	6.98	cereal
4	5	78	42	42	20.13	81.60	7.63	cereal
...	...	...	...	...	...	...	...	...
1284	1295	42	59	55	40.10	94.35	6.98	vegetables
1285	1296	43	64	47	38.59	91.58	6.83	vegetables
1286	1297	35	67	49	41.31	91.15	6.62	vegetables
1287	1298	56	59	55	37.04	91.79	6.55	vegetables
1288	1299	39	64	53	23.01	91.07	6.60	vegetables

1289 rows × 8 columns

Fig. 5 Dataframe/dataset to train the model

e. Model training

We have tried different models but here only two are shown.it means after trying many models we have to choose the one with highest accuracy (model validation).

1) Classification model using logistic regression algorithm,

```
# model training

# train=70%   while   test=30%


from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x, y, test_size=0.30, random_state=42)
```

```
# logistic regression.
from sklearn.linear_model import LogisticRegression
LRmodel=LogisticRegression()
LRmodel.fit(x_train, y_train)
```

```
LogisticRegression()
```

```
prediction_LRmodel=LRmodel.predict(x_test)
```

```
# The accuracy of this model is 90.69% which is quit good
accuracy_LRmodel=accuracy_score(y_test,prediction_LRmodel)*100
```

```
print(accuracy_LRmodel)
```

```
90.69767441860465
```

Fig. 6 Prediction Result of Logistic Regression (Which Is 90.69%)

2) Classification model using Decision tree classifier

```
#prediction by using second model which is Decision Tree

from sklearn.tree import  DecisionTreeClassifier
DTmodel = DecisionTreeClassifier()
DTmodel.fit(x_train, y_train)

DecisionTreeClassifier()

prediction_DTmodel=DTmodel.predict(x_test)

accuracy_DTmodel=accuracy_score(y_test,prediction_DTmodel)*100

# This  model od Decision Tree  is more accurate than Logistic Regression with 90.22%
print(accuracy_DTmodel)

99.2248062015504
```

Fig. 7 Prediction Result of Decision Tree Classifier (Which Is 99%)

f. Model validation

This is the process of choosing the model which is achieving the intended output among many models.it means after trying different models you can end up with no reliable model but for our case, we are choosing the Decision tree classifier which has accuracy of 99.2%.

```
# Below shows  two output examples  when you path any random input to the model

x_DT=53,60,55,37,92,6.3

x_DT

(56, 60, 55, 37, 92, 6.7)

x_DT=np.array([[53, 60, 55, 37, 92, 6.3]])
x_DT_prediction=DTmodel.predict(x_DT)

x_DT_prediction

array(['vegetables'], dtype=object)

x_DT=91,42,43,20.88,82,6.7

x_DT

(91, 42, 43, 20.88, 82, 6.7)

x_DT=np.array([[91, 42, 43, 20.88, 82, 6.7]])
x_DT_prediction=DTmodel.predict(x_DT)

x_DT_prediction

array(['cereal'], dtype=object)
```

Fig.8 Prediction Results of Validated Model

IV. RESULT AND ANALYSIS

In this chapter we discuss the outcome of the algorithm and its efficiency and if it can be trusted at what percentage to be realistic on the field.

a) system result

With validation of decision tree classifier which has accuracy of above 99%, we can say that the model is reliable and it can be deployed for use by end users.

```
print(classification_report(y_test,prediction_DTmodel))
```

	precision	recall	f1-score	support
cereal	0.94	1.00	0.97	48
fruits	1.00	0.99	1.00	143
legume	1.00	0.99	0.99	167
vegetables	1.00	1.00	1.00	29
accuracy			0.99	387
macro avg	0.99	1.00	0.99	387
weighted avg	0.99	0.99	0.99	387

Fig.9 Prediction summary each crop type and overall accuracy

This table above shows the precision, recall, score and support if each croptype and overall accuracy. For details we are taking one crop type which is cereal.

Precision indicates percentage of relevant items among retrieved items.it is taken as measure of quality. Precision is equal to true positives over all true positives including those not identified correctly. take an example for cereal, given the test size(support) of 48 which was named true positive ,45 out of 48 which is equal to 0.94 or 94% is well predicted while 13/48 was wrongly pressed named.

Recall indicates the percentage of relevant items that were retrieved. Recall is equal to the true positives over all samples that should be identified as positives. Take an example of cereal, given the support of 48, since it is 100% it means all items retrieved was relevant.is taken as measure of quantity.

F1-score is weighted average of precision and recall (94%+100%)/2=97% as it is in the table.

Support indicates the total samples. For example in the table above the support is mentioned 30% chosen as test data.

Accuracy indicates the number of correctly predicted observations to the total observations. The higher the accuracy the best the model. Since the accuracy of our model is 99%, it is best and it is validated to be used by beneficiaries.

b) Hypotheses and objectives approval

Based on the results, we can approve the hypothesis because prototype was made and it is capturing and displaying real time data and data is being sent to the database which means that a farmer can keep monitoring the reduction of soil nutrients/properties in the garden and be able to add fertilizers only when needed and estimate the quantity and the type of fertilizer missing, again the model was made using Decision tree classifier as machine learning algorithm to predict the best fit crop type on existing parameters.

All objectives was achieved as expected though few challenges was mentioned.

c) Discussion

In this section we discuss the status of the use of IOT in agriculture in general and particularly in Rwanda. IOT was emerged in 1990s and its main practical functions was to expand the theory of Services Sciences, Management

Engineering (SSME) and this lead it to infiltrate in all walks of life in different fields. Agriculture is one of the field interested in due to its importance for life.as mentioned in literature review, there are different ways of applying IOT in this field and the main purpose is to increase efficiency means small land, high crop productivity in short time [12],[13].

In agriculture most of applications as related to what is mentioned is about monitoring water, soil fertility of a given crop type and system reaction base on real time data for example when water decreases to a certain level irrigation system starts based to only temperature and moisture for a specific crop example maize or rice. Particularly in Rwanda there are very few application of IOT systems and most of them are based on irrigation and many of them including the mentioned ones in related work none with database to monitor not only real time data but also store on database for long term planning by government and none with predictions using machine learning.

The main issue is that the system (IOT) is new technology in Rwanda and IOT components to be used are very expensive, for example I bought only two sensors at $300 which means the beneficiaries cannot afford to pay easily. The other issue is about few knowledgeable people in the field but thanks to the government that placed ACEIOT to solve such problems and soon progress will occur.

This work will contribute more in academic field and in the Rwandan community on the level of farmers as well as on the level of

decision makers like RAB since it combines collecting many (six) essential soil properties needed by a plant to grow and save those data on database hosted on the cloud and provide a model for predictions of best use of land using machine learning.

V. CONCLUSION

In this work we have presented the results of IoT based monitoring system of soil properties and crop type prediction using machine learning algorithm. The prototype allows real time measurement of data and storing those data on cloud database which was used in training and testing model.

The most challenge faced was lack of sensors on time because we did not find them on local market of Rwanda and we needed to order from outside. Again the issue of calibration of those sensors which are new to the market is very difficult.

The other challenge is that there is a need of advanced IT skills which was difficult to handle tasks related to database system which like creating user interface so that the end user can monitor data in the database on back end where he can only monitor and visualize information.

The challenge of deploying model which is created using python was a problem because most of deploying is using apache database management system and it is not easy to accomplish this task of model deployment.

REFERENCES

[1] Keyur K Patel and Sunil M Patel,.Internet of Things-IOT:Definition,Characteristics, Architecture, Enabling Technologies, Application & Future Challenges. *International journal of engineering science and computing*, India: Vadodara, Gujarat, May, 2016.

[2] FAO, IFAD, UNICEF, WFP and WHO, *Transforming food systems for food security, improved nutrition and affordable healthy diets for all*, FAO, Italy: Rome, 2021.

[3] Kun Han, Shurong Liu, Dacheng Zhang, Ying Han, Initially Researches for the Development of SSME under the Background of IOT, *Physics Procedia*, Vol. 24, ISSN 1875-3892, pp. 1507-1513, 2012.

[4] A. Pathak et al., IoT based Smart System to Support Agricultural Parameters, *Procedia Computer Science*, Vol.155, ISSN 1877-0509, pp. 648-653, 2019.

[5] Collins Amoah-Antwi et al, Restoration of soil quality using biochar and brown coal waste. *Science of Total Environment*, Vol.722, ISSN 0048-9697, 2020.

[6] Eunice A. Mutuku, Dries Roobroeck, Bernard Vanlauwe, Pascal Boeckx, Wim M. Cornelis, Maize production under combined Conservation Agriculture and Integrated Soil Fertility Management in the sub-humid and semi-arid regions of Kenya. *Field Crops Research*, Vol.254, ISSN 0378-4290, Sept 2020.

[7] Jash Doshi, Tirthkumar Patel, Santosh kumar Bharti, Smart Farming using IoT, a solution for optimally monitoring farming conditions. *Procedia Computer Science*, Vol. 160, ISSN 1877-0509, pp. 746, 2019.

[8] Tibbett, Mark, et al., Long-term acidification of pH neutral grasslands affects soil biodiversity, fertility and function in a heathland restoration, *CATENA*, Vol.180, ISSN 0341-8162, pp. 401-415, Sept 2019.

[9] Xue-yun YANG, Ben-hua SUN and Shu-lan ZHANG, Trends of Yield and Soil Fertility in a long-term Wheat-maize System. *Journal of Integrative Agriculture*, Volume 13, ISSN 2095-3119, pp. 402-414, Feb 2014

[10] Vladimir Pavelčík and Erik Kuba, Application of basic machine learning algorithms in railway brake disc temperature prediction, *Transportation Research Procedia*, Vol. 55, ISSN 2352-1465, pp. 715-722,2021.

[11] Sathya, Ramadass, and Annamma Abraham, *Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification*. India, Bangalore, International Journal of Advanced Research in Artificial Intelligence, 2013.

[12] Jaydeep Karandikar, Tony Schmitz and Scott Smith., *Logistic classification for tool life modelling in machining*. USA: University of Tennessee, Knoxville, 2020.

[13] Priyanka Rajendra and Shahram Latifi,Prediction of diabetes using logistic regression and ensemble techniques, *Computer Methods and Programs in Biomedicine Update*, Vol. 1, ISSN 2666-9900, 2021.