How to Cite:
Rajkumar, S., Kaviya, P., & Mounica, M. (2022). A competent approach to predict the crop yield based on soil condition using CNN algorithm. *International Journal of Health Sciences*, *6*(S4), 5006–5016. https://doi.org/10.53730/ijhs.v6nS4.9230

A competent approach to predict the crop yield based on soil condition using CNN algorithm

Rajkumar S
Assistant Professor, K.S.Rangasamy College of Technology, Tiruchengode-637 215, Namakkal District, Tamil Nadu, India

Kaviya P
Students, Department of Computer Science and Engineering, K.S.Rangasamy College of Technology, Tiruchengode-637 215, Namakkal District, Tamil Nadu, India

Mounica M
Students, Department of Computer Science and Engineering, K.S.Rangasamy College of Technology, Tiruchengode-637 215, Namakkal District, Tamil Nadu, India

***Abstract*---**The quality of data collected from sensors greatly influences the performance of the modeling processes using ML algorithms. These elements (e.g., BD, ML) have been used extensively to improve all aspects of rice production processes in agriculture, ushering in a new era of rice smart farming or rice precision agriculture by transforming old rice farming practices into a new era of rice smart farming or rice precision agriculture. A review of the most recent studies on intelligent data processing technology in agriculture, with a focus on agricultural production forecasting. Machine learning is a useful decision-making tool for predicting crop yields, as well as for deciding what crops to plant and what to do during the crop's growth season. To extract and synthesize the techniques and features that have been employed in agricultural production prediction research, a systematic literature review (SLR) was conducted. According to analysis, the most used features are temperature, rainfall, and soil type, and the most applied algorithm is Artificial Neural Networks in these models.

***Keywords*---**crop yield, machine learning, CNN algorithm.

Manuscript submitted: 27 March 2022, Manuscript revised: 18 May 2022, Accepted for publication: 9 June 2022

Introduction

A literature study was conducted to gain an overview of what has been done on the application of machine learning in crop yield prediction (SLR). A Systematic Literature Review (SLR) shows the potential gaps in research on a particular area of problem and guides both practitioners and researchers who wish to do a new research study on that problem area. All relevant papers are acquired from electronic databases, synthesized, and presented to response to research questions stated in the study by following a methodology in SLR. An SLR study opens up new insights and aids new scholars in the subject in gaining a better understanding of the current state of the art. Machine learning (ML) methodologies are utilized in a variety of fields, from evaluating customer behavior in supermarkets to predicting customer phone usage. Machine learning has also been employed in agriculture for numerous years. Crop yield prediction is one of precision agriculture's most difficult problems, and numerous models have been suggested and confirmed thus far. Because crop production is affected by a range of factors such as climate, weather, soil, fertilizer use, and seed variety, this challenge necessitates the use of many datasets. This suggests that predicting agricultural yields is not a simple operation; rather, it entails a series of complex stages. Depending on the study challenge and research objectives, an ML model might be descriptive or predictive. Predictive models are used to make forecasts in the future, whereas descriptive models are used to obtain knowledge from the collected data and describe what has transpired. When attempting to create a high-performance prediction model, ML research provide a variety of problems. It's critical to choose the correct algorithms to tackle the problem at hand, and the algorithms and underlying platforms must be able to handle large amounts of data.

Literature Survey

Ishfaq Ahmad et.al., has proposed the Policymakers benefit from real-time, accurate, and trustworthy maize yield estimates when making decisions. The current research was designed to estimate spring maize yields using remote sensing and crop modelling. The CERES-Maize model was validated and evaluated with field experiment data in crop modelling, and then used to forecast maize production following calibration and evaluation. In Faisalabad, a field survey of 64 farms was done to collect information on initial field conditions and crop management. These data were utilized to anticipate maize yield in farmers' fields using a crop model. Peak season Lands at 8 photos were identified for land cover categorization using a machine learning technique using remote sensing.

Iftikhar Ali et.al., has proposed the grasslands are one of the major and crucial components of the terrestrial ecosystem and most prevalent and widespread global land cover types. Grasslands cover about 40.5% of the Earth’s surface and after forests, grasslands are the major source (about 30%) of carbon sink and thus play a very important role in regulating the global carbon cycle. The demand and consumption of dairy products are increasing globally in order to meet this demand, an equivalent growth in livestock has to be maintained. Grasslands encompass around 40.5 percent of the Earth's surface and are the second-largest source of carbon sink (about 30 percent), hence they play a critical role in regulating the global carbon cycle. Dairy product demand and consumption are rising internationally, and in order to supply this demand, livestock must also grow at a similar rate. Grasslands are the primary source of feed for grazing cattle, and the amount of above-ground biomass in a pasture determines it carrying capacity—the maximum number of animals that can graze it for a certain length of time without hurting it. Because grazed grass is the cheapest feed for cattle, it is critical to maintain grasslands, as better management will result in low-cost, high-quality feed. Unmanaged (natural) and Managed (agricultural pastures) grasslands. Weed control, removing dead plants, mowing, trimming, assessing growth rate, grazing length, and grassland usage are all examples of "grassland management" in the context of this study. Both ground-based traditional methods and remote sensing technology can be used to assess grassland biomass. The following are examples of ground-based methods: Visual: visual evaluation by human eye (expert or farmer); this method is spatially scarce and performs poorly. Cut and dry (Clipping): grass from the paddock is collected, dried, and weighed to determine the dry matter (DM) yield. Rising plate meter both mechanical and electronic, work on the same principle, with the plate rising up and down the shaft to measure grass height. Field spectrometry can also be used to estimate above-ground biomass by converting acquired spectra into reflectance and calibrating with biomass samples. Ground-based approaches are subjective, time-consuming, and only viable (or relevant) for small-scale grassland evaluation and monitoring. Remote sensing data is used in more complex and spatially wide grassland monitoring approaches. Sensors (optical and/or radar) deployed on various platforms can collect remote sensing data. For example, in airborne remote sensing, sensors are mounted on aero planes, helicopters, or unmanned aerial vehicles, whereas in satellite remote sensing, the sensor(s) are mounted on a spacecraft. Because the aircraft can fly under the cloud cover at an ideal period for data collection, airborne remote sensing is good for cloud-free data capture at a small scale.

Hulya Yalcin et.al., has proposed smart farming and precision agriculture are becoming increasingly important to cope with challenges due to the growth of world population. To ensure food security and sustainability in agricultural production, accurate crop yield prediction is an essential aspect of current agricultural technologies. Because environmental factors have such a large impact on a plant's growth, a precise estimate of agricultural yield can provide a wealth of information that can be used to improve crop quality. To estimate crop yield in field pictures, a deep learning architecture is used [7]. Cameras set on the ground agricultural stations capture plant photos every half hour. This method uses the intermediate outputs of deep learning architectures to create a measure for estimating agricultural yield. This estimate is a relative measure of crop yield estimations in comparison to the high crop production projections in agricultural parcels used to train the deep learning architecture. Thus experimented the approach on sunflower image sequences collected from four different parcels and obtained promising results.

Limitations

Neural network forecasting models are combined to increase the accuracy of prediction. In practice, this combination forecasting model is less dangerous, more obvious, and practicable. The models were at first trained on the correlation between previous environmental climate and crop yield rate. Finally, the models are compared to justify the accuracy.

Platform Prerequisite

JUPYTER NOTEBOOK: The Jupyter Notebook is a free and open-source web tool that lets you create and share documents with live code, equations, visualizations, and narrative text. Data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and many other applications are all possible. The software requirement specification is created at the end of the analysis task. As part of system engineering, the function and performance assigned to software are developed by creating a complete information report that includes a functional representation, a representation of system behavior, an indication of performance requirements and design constraints, and appropriate validation criteria.

ANACONDA CLOUD: Anaconda Cloud is a package management service by Anaconda. Cloud makes it easy to find, access, store and share public notebooks, environments, and condo and Py PI packages. Cloud also makes it easy to stay current with updates made to the packages and environments are using. Cloud hosts hundreds of useful Python packages, notebooks, projects and environments for a wide variety of applications. Not need to log in, or even to have a Cloud account, to search for public packages, download and install them and can build new conda packages using conda-build, then upload the packages to Cloud to quickly share with others or access yourself from anywhere. The Anaconda Cloud command line interface (CLI), anaconda-client, allows to manage the account - including authentication, tokens, upload, download, remove and search. Connect to and manage the Anaconda Cloud account. Upload packages have created. Generate access tokens to allow access to private packages. To build new conda packages using conda-build, then upload the packages to Cloud to quickly share with others or access yourself from anywhere. The Anaconda Cloud command line interface (CLI), anaconda-client, allows to manage the account - including authentication, tokens, upload, download, remove and search. Connect to and manage your Anaconda Cloud account. Upload packages have created. Generate access tokens to allow access to private packages.

PYTHON: Python is a high-level programming language designed to be easy to read and simple to implement. It is open source, which means it is free to use, even for commercial applications. Python is available for Mac, Windows, and Unix platforms, as well as Java and.NET virtual machines. Python, like Ruby or Perl, is a scripting language that is frequently used to create Web applications and dynamic Web content. Python is also supported by a variety of 2D and 3D imaging tools, allowing users to write custom plug-ins and extensions. GIMP, Inkscape, Blender, and Autodesk Maya are examples of software that provide a Python API. Python scripts (.PY files) can be parsed and executed right away. They can also be saved as compiled programs (.PYC files), which are commonly used as programming modules that other Python programs can reference.

Proposed Solution

The forecasting of crop production is done by using the time series data set precisely than the existing models. By using Boost technique, ensemble models such as are developed. To bring weak learners who are slow in learning, Prediction technique helps their understanding. CNN when joined with Prediction will make superior classification by giving weak learners with appropriate training. A like method is used for CNN classifier in which Prediction based CNN is used to generate superior classified data.

Validating the model is done by comparing the result of existing and the proposed technique. Forecasting of crop production is done by using the time series data set precisely than the existing models. Accuracy in time series analysis is the value forecasted which is very near to the actual value. In time series analysis, accuracy refers to the predicted value being very close to the actual value. Crop yield dataset is given as the input. Solving the problem in Indian context. The data about farm yield from different parts of region for particular time. It will be using this data for training models for crop yield prediction.

A convolutional neural network (CNN) is a class of artificial neural network, most commonly applied to analyze the dataset. Based on the shared-weight architecture of the 29 convolution kernels or filters that slide along input features and give translation equivariant responses known as feature maps, they are also known as shift invariant or space invariant artificial neural networks (SIANN). Data about agricultural yields from various parts of the region for a specific period of time. This information will be used to train crop yield prediction models.

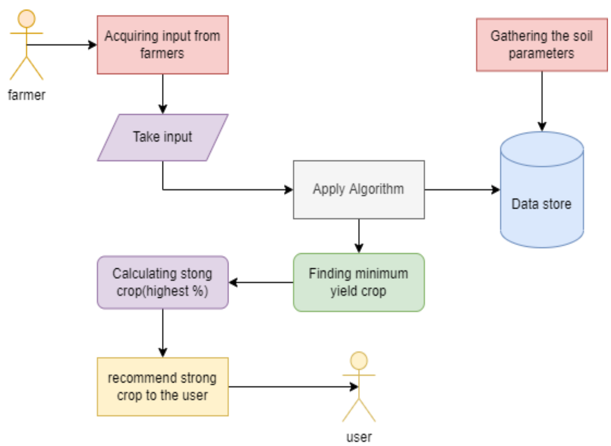

Figure 1: Architecture diagram of proposed system

Data Pre-Processing

Data pre-processing is the procedure for preparing raw data for use in a machine learning model. It's the first and most important stage in building a machine learning model. It is not always the case that clean and formatted data is available when working on a machine learning project. And, before doing any data-related activity, it is necessary to clean the data and format it. As a result, the data pre-processing activity is required. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. In the final data frame, there are two categorical columns in the data frame, categorical data are variables that contain label values rather than numeric values. The number of possible values is often limited to a fixed set, like in this case, items and countries values. Many machine learning algorithms cannot operate on label data directly, they require all input variables and output variables to be numeric.

Here the raw data in the crop data is cleaned and the metadata is appending to it by removing the things which are converted to the integer. So, the data is easy to train. First load the metadata into this and then this metadata will be attached to the data and replace the converted data with metadata. Then the data will be moved further and remove the unwanted data in the list and it will divide the data into the train and the test data. The splitting of the data into train and test need to import train test split which in the scikit-learn this will help the pre-processed data to split the data into train and test according to the given weight given in the code.

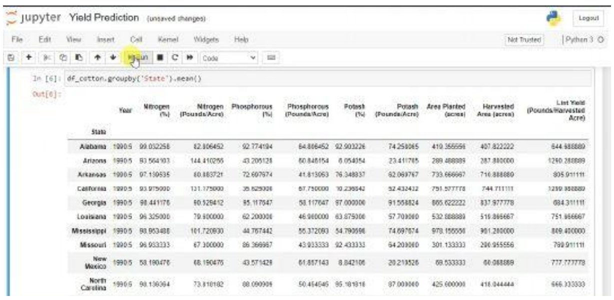

Figure 2: The crops are grouped and then shown in the particular value of the crops.

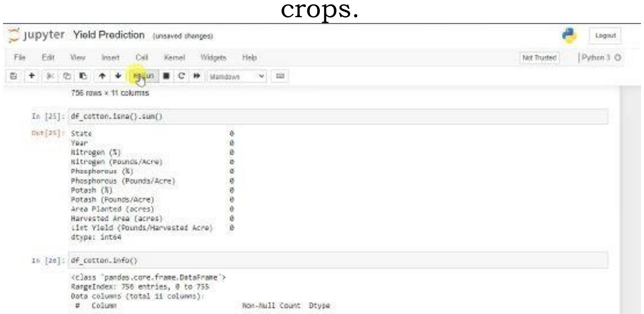

Figure 3: There is no null values in group by field. By using isna() function its shows the null value of the crop.

Feature Selection

When creating a predictive model, feature selection is the process of minimizing the number of input variables. The number of input variables should be reduced to lower the computational cost of modelling and, in some situations, to increase the model's performance. The relationship between each input variable and the goal variable is evaluated using statistics, and the input variables with the strongest link with the target variable are selected. These methods can be fast and effective, although the choice of statistical measures depends on the data type of both the input and output variables. The difference has to do with whether features are selected based on the target variable or not. Unsupervised feature selection strategies, such as methods that use correlation to remove redundant variables, disregard the target variable. Techniques that use the target variable, such as methods that remove unnecessary variables, are supervised feature selection techniques.

The process of applying statistical tests to inputs in order to produce a specific output is referred to as feature selection. The goal is to figure out which columns are more likely to anticipate the outcome. Apply the machine learning techniques which are helpful for finding crop yield for any of new data occurred in the data. A feature is an attribute that has an impact on a problem or is useful for the problem, and choosing the important features for the model is known as feature selection. Each machine learning process depends on feature engineering, which mainly contains two processes; which are Feature Selection and Feature Extraction.

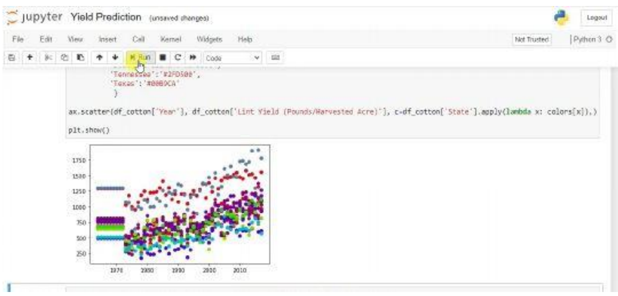

Figure 4: The dot graph shows the area scatter of the crop yield in the state in the particular year.

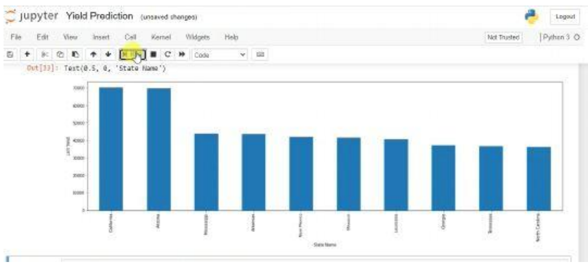

Figure 5: The histogram indicates that in which state the crops yielded highly

Feature Extraction

Feature extraction is the process of decreasing the number of resources needed to describe a huge amount of data. As a result of the analysis, the proposed model for determining crop production is more efficient than the existing model. The establishment of the above method would aid in the betterment of our country's agricultural practices. It can also be used to help farmers reduce their losses and increase crop yields in order to increase their capital in agriculture. The process of translating raw data into numerical features that may be handled while keeping the information in the original data set is known as feature extraction. It produces better outcomes than applying machine learning to raw data directly. Feature extraction is the process of transforming the space containing many dimensions into space with fewer dimensions. This approach is useful when want to keep the whole information but use fewer resources while processing the information.

The extraction for image data represents the interesting parts of an image as a compact feature vector. Previously, specialized feature identification, feature extraction, and feature matching algorithms were used to do this. The interesting sections of an image are represented as a compact feature vector in image data extraction. Previously, this was done using specialized feature identification, feature extraction, and feature matching algorithms. Deep learning is currently widely used in image and video analysis, and it is well-known for its capacity to process raw picture data without the need for feature extraction. Deep learning is frequently utilized in image and video analysis, and it is well-known for its ability to analyze raw image data without requiring feature extraction. Regardless of which approach you use, computer vision applications like image registration, object detection and classification, and content-based image retrieval all require effective representation of image features, either implicitly through the first layers of a deep network or explicitly through the use of long-established image feature extraction techniques.

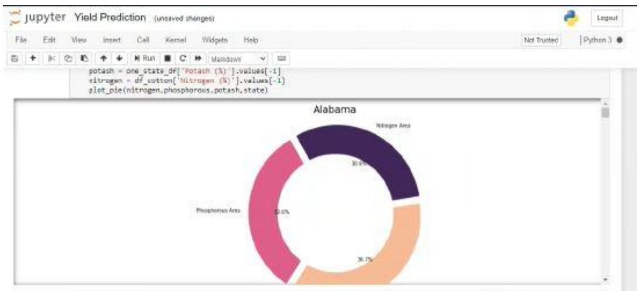

Figure 6**:** The pie chart represents the percentage of the particular crops yield in the state

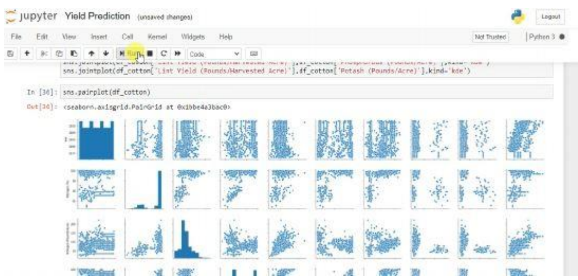

Figure 7. The seaborn jointplot displays the bivariate and uni-variate graphs.

Crop Yeild Prediction Analysis

Crop yield refers to the amount of agricultural production reaped in one unit of land area. The measurement is normally in tones or pounds per acre and is used for grains and cereals. For measurement, agricultural producers consider the amount of harvest per unit area. The entire farm's extrapolation is then done based on the crop's collected weight. Climate and its unpredictability are some of the factors behind the reduced rate of crop production. Weather forecasting is thus essential for improved management of crops. Agriculture has several other related industries like the sugar industry that depends on the sugarcane farmers.

Agricultural data is being produced constantly and enormously. As a result, agricultural data has come in the era of big data. Smart technologies contribute in data collection using electronic devices. The back propagation algorithms help in identifying the appropriate weigh value of the yield to calculate the error derivative. Accuracy of crop yield estimation is significant for agronomic production reasons. As a result, crop output prediction is critical for the global food production ecology. It becomes feasible to make educated decisions with improved data. Crop production forecast data is also useful to government organizations, as it allows them to prepare ahead for national food security. Crop yield forecast is a critical challenge for decision-makers at all levels, including national and regional (e.g., EU) decision-making. Farmers can use an accurate crop yield prediction model to assist them determine what to grow and when to grow it. Crop production prediction can be done in a variety of ways.

Figure 7. Analyzing the weather condition and soil condition of the land is good to yield the crops.

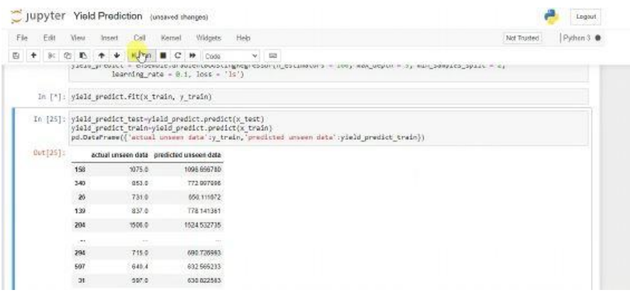

Figure 8. The yield prediction is done by testing and training data.

Conclusion

The way to improve the overall efficiency of the rice production system. The potential benefits lead to an improvement in the return of investment (ROI) for all rice production systems by minimizing the losses or costs involved in the production of rice. The choice of features is dependent on the availability of the dataset and the aim of the research. Studies also stated that models with more features did not always provide the best performance for the yield prediction. To find the best performing model, models with more and fewer features should be tested. Many algorithms have been used in different studies. The results show that no specific conclusion can be drawn as to what the best model is, but they clearly show that some machine learning models are used more than the others. Since Neural Networks is the most applied algorithm, we also aimed to investigate to what extent deep learning algorithms were used for crop yield prediction. There are also other kinds of algorithms applied to this problem.

Future work

In future aimed to build on the outcomes of this study and focus on the development of a DL-based crop yield prediction model.

References

[1] Ahmad, I., Fahad, M., Habib-ur-Rahman, M., Judge, J., Saeed, U., Ullah, A.,2018. Yield forecasting of spring maize using remote sensing and crop modeling in Faisalabad-Punjab Pakistan. J. Indian Soc. Remote Sens. 46 (10), 1701– 1711.

[2] Bargoti, S., Underwood, J.P., 2017. Image segmentation for fruit detection and yield estimation in apple orchards. IEEE J. Field Rob. 34 (6),1039–1060.

[3] Cawkwell, F., Dwyer, E., Green, S., 2017. Modeling managed grassland biomass estimation by using multi temporal remote sensing data—a machine learning approach. IEEE J. Sel. Top. Appl. Earth Obs. Remote .10 (7), 3254–3264.

[4] Chukalla, A.D., Hoekstra, A.Y., Krol, M.S., 2018. Grey water footprint reduction in irrigated crop production: effect of nitrogen application rate, nitrogen form, tillage practice and irrigation strategy. Hydrol. Earth Syst. Sci. 22,3245–3259.

[5] Cassman, K.G., Edreira, J.I.R., Guilpart, N., R.L.,Grassini, P., Sadras, V., Schils, Van Ittersum, M.K., 2018. Water productivity of rainfed maize and wheat: a local to global perspective. Agric. For. Meteorol. 259,364–373

[6] Filippi, P., Jones, E.J., Pozza, L.E., Somarathna, P.D.S.N., Ugbaje,S.U., T.F.A., Wimalathunge, N.S., 2019b. An approach to forecast grain crop yield using multilayered multi-farm data sets and machine learning. IEEE Precis.Agric.

[7] Goldstein, A., Lutenberg, O., Meitin, A., Ravid, G., 2018. Applying machine learning on sensor data for irrigation recommendations: revealing the agronomist's tacit knowledge. IEEE Precis. Agric. 19 (3),421–444.

[8] Gaku Yokoyama, Jiaojiao Feng, Keisuke Minami, Kensuke Kimura, Masaharu Kitano., Yueru Wud, 2021, Evaluation of the physiological significance of leaf wetting by dew as a supplemental water resource in semi-arid

cropproduction.
[9] Haneen Ali, Huiyang Li, Piterson Lisboa Department of Systems Science and Industrial Engineering, State University of New York atBinghamton.
[10] Jia, Z., Wu, W., Zhao, Z., Zhang, Q., 2019. Dew yield and its influencing factors at the western edge of Gurbantunggut Desert, China. Water 11,733.

[11] K., Du, J., Poku, J., 2018. Causal relationship between agricultural production and carbon dioxide emissions in selected emerging economies. Environ. Sci. Pollut. Res. 25,24764–24777.
[12] Lecourt, J., 2018. Advances in non-destructive early assessment of fruit ripeness towards defining optimal time of harvest and Yield prediction—a review. IEEE Plants 7(1).
[13] Lin, B., 2017. Analysis of energy related carbon dioxide emission and reduction potential in Pakistan. J. Clean. Prod. 143,278–287.
[14] Moret, E.; Muntané, E.; Ricart, P.; Bonafont, X European JournalofAnaesthesiology:June2011– Volume28– Issue-p218
[15] N.R., Moore, M. R., Schaible, G., 2019. Modeling Western Irrigated Agriculture and Water Policy: Climate-Change Considerations. In Economic Issues in Global Climate Change(pp.148–167).
[16] Osman, T., Psyche, S.S., Rahman, R.M., Tamanna, F., 2017. Predicting early crop production by analysing prior environment factors, IEEE pp.470–479.
[17] Ranjan, A.K., 2019. Paddy acreage mapping and yield predicton using sentinel-based optical and SAR data in Sahibgani district, Jharkhand (India). IEEE Spatial Inf. Res.
[18] Tan, M., Zheng, L., 2017. Different irrigation water requirements of seed corn and field corn in the Heihe River Basin. Water 9, 606.
[19] Wu, X., Xu, X., Zhu, X., 2019. Design of an integrated climatic assessment indicator (ICAI) for wheat production: a case study in Jiangsu Province, China. IEEE Ecol. Ind. 101, 943-953.
[20] Yalcin, H., 2019 . An approximation for a relative crop yield estimate from image using deep learning. IEEE, pp. 1-6.
[21] Yang, Q., Zha, Y., Zhu, P., 2019. Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images. Field Crops Res. 235, 142-153.
[22] Zhang, Q., 2019. Variation characteristics of non-rainfall water and its contribution to crop wate requirements in China's summer monsoon transition zone. J. Hydrol. 578, 124039
[23] Zafeiriou, E., 2017. CO2 emission and economic performance in EU agriculture: Some evidence from Mediterranean countries.
[24] Rinartha, K., & Suryasa, W. (2017). Comparative study for better result on query suggestion of article searching with MySQL pattern matching and Jaccard similarity. In *2017 5th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1-4). IEEE.
[25] Rinartha, K., Suryasa, W., & Kartika, L. G. S. (2018). Comparative Analysis of String Similarity on Dynamic Query Suggestions. In *2018 Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)* (pp. 399-404). IEEE.
[26] Lopez, M. M. L., Herrera, J. C. E., Figueroa, Y. G. M., & Sanchez, P. K. M. (2019). Neuroscience role in education. *International Journal of Health & Medical Sciences*, *3*(1), 21-28. https://doi.org/10.31295/ijhms.v3n1.109