002 Nilesh Team

by Nilesh Team

Submission date: 09-Nov-2022 10:02PM (UTC+0530)
Submission ID: 1949295400
File name: J_Component_Final_Report_3_without_plag.docx (1.42M)
Word count: 8196
Character count: 41586

Agriculture is the backbone of any economy; it doesn't matter whether the country is developing or developed. In countries such as India, where more than 50% of the population is dependent on agriculture for their source of income, its importance increases manifolds.

In recent years, we have observed that there has been a strain placed on the existing lands as well as the farmers in those areas to produce more crops in order to satisfy the requirements of the growing population. Due to a lack of information on the part of the farmers about how the weather and the conditions of the soil affect crop growth, the actual yield does not achieve the target that was expected. Here, we mean that they use some traditional ways that in May Summer will be there, December for Winters etc. but weather can change too. Unexpected rains in Summer, High Temperature in Winter drastically affect the crop.

In addition to this, policy makers themselves are unable to establish proper policies due to a lack of knowledge of the factors that are being discussed here.

We are attempting to create a Deep Learning model that can estimate the amount of produce that can be harvested from a particular plot of land over a particular amount of time. In order to construct this model, we will first conduct research over an extended period of time on a variety of elements, including temperature, humidity, soil conditions, and rainfall. In this discussion, we will look at the situation of the paddy crop in Tamil Nadu.

Our model will be of assistance to farmers in their efforts to deduce processes that will boost the agricultural output on their respective farms. In addition to this, it will be helpful to policymakers as they formulate measures to ensure the state's food security.

Alongside various representations of the analysis performed on the historical data, the forecast will be presented in a user-friendly user interface. This will be done when the forecast has been generated. In addition to this, it will provide some prescriptive methods that can be taken to increase the crop production of the area.

The role that technology plays in today's society is rapidly expanding and playing an increasingly important role. Technology is currently the primary factor in the success of any endeavor, including commercial enterprises, educational institutions, and medical practices. By automating processes that were previously labor-intensive, this has made a significant improvement in everything's efficiency and smoothness. On the other hand, it is common knowledge that technology generates large amounts of data, which has resulted in the development of a number of sectors that are all dependent on data and its ever-increasing significance. The analysis and visualization of data are two examples of this.

There is no hiding the fact that we are currently living in the age of big data, which is a time when information is vital, especially in the corporate world. In a circumstance like this, data analytics and visualization take on increased significance. Despite the fact that the majority of people are likely to be inexperienced with these concepts, investing in data analytics and visualization could be the difference between a firm being successful or unsuccessful.

Agriculture is the sector that supports both developing economies and those that are more developed. Its relevance is magnified in countries such as India, where about half of the population relies on it for their own existence, and nowhere else in the world.

Recent years have witnessed a rise in the amount of pressure placed on farmers and land that is already in use to enhance food production in order to meet the needs of an expanding population. The production falls short of the target unfortunately due to a lack of understanding on the part of the farmers regarding the conditions of the weather and the soil. In addition, policymakers themselves are unable to establish accurate policies since they do not have adequate knowledge of the relevant parameters.

The most important industry in India is agriculture, which is also absolutely crucial to the survival of rural areas and the overall economy of the country. Agriculture accounts for around 70 percent of both the primary and secondary economic sectors. As a direct consequence of this, a significant number of farmers are beginning to incorporate contemporary methods and equipment into their farming operations. On the other hand, there are a lot of people who aren't aware of how important it is to produce crops at the appropriate time and in the appropriate location. The identification of crop adaptation and yield using a number of elements that influence production can, in this instance, boost crop quality and yield, which can lead to better economic growth and profitability. Crop development, despite being challenging, is one of the phenomena that agricultural input parameters advise on. Data mining is the process of extracting previously unanticipated information from large databases and is a commonly used term for this practice. The ability to mine data helps firms analyze future behavior and patterns, which in turn enables them to make decisions that are more informed. The process of examining, cleaning, and modeling data in order to obtain insightful knowledge and meaningful conclusions is referred to as data analysis.

We are now working on developing a deep learning model in order to achieve our goal of accurately predicting the agricultural output of a certain parcel of land for a specific period of time. This model will be developed as we conduct research on a wide range of factors over an extended period of time, including

temperature, humidity, the quality of the soil, and rainfall. In this particular situation, we'll use the example of the paddy crop that is grown in Tamil Nadu.

Our method will be of assistance to farmers in the process of formulating plans to increase agricultural production on their own farms. In addition to this, it will be helpful in the process of formulating policies that will ensure the food security of the state. The analysis that was done on the historical data will be displayed, together with some visualizations of the results, in a user interface that is quite straightforward. In addition to this, it will provide some recommendations on how the agricultural yield might be increased on the land.

The proliferation of technology as an essential driving force behind each and every industry in the modern world has led to the generation of massive amounts of data. The interpretation of this information is impossible for organizations to accomplish without data analytics. The term "analytics" refers to a wide-ranging discipline that incorporates several different subfields. It explains all of the techniques and processes that are utilized in order to perform data analysis and provide meaningful conclusions and interpretations. It is essential to keep in mind that data analytics is dependent on the use of specific computer programs and tools in order to collect data and perform suitable analysis on it before appropriate business decisions can be made.

Data analytics is utilized extensively in the business world since it helps companies gain a deeper understanding of their customer base and improve the effectiveness of their marketing campaigns. This sector is always evolving, as seen by the plethora of new advancements that are regularly brought to our attention. At the time that this article was written, Data Analytics had become a mechanized industry that relied on computer algorithms to analyze raw data and draw accurate conclusions.

The application of a deep learning algorithm is currently underway in order to forecast the crop yield of a specific parcel of land. This study is therefore both predictive and prescriptive in nature. In order to train the model, we will use-

- Recurrent Neural Network (RNN)
- Long Short Term Memory (LSTM)

In these kinds of situations, feed forward neural networks are typically the best choice. However, it does have some restrictions. That is what:

• Cannot process sequential data

• Only takes into account the most recent input

• Is unable to remember the most recent input

The RNN is the answer to these problems and concerns. An RNN is able to process sequential data by taking in both the data that is currently being input and the data that was input in the past. RNNs have their own internal memory, which allows them to remember prior inputs.

Despite the fact that RNNs are highly effective, they are plagued by the issues of exploding and vanishing gradients. In the situation of exploding gradients, the algorithm assigns greater weights than are necessary to the inputs. On the other hand, in the event of vanishing gradients, the gradients become too small, and after a certain number of iterations, the model stops learning. In order to solve this issue, LSTM is combined with RNN as an analytical tool.

Neural networks, which are collections of algorithms that are designed to closely mirror the functioning of the human brain, are utilized so that patterns may be recognised. They recognise or categorize raw input in order to use machine perception to understand sensory data. They are able to recognise the numerical patterns that are present in the vectors that need to be translated in order for any real-world data (images, music, text, or time series) to exist. These patterns can be found in the vectors that must be translated in order for any data to exist. Artificial neural networks are made up of a large number of processing units that are extremely interconnected and work together to solve a problem. These processing units are termed neurons.

In an ANN, the processors are often organized in tiers and involve a high number of parallel processors. The first layer, which functions similarly to the optic nerves in the human visual processing system, is where the raw input data is received. In a manner analogous to that in which neurons located further away from the optic nerve receive signals from those located closer to it, each succeeding layer instead receives the output from the tier that came before it rather than the raw input. The output of the system is produced by the third and final stage.

RNNs are a powerful and dependable sort of neural network, and they are one of the most promising ones that are currently being employed. RNNs are the only type of neural network that has an internal memory.

One prevalent strategy for deep learning that has been around for some time is known as recurrent neural networks. Even though they were first invented in the 1980s, it wasn't until very recently that we fully understood their prospective applications. The rise in processing power, the vast amounts of data to which we now have access, and the advent of long short-term memory (LSTM) in the 1990s have all contributed to the increased interest in RNNs.

Because RNNs have their own internal memory, which allows them to remember important information about the input they were given, they are able to make accurate predictions about what will happen in the future. They are the algorithm of choice for processing sequential data of many different forms, including but not limited to text, financial data, weather, time series, speech, audio, video and many more. Recurrent neural networks are able to create a significantly better understanding of a sequence and the context in which it exists when compared to other methods.

Recurrent neural networks, often known as RNNs, are a type of neural network that can be utilized for the simulation of sequence data. RNNs, which are created from feedforward networks, show behaviors that are analogous to those of human brains in certain respects. To put it more simply, recurrent neural networks have the ability to anticipate outcomes in sequential data, something that other algorithms do not possess.

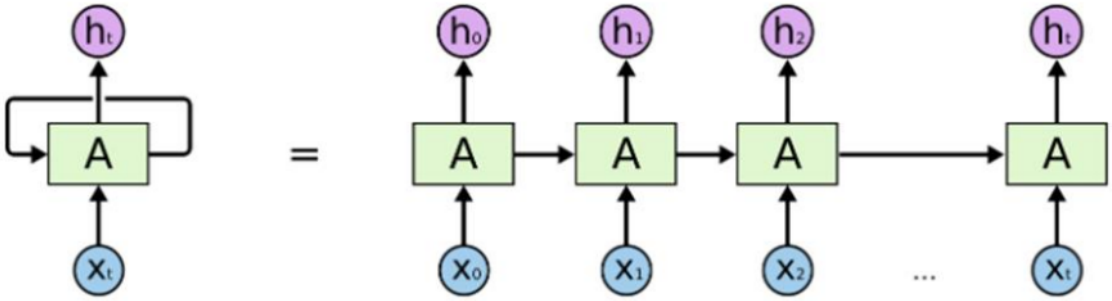
When there is a sequence of data and a temporal dynamic that connects the data, the geographical content of each individual frame becomes less significant. This is the case whenever there is a map. As a result of their incorporation into the software that runs Siri and Google Translate, recurrent neural networks, also known as RNNs, are widely encountered in everyday life. Traditional neural networks do not have the capacity to store information from the network's previous experiences.use past information to make predictions about future values

This issue was solved by using Recurrent Neural Networks, which are very similar to Neural Networks. Looping networks were also utilized. The following is what an illustration of RNN will look like after its loop is uncoiled:

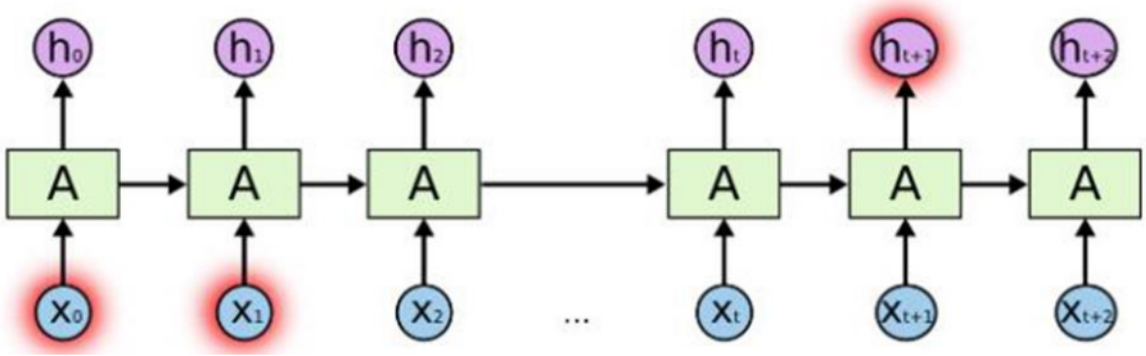
From this vantage point, we are able to see the neural network. Learning for A takes place at time t as a direct result of data x(t) transmits the output to the result while simultaneously feeding it back to itself in order to make it compatible with incoming input.

Now, when we limited our analysis to just the prediction based on this, we found that it was actually quite helpful data recently. Imagine you are in a position where you need to make a prediction utilizing highly outdated information. It will not work, and the effects will be unpredictable. The phrase "vanishing" refers to the situation in which historical data are no longer taken into consideration while making forecasts about the future.

Graduation as well as the allocation of an extraordinarily high weight. The following is the rationale behind the utilization of LSTM, which is an abbreviation for long short-term memory:

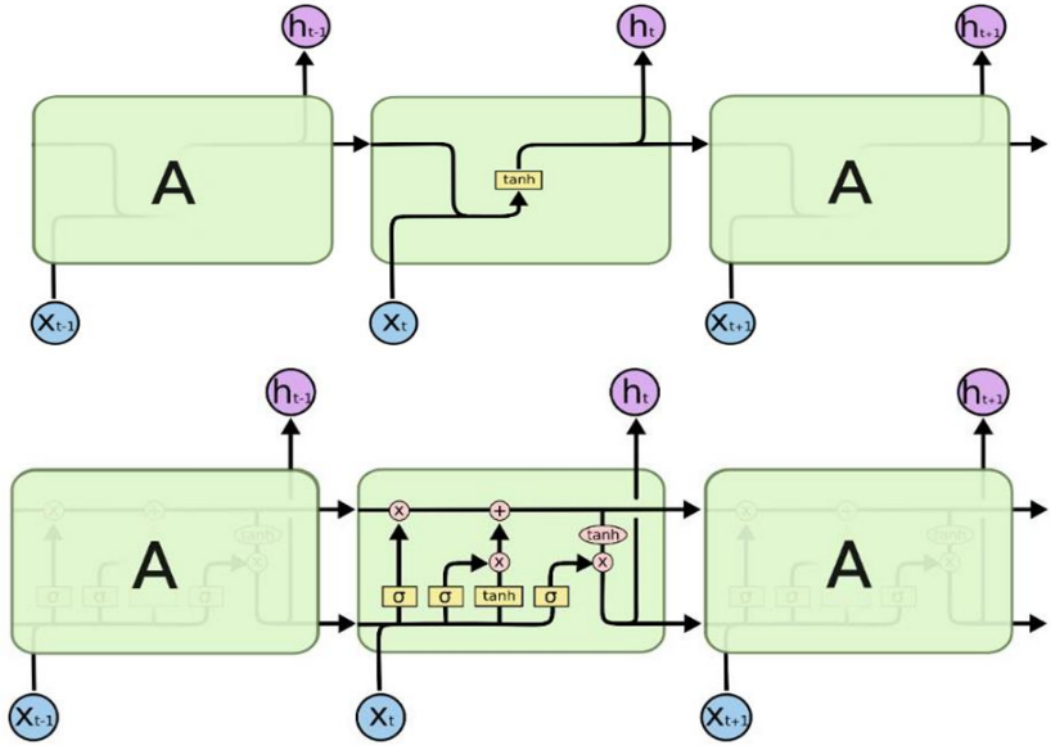


The retention of information is made possible by an improved RNN or sequential network that uses a technique known as long short-term memory network. It is able to solve the problem of vanishing gradients that was presented by the RNN. When it comes to persistent memory, RNNs, which are also known as recurrent neural networks, are the way to go.



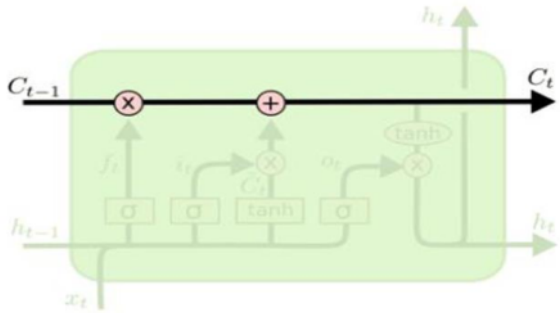
Suppose you can remember the scene you saw before this one while watching a movie or the events from the chapter you read before this one while you were reading a book. In a manner analogous to that which is performed by RNNs, these systems remember previously acquired information and make use of it while processing newly received input. The inability of RNNs to remember long-term dependencies is caused by the fact that gradients tend to decrease over time. When building LSTMs, designers make it a point to steer clear of problems involving long-term dependencies. In order to maintain the connection, it is preconfigured to store a record of the data it has previously received. A

This is made abundantly evident when we examine the structure of an RNN and an LSTM from the inside out understanding of coming out on top in the end.



When compared to the LSTM, the RNN (first image) consists of only one layer while the LSTM consists of numerous layers.

There are a lot of layers here.



The ability of agriculture to accurately anticipate crop yields is essential. When it comes to planning and decision-making, farmers and policymakers alike stand to gain a great deal by having access to timely and accurate estimates of crop production.

Statistical models are frequently utilized, which is a process that is time-consuming and labor-intensive to forecast agricultural yield. Both deep learning and machine learning are relatively recent developments that have had a substantial impact on the industry. Because deep learning models inherently have the ability to extract features from large datasets, they are better suited for making predictions than traditional machine learning models. In this study, a Recurrent Neural Network (RNN) model that is based on deep learning is utilized to make predictions regarding the wheat crop yield in the northern region of India. In addition, LSTM was utilized in this investigation to address the vanishing gradient problem that was caused by the RNN model.

Experiments utilizing a benchmark dataset spanning 43 years were used to evaluate the performance of the proposed machine learning model in comparison to three other models. The efficiency of the model was clearly proved by the findings of the RNN-LSTM model, as well as those of the Artificial Neural Network, Random Forest, and Multivariate Linear Regression. In addition, it was demonstrated that the work that was suggested was effective by demonstrating that the RNN-LSTM model's anticipated crop production values were proved to be more accurate than true values.

Agriculture is important in developing countries. The majority of people in India are dependent on agriculture for their livelihood. Agriculture-related issues constitute a key barrier for developing countries.

The most effective solution to this problem would be to use "smart agriculture," which makes use of contemporary farming practices. Precision farming can be accomplished through the application of information and communication technology (ICT) in the agricultural sector. The term "smart agriculture" refers to a set of practices that include the use of sensors to monitor a field, the processing of data, and the development of a crop management system.

In addition to this, it provides intelligent irrigation as well as fertilizer supply that is based on the requirements of the land. Data is a vital component in precision farming. Data can be collected in a variety of ways, including manually, through satellite sources, or from sensor sources. The information that is acquired might reveal the many types of nutrients that are found in soils. Utilization of this study may make it feasible to increase the output of agricultural production. Cloud storage allows dairy farming activities to be stored, which improves data accessibility, storage, and financial advantages for farmers.

Forecasting the output of crops is absolutely necessary if one wishes to keep the actual demand and supply of food in equilibrium.

The most important industry in India is agriculture, which is also absolutely crucial to the survival of rural areas and the overall economy of the country. Agriculture accounts for around 70 percent of both the primary and secondary economic sectors. As a direct consequence of this, a significant number of farmers are beginning to incorporate contemporary methods and equipment into their farming operations.

On the other hand, there are a lot of people who aren't aware of how important it is to produce crops at the appropriate time and in the appropriate location. The identification of crop adaptation and yield using a number of elements that influence production can, in this instance, boost crop quality and yield, which can lead to better economic growth and profitability. Crop development, despite being challenging, is one of the phenomena that agricultural input parameters advise on. Data mining is the process of extracting previously unanticipated information from large databases and is a commonly used term for this practice. The ability to mine data helps firms analyze future behavior and patterns, which in turn enables them to make decisions that are more informed. The process of examining, cleaning, and modeling data in order to obtain insightful knowledge and meaningful conclusions is referred to as data analysis.

The yield of a crop is a highly complicated characteristic that is controlled by a number of factors, including its genotype, environment, and the interactions between the two. A fundamental understanding of the functional link that exists between yield and various interaction components is required in order to make an accurate prediction of yield. In order to uncover such a relationship,

extensive datasets and effective algorithms are required. Syngenta distributed numerous huge datasets as part of the 2018 Syngenta Crop Challenge.

These datasets captured the genotype and yield performances of 2,267 maize hybrids that were planted in 2,247 locations between 2008-2016. Participants were challenged to forecast the yield performance in 2017. We were one of the winning teams, and one of the things that helped us win was the deep neural network (DNN) approach that we developed. This approach made use of cutting-edge modeling and solution approaches. Our model was shown to have better prediction accuracy, with a root-mean-square-error (RMSE) of 12% of the average yield and a standard deviation of 50% for the validation dataset including projected meteorological data. This conclusion was reached because of the fact that RMSE was computed using the average return and standard deviation.

If the weather data were perfect, the relative standard error (RMSE) would be lowered to 11% of the average yield but the standard deviation would be 46%. Additionally, we used feature extraction based on the training DNN model, which was successful in reducing the input space's size without noticeably lowering prediction accuracy. This was a successful outcome. According to the findings of our computations, this model appeared to perform noticeably better than a number of other well-known methods, including Lasso, shallow neural networks (SNN), and regression trees (RT). The findings also demonstrated that environmental conditions had a bigger influence on crop output than the genotype of the plant.

Predicting crop yield by taking into account environmental, soil, water, and crop parameters has been identified as a potential area of research. Extraction of significant crop features for the purpose of prediction is made widespread use of models based on deep learning. Despite the fact that these methods have the potential to solve the problem of yield prediction, there are still the following.

Inadequacies include not being able to create a direct non-linear or linear mapping between the raw data and the crop yield values; the performance of those models heavily depends on the quality of the extracted features; and the inability to create a direct mapping between the raw data and the crop yield values. The aforementioned deficiencies can be addressed with greater direction and motivation thanks to deep reinforcement learning.

Deep reinforcement learning is a method that builds a comprehensive framework for crop yield prediction by combining the intelligence of reinforcement learning and deep learning. This framework is able to map raw data to crop prediction values. In the work that is being proposed, Deep Recurrent Q-Network model will be built. This model will forecast crop yield using a Recurrent Neural Network deep learning algorithm that will be applied on top of a Q-Learning reinforcement learning algorithm. The data parameters are what provide food for the recurrent

neural network's layers that are stacked sequentially. A setting for crop yield prediction is fabricated by the Q-learning network on the basis of the parameters that are fed into it. A linear layer acts as a mapping between the Q-values and the output values of a recurrent neural network. The reinforcement learning agent takes into account a number of parametric features in addition to the threshold, both of which contribute to the process of crop yield prediction. At the end of the process, the agent is awarded an aggregate score based on the actions taken to reduce error while simultaneously improving the accuracy of the forecast.
The proposed model accurately predicts the crop yield with a level of precision that is 93.7% higher than that of the existing models, which it does by maintaining the distribution of the original data.

In recent years, the importance of the part that data analytics plays in the process of converting raw data and information into actionable decisions in the field of agriculture has been steadily growing. In recent years, the concept of data analytics has been widely used in a variety of sectors for the purpose of efficient decision making, and more and more people are becoming aware of the significance of the value that data analytics provides. As a result, it has been demonstrated that there is hardly any sector in which data analytics cannot play a role. In this context, machine learning also has a role to play, particularly in the agricultural sector, in addition to the roles it plays in the fields of education, banking, healthcare, and retail. In order to make progress along this route of transformation, it is essential to keep in mind that the primary challenge lies in the transformation of data from an unstructured to a structured format, and this must occur even before the application of machine learning algorithms.

In the event that farmers are given the support of effective decision making options through the application of data analytics, this would result in a contribution to the level of the economy. As a result, the researchers who participated in this study attempted to suggest a platform for data analytics that would cater to the requirements of the stakeholders in the agricultural sector while also bearing in mind the concept of an "intelligent agricultural decision system." The findings of this study would guarantee the platform that has been suggested, in addition to paving the way for additional research on more sophisticated decision-making in the field of agriculture.

Although Deep Learning has been applied to the problem of predicting crop yields, there is not yet a systematic analysis of the studies that have been conducted. As a result, the purpose of this investigation is to provide an analysis of the most recent developments in the field of crop yield prediction using Deep Learning. We conducted a Systematic Literature Review (SLR) in order to locate and evaluate the papers that were most pertinent to our study. Following the application of selection and quality assessment criteria to the relevant studies, we were able to retrieve a total of 456 primary studies, from which we ultimately chose 44 primary studies for further investigation. An in-depth analysis and synthesis of the primary studies were carried out with regard to the

primary motivations, the target crops, the algorithms used, the features that were utilized, and the data sources that were utilized.

We found that the Convolutional Neural Network, abbreviated as CNN, is the algorithm that is used the most, and it also has the best performance in terms of Root Mean Square Error (RMSE).

One of the biggest and most essential problems is the absence of a huge training dataset, which increases the possibility of overfitting and, ultimately, worse model performance in real-world applications. It is useful to highlight the challenges that are now being encountered as well as the possibilities for further exploration since researchers in this discipline have a propensity to focus on the relevance of unresolved research issues.

The agricultural sector plays a significant role in India's economy. Agriculture is the sole sector that contributes to India's economy. In the coming years, a significant problem that arises is the rapidly growing population as well as the diminishing quality of the land. In order to fulfill the requirement, there is a need to boost agricultural production relative to the population. To put it another way, we will have to figure out how to produce more with fewer resources. There have already been two major revolutions in the agricultural industry. The first one occurred during the time of the Industrial Revolution, when production began to be mechanized. The second one occurred during the Green Revolution, when pesticides and other types of agrochemicals became widely used. The Big Data revolution is the third one that is currently taking place.

The application of Big Data in farming therefore has the potential to increase output while decreasing inputs. Precision agriculture is the answer to this problem. It is nothing more than the application of data mining technology combined with big data to agricultural data. When agricultural data is combined with technological advancements, the result is an increase in crop yield production. This occurs despite the presence of factors that can have an impact on agricultural production, including the climate, the composition of the soil, and various kinds of pests.

The actual practices need to be carried out on a relatively small area before moving on to a larger one. The findings might make it easier for farmers to embrace new forms of technology. The field was broken up into subunits using the information that was previously accumulated, taking into account the soil texture, the contour of the land, and the water level [8]. Different seeds will need to be planted in each subdivision of the land because the terrain varies so much. This will contribute to an increase in the production of crop yields.

The yield of maize is highly variable from year to year, and climate change is responsible for 39% of that variation (Ray et al., 2015). In regions where maize serves as a primary source of nutrition, adaptation to climate change and its effects must be prioritized if food security is to be maintained.

The majority of the world's maize is grown by small-scale farmers, who account for 82 percent of the total global maize production area.

The majority of the world's maize is grown in areas that are not considered major maize growing areas. On farms that are less than 2 ha in size, approximately one quarter of the world's food supply is produced. Small-scale farms in developing countries frequently face severe financial and infrastructural constraints. However, only a small fraction of these farms have access to newly developed digital agricultural technologies that are commercially available, and the majority of farms still do not have internet access.

In addition, small-scale farmers in developing countries do not routinely keep records of their farms and do not have easy access to information regarding the weather. As a result, they are typically unable to perform an analysis of what took place in the past and come to data based conclusions about what will take place in the future. In this context, many small-scale farmers get their knowledge about which crops to grow and how to manage them from talking to other farmers and receiving assistance from extension services. This information sharing between farmers is a more efficient and cost-effective alternative to traditional methods of farmer training.

Several case studies have used machine learning to estimate crop productivity. Data and methodology may not apply to other crops or regions.. Machine learning has great promise when enormous amounts of public data are collected. Combining agronomic crop modeling with machine learning, we established a machine learning baseline for large-scale agricultural yield forecasting. This improved agricultural yield predictions. Baseline is a correct, modular, and reusable procedure. To achieve accurate results, we developed explainable predictors and applied ML to limit information leakage. We used crop simulation outputs and MCYFS weather, remote sensing, and soil data to build features. We focused on designing a modular, repeatable workflow that could handle different crops and countries with minimum configuration changes. Using standard input data, the method can be used to do repeatable experiments (such as season forecasts) and produce reproducible results. The findings can help optimize future attempts.

We forecasted regional yields for five crops in the Netherlands, Germany, and France in the case study we have taken. Soft wheat, sunflower spring barley, potatoes and sugar beets. We tested our algorithm to a simple technique without predictive abilities that merely predicted a linear yield trend or the training set mean. We collected all the projections into one national report and compared it to past MCYFS estimates. Normalized RMSE (NRMSE) for NL (all crops), DE (all except soft wheat), and FR (soft wheat, spring barley, sunflower) were equivalent 30 days after planting. The Netherlands' NRMSE for soft wheat was 7.87 (6.32 for MCYFS), and Germany's was 8.21. (8.79 for MCYFS). Soft wheat (Germany), sugar beet (France), and potatoes (France) had NRMSEs double MCYFS. DE and FR's end-of-season NRMSEs were worse than NL's

MCYFS. Additional data sources, predictive features, and machine learning algorithms can improve the baseline. The baseline will help apply machine learning to crop yield predictions.

To achieve our objectives, first of all it is required to collect the data relevant to our project. As we are talking about crop prediction in agriculture, we require those factors which affect the growth and production of crop. From our basic understanding and research, some of the major factors are: Sunshine, Humidity, Air Speed, Air Quality, Soil Moisture, Temperature, Soil Nutrition, Soil Type etc. Now, for these factors, we need to find a proper dataset which caters to our needs by having all these factors with proper metadata. By metadata here, we mean that it should describe how the data was collected, the authenticity of data, the units in which the data was stored etc. Also, we have to decide the time frame for which the data we want. Generally, for such prediction it is desirable to have data of at least 15-20 years but rest depends on the availability of data. Plus, it is also needed to have data sampled on some time units, like are we going to have a data which corresponds to the whole year or is it in daily, monthly or quarterly form? For crops, it is desirable to have the data which corresponds to the crop season and here, we find that it is foremost important to define the crop for which the whole project is being made. Also, we have to ensure that we choose a region for our project because if we go on to analyze the data without any specificity, then our project will be considered absurd. Along with the factors, we will require the Crop Yield of that particular crop during that year as we will test our model by predicting yield and comparing it against the provided yield. After the collection of data, ensure that there is no inconsistency in the data; inconsistencies can lead to unwanted results which may tamper our model's predictions. Such inconsistencies can be taken care by either removing those attributes which have lot of inconsistencies or we can work on correcting those data by inputting our own data there. This is called Synthetic Data Creation, although we are not creating the whole data but filling out some data based the understanding of the real dataset. After all this processes are over, we need to segregate the data into train and test datasets; train will be the one which will be fed to the model to learn the prediction from the data. After that, it will be tested against the test data to get the comparative study. After this splitting, we are ready to move forward with the next step.

After the data collection and pre-processing part, we proceed with the algorithm selection part for our project. As discussed above we are going with Long-Short-Term-Memory Algorithm, which is a modified-but advanced version of Recurrent Neural Networks (algorithm). The main reason for which we chose this algorithm is its ability to remember the past data properly and accordingly work it and make its importance visible when predicting new data (i.e. removing the vanishing gradient to a greater extent). Now for implementing this algorithm, we can use various tools, most famous being writing a python code for the same using various built-in libraries. Discussion on more about the coding aspect will be discussed in the 'Experiments Section'. Now, to check our model's competency, we will be doing two things; first of all, create one more different machine learning model and second, choose some standard performance metrics which are globally accepted for such research works. Now, we will compare the two models on the basis of these performance metrics and also try plot the values of test and train results of the dataset's splits done above to know which model is better.

From this, our discussion changes to the point of selecting a model which we can use in this scenario. It is understood that a machine learning model can be used for different scenarios and for a single scenario, there can be various accepted machine learning models. We just have to look into the application and decide the model. As discussed above, we chose LSTM because this topic of Crop Prediction is time-dependent; huge past data are required to predict data of 20-30 years later, and it has the capability to do that. From our research and reference from the base paper, we came across two machine learning models which can be compared against LSTM. They are:

1. Multivariate Linear Regression: It is categorized under supervised machine learning. It is all about using an equation given by following, equivalent to the general equation y = mx+c.

$$y = \theta_0 + \sum_{i=1}^{n} \theta_i * x_i$$

Here, there are various number of predictors, nearly n predictors are there. For each combination of these predictors, targets values are found and an attempt is made to find the best fitted line. Target value (here Crop yield) is the dependent variable (y) and Predictors (here the crop prediction factors/attributes) are the independent variables, x_i. Summation is being used as there can be 1 to n predictors.

2. Random Forest: In literal sense, it is a cumulative-averaged result of various randomized combinations of decision trees, which are themselves used for prediction and classification problems. Each of the random tress gives us a result and then the whole result is averaged out to get the final result.

$$f_{b=1}^{B} = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$$

Here, B represent the number of trees our random forest consists of and $T_b(x)$ is the function of the bth random tree. This equation will help us in finding a new prediction at point x.

Of these two, we will go with Multivariate Linear Regression Model. Our selection is based on the fact that its equation is easy to understand, plus in the latter many trees will be created consuming more processing power. For longer research it can be considered but for the scope of our project, Multivariate Linear Regression is most suitable for comparing efficiency against LSTM.

After choosing the algorithm, now it is the time to decide the performance metrics which will be used to compare the models' efficiency. Here, we have thought of analyzing the performance based on the following four standard metrics:

1. MAE (Mean Average Error): As the name suggests, it the average of all the errors (difference between predicted and actual value).
2. MSE (Mean Square Error): It is similar to MAE, difference being that here the average is of the square of errors. That is, sum of square of individual errors, divided by the total number of observations.

3. RMSE (Root Mean Squared Error): Mathematically, it is the square root of Mean Square Error. Talking about their physical significance, they tell us how far the data points are from the line of best fit.
4. Accuracy: Simply, it is the number of correct predictions against the total number of predictions

These 4 are the standard metrics widely accepted for any machine learning model. Hence, this is our proposed methodology to proceed on the project.

First of all, talking about the data, we were unable to find any suitable dataset that catered to our need as from all the sources, there was one or more data missing or presence of inconsistencies in large numbers. Also, we thought of combining datasets as well but that was of no use as there was a mismatch in the data units or metadata was absent in a major dataset, hence can't have any knowledge about the units of data in the dataset. So finally, we made our own dataset with the data collected from the website [11]. Before talking about how the dataset was created, we will clarify some of the topics were to be researched upon. So basically, we went on to do the data collection for the region of Tamil Nadu for the paddy crop. The reason to go for paddy crop is the curiosity we had when we found from our surroundings and research that most of the people here eat rice and rice-made products like dosa, idli uttappam and hence, wanted to know whether the climatic conditions of Tamil Nadu are really such that they are able to produce rice (paddy) in so much of amount to cater to the people's need. Also, when we went to fix the time frames, we realized that different types of rice are grown in different times of year, which is like throughout the year some variety is being grown. Hence, we couldn't fix any one period for the crop and went to select all the months of the year, where the attributes will be averaged value over the month and the predictor value (crop yield) will be of the whole year. Here, we are considering financial year. For your reference,

Financial Year 2009 refers to the duration from April 2008 to March 2009.

Hence, for all the months of the financial year duration, we iterated over the months and collected following data to store in excel along with their units:

- Min and Max Temperature: Degree Fahrenheit
- Wind Speed: miles per hour
- Rainfall: mm

We reduced the number of attributes being used due to the unavailability of all the attributes in a single place, and hence, went for the ones which we think are the most important attributes with which we can move forward in our project.

Now talking about the Crop Yield, we got the data from here [12]. Here the data was present for the financial years 2009-2020, so we also have to consider those years only when collecting the climatic conditions data. Also the unit of crop yield is million metric tons. Finally, we aggregated all this data into a single place (an excel spreadsheet) and here is our dataset [13].

Now here a general question arises, can our dataset be trusted? Is it the right way to use synthetic data? To answer them, first of all the data is not synthetic; it is already present on a reputed website. Also, the yield data present on statista can be trusted too and between these two, there is no difference in units or any other parameters. Hence, we were able to combine the data easily and form the dataset. For sure, the dataset could have been better but with resources in hand, this was the best possible result we could have done, comprising of our *novelty in this project. Now finally, we segregated our data into train and test datasets by using ratio 1:1 i.e., 50% of that data that is of 6 years be used as training dataset and remaining as testing dataset.

Hence after the data collection/segregation part comes the implementation of the algorithm and performance metrics. After much discussion, the project was finally built using Python Language but on Google Colab (Colaboratory), a development environment for python that runs on Google Cloud and provides great computing capacity and flexibility to build such projects collaboratively with other peer mates.

Before we can apply any machine learning model, we need to import our Dataset which is in csv format into our colab file by using dataframes from pandas library. After that, we will convert it to arrays as models will accept only array, which will be done using oandas and numpy.

Now talking first about the LSTM model, we decide the number of layers we have to give to them. Generally, the deeper the model is, accurate will be the results provided by it. Taking inspiration from our base paper [1] we go with 4 LSTM layers, in which one layer will provide data to the next one in sequential manner and one output dense layer. After that for each layer we fix the Regularization Technique, which is Dropout here. In this, the nodes/inputs are randomly dropped and forces the network to learn from the remaining data; this is done to avoid overfitting as it may happen in smaller datasets, the subsequent layers may see inconsistencies as errors and may correct them in their prediction turn. The rate of dropout is set to 20% (0.2) and 50 neurons are assigned to each layer to classify information based on the pre-defined architecture. Now in deep learning models, it is necessary to use optimizer techniques which keeps a check on providing learning rates and weighs to layers and epochs. The most efficient one, Adam optimizer is used in our project. The name is derived from Adaptive Moment Estimation and in this the weight is assigned to each component individually with a feature that it utilizes both first and second moment of gradients when giving weights to the epochs. Now epochs and batch size are closely related. Batch Size determine the number of samples which will go for iteration, whereas the number of epochs define the number of times model will train on given dataset. We set epochs to 200 and batch size of 20. This whole model setup is done using Keras Library of python. Also, we will model MLR according to its methodology and predict the yield of crop.
After this, we find all four-performance metrics for our model and print them.
Full code of our project is here [14].

In real-time, our application can be used by the farmers or policy makers to predict the yield of rice crop but with some modifications. As what we have made is a python code, it must be given a form of interface or an application, which can then be used by the people to predict the yield. What it can look like is a specific person assigned by the government for an area can upload the datasets in a specific format and the model will tell the yield prediction. Accordingly, the decision and steps can be taken by both the parties. The format of data should be pre-defined in accordance with the model created for the purpose.

For that, the model is surely in its beginning stage, but with due course of time, we can improve the model efficiency and add more parameters like soil conditions and humidity, depending on the dataset we get.

Also, there is a requirement to have more authentic datasets as discussed above, we were forced to aggregate the data due to absence of datasets for those regions with appropriate attributes. Hence, it also opens up an opportunity for the data engineers and data scientist along with IoT experts, which can work in sync and collect various data using sensors and perform various data pre-processing techniques and form a proper dataset for public usage.

During the course of the project, we ran into a number of challenges, the most significant of which was identifying a dataset that was suitable for our purposes. As a result, we came to the conclusion that we needed to integrate a few different datasets in order to get the dataset we required.

The second challenge was to gain an understanding of neural networks, which is a big subject and, without which, it is difficult to gain an understanding of what we need to do in order to move forward. After looking through a variety of different ideas and publications, we decided that the best algorithm for our needs would be LSTM, which is a modified form of RNN.

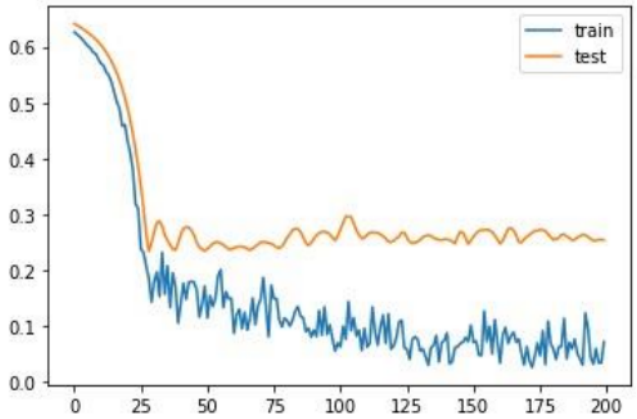
Following this, we went through a large number of tutorials and websites to educate ourselves on Python implementation; despite the fact that there are libraries available, we are unable to simply import everything because a particular methodology needs to be adhered to in order to achieve the best possible results.

Finally, we were able to achieve our result, which represents a key step towards expanding the scope of this project in the future. As was mentioned before, expanding the scope of this project can mean using it as a full-fledged model to assist in different ways..

The scope of the model can be broadened to encompass the prediction of agricultural yields for a range of plant species. Also as mentioned above, the additional elements, such as the level of moisture in the soil and the accessibility of irrigation facilities being taken into account can in turn also help in realizing the amount of water resources available/required for irrigation and to calculate the kinds of nutrients that the soil will require in order to be prepared for the coming year.

Farmers will be able to select crop-rotation plans thanks to our ability to implement functionality for several crops as this will serve dual purpose; growing multiple crops on a piece of land and helping in replenishing soil nutrients by natural methods.

The need of the day is to educate Indian farmers. For a smart India, we also need smart farmers, and eventually, that means educating them. The real secret to the upliftment of the entire nation is concealed in the farmers, who form the foundation of the country. Our project is just an effort in helping them by providing them a way to predict the yield of their crops, which can help both the farmers and the government/policy makers to be prepared for the upcoming farming season.




```
   var1(t-1)  var2(t-1)  var3(t-1)  ...  var47(t)  var48(t)  var49(t)
1   0.547758   0.157142   0.072308  ...  0.615778  0.703412       0.0
2   0.643275   0.448979   0.729231  ...  0.000000  0.540683       0.0
3   0.666667   0.375510   0.435385  ...  0.615778  0.721785       0.0
4   0.992203   0.136735   0.000000  ...  0.081427  0.666667       1.0
5   0.327485   0.000000   0.553846  ...  0.279900  0.771654       0.0

[5 rows x 98 columns]
(6, 1, 97) (6,) (5, 1, 97) (5,)
Epoch 1/200
1/1 - 7s - loss: 0.6263 - val_loss: 0.6417 - 7s/epoch - 7s/step
Epoch 2/200
1/1 - 0s - loss: 0.6224 - val_loss: 0.6388 - 38ms/epoch - 38ms/step
Epoch 3/200
1/1 - 0s - loss: 0.6182 - val_loss: 0.6359 - 34ms/epoch - 34ms/step
Epoch 4/200
1/1 - 0s - loss: 0.6135 - val_loss: 0.6328 - 34ms/epoch - 34ms/step
Epoch 5/200
1/1 - 0s - loss: 0.6077 - val_loss: 0.6295 - 36ms/epoch - 36ms/step
Epoch 6/200
1/1 - 0s - loss: 0.6017 - val_loss: 0.6260 - 37ms/epoch - 37ms/step
Epoch 7/200
1/1 - 0s - loss: 0.5979 - val_loss: 0.6222 - 34ms/epoch - 34ms/step
Epoch 8/200
1/1 - 0s - loss: 0.5908 - val_loss: 0.6181 - 36ms/epoch - 36ms/step
Epoch 9/200
```

```
Epoch 195/200
1/1 - 0s - loss: 0.0448 - val_loss: 0.2560 - 44ms/epoch - 44ms/step
Epoch 196/200
1/1 - 0s - loss: 0.0324 - val_loss: 0.2534 - 42ms/epoch - 42ms/step
Epoch 197/200
1/1 - 0s - loss: 0.0603 - val_loss: 0.2531 - 41ms/epoch - 41ms/step
Epoch 198/200
1/1 - 0s - loss: 0.0351 - val_loss: 0.2552 - 40ms/epoch - 40ms/step
Epoch 199/200
1/1 - 0s - loss: 0.0347 - val_loss: 0.2553 - 38ms/epoch - 38ms/step
Epoch 200/200
1/1 - 0s - loss: 0.0723 - val_loss: 0.2545 - 31ms/epoch - 31ms/step
1/1 [==============================] - 1s 1s/step
[0.5477582  0.6432749  0.6666666  0.99220276 0.3274854  0.5808966 ] [[0.65497077]
 [1.        ]
 [0.        ]
 [0.8323586 ]
 [0.7329434 ]]
Test RMSE: 0.355
```

During the review 2, we were given some points to discuss upon and I would like to present the answer/solution/explanation to each of them.

1. Architecture: For that, we have made a flowchart showcasing the whole project process.
2. Novelty: Stating the fact that there is no proper dataset already available, the fact that we collected the data and aggregated it to form a dataset is itself a novelty. Due to time and resource constraint we were unable to do the same using sensors, which we have included in our future scope.
3. Performance Metrics: We have used 4 standard performance metrics: MAE, MSE, RMSE and Accuracy.
4. Comparison: We have implemented Multivariate Linear Regression for comparing with LSTM.
5. Visualization: Some visualizations are present above in Appendix 1.
6. Synthetic Dataset: We have justified out dataset creation above sir under Implementation

002 Nilesh Team

ORIGINALITY REPORT

11%	6%	7%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	3%
2	Daniel Jiménez, Sylvain Delerce, Hugo Dorado, James Cock, Luis Armando Muñoz, Alejandro Agamez, Andy Jarvis. "A scalable scheme to implement data-driven agriculture for small-scale farmers", Global Food Security, 2019 Publication	1%
3	Submitted to University of Wales Institute, Cardiff Student Paper	1%
4	research.wur.nl Internet Source	1%
5	Submitted to Higher Education Commission Pakistan Student Paper	1%
6	louisdl.louislibraries.org Internet Source	<1%
7	ziraat.ahievran.edu.tr Internet Source	<1%

8	Submitted to University of Essex Student Paper	<1 %
9	Submitted to University of Stellenbosch, South Africa Student Paper	<1 %
10	Dhivya Elavarasan, Durai Raj Vincent P M. "Crop Yield Prediction using Deep Reinforcement Learning Model for Sustainable Agrarian Applications", IEEE Access, 2020 Publication	<1 %
11	Submitted to Queen Mary and Westfield College Student Paper	<1 %
12	Nishu Bali, Anshu Singla. "Deep Learning Based Wheat Crop Yield Prediction Model in Punjab Region of North India", Applied Artificial Intelligence, 2021 Publication	<1 %
13	Dilli Paudel, Hendrik Boogaard, Allard de Wit, Sander Janssen, Sjoukje Osinga, Christos Pylianidis, Ioannis N. Athanasiadis. "Machine learning for large-scale crop yield forecasting", Agricultural Systems, 2020 Publication	<1 %
14	vtechworks.lib.vt.edu Internet Source	<1 %

15	"Artificial Intelligent Techniques for Wireless Communication and Networking", Wiley, 2022 Publication	<1 %
16	mdpi-res.com Internet Source	<1 %
17	Submitted to National University of Ireland, Galway Student Paper	<1 %
18	publik.tuwien.ac.at Internet Source	<1 %
19	Submitted to Liverpool John Moores University Student Paper	<1 %
20	Submitted to The British College Student Paper	<1 %
21	Submitted to Harper Adams University College Student Paper	<1 %
22	Lontsi Saadio Cedric, Wilfried Yves Hamilton Adoni, Rubby Aworka, Jérémie Thouakesseh Zoueu et al. "Crops Yield Prediction Based on Machine Learning Models: Case of West African Countries", Smart Agricultural Technology, 2022 Publication	<1 %
23	www.ncbi.nlm.nih.gov Internet Source	

<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		