Analysis of SVM and RNN-LSTM on Crop Datasets

Kusum Lata[1], Sajidullah S. Khan[2], Onkar Kemkar[3]
[1, 2, 3]*School of Computer Sciences and Engineering,*
Sandip University, Nasik, Maharashtra, India
E-mail: [1]kusumrao444@gmail.com, [2]sajidullah.khan@sandipuniversity.edu.in,
[3]onkar.kemkar@sandipuniversity.edu.in

***Abstract*—The increasing population is raising the concern for food security due to limited agricultural resources. The technical advancements in the agricultural domain have strengthen the crop management. The crop yield prediction with machine learning techniques plays a role for sustainable agriculture and will help the farmers. The main motive of this paper is to compare two machine learning techniques namely support vector machine (SVM) and recurrent neural network (RNN) with long short-term memory (LSTM). The techniques are evaluated in terms of percentage in prediction accuracy. The models are framed on a single historical dataset obtained from Indian government website and a research station where the influencing attributes are temperature, rainfall, humidity etc. The training and test set are split in the ratio70:30 whereas the training is done in 10 folds for cross validation. The outcome the analysis exhibit that the SVM model has marginally higher accuracy when compared with RNN-LSTM in the instant case. However, the results are dependent on the implementation when dealing with real-world application.**

Keywords: SVM, RNN-LSTM, Agriculture, Crop Yield.

I. Introduction

Agriculture domain is one of the main research areas as this is an important source of income for masses including farmers. The population of the world is increasing day by day and the demand of food will also increase accordingly. The agricultural land and other resources are very limited to fulfil the demand. In India approximately seventy percent of the total population is living in villages and agriculture is a form employment for the farmers. The future of the farmers depends on the agricultural output. But the farmers are not achieving the expected crop yield due to various issues in the agricultural domain. Government has launched various web portals and apps in favor of farmers in order to educate them and to resolve their issues. Similarly, machine learning is a potential technique and can be applied in crop disease detection, weather forecasting and crop yield prediction etc. The crop yield predictions in advance will strengthen the farmer's community and further minimize the losses. In this paper we will explore and compare the two machine learning techniques Support Vector Machine (SVM) and Recurrent Neural Network (RNN-LSTM). We will explore that which classifier has accurate predictions using a historical crop dataset. The machine learning classifiers SVM, RNN-LSTM will be compared. The basis of the paper is to use existing algorithms with the help of python and available libraries and not to develop a new model for the implementation. The analysis is restricted to the two models for clear implementation and evaluation. The dataset training and testing are performed on a single historical crop dataset. Crop yield prediction mostly depends on the weather and non-weather parameters. This paper is categorized in various sections. Accordingly, Section 2 describes the background of machine learning and Section 3 is about the literature survey. The implementation is covered in Section 4 and Section 5 describes the results and discussion. Section 6 concludes the work with future improvements.

II. Background

A. Machine Learning

Machine learning is a field in computer science evolved from artificial intelligence. It is subset of artificial intelligence. In machine learning computer programs are formulated to improve and learn from experience to make future decisions. It can make predictions based on the input data. Machine learning has variety of applications are below [11]:

1. Bioinformatics
2. Computational finance
3. Computer vision
4. credit card fraud detection
5. Information retrieval
6. Internet fraud detection
7. Medical diagnosis
8. Recommendation systems
9. Sentiment analysis
10. crop health monitoring
11. Crop yield prediction

ISBN: 978-93-91355-11-1

B. Types of Machine Learning

Machine learning mainly categorized into four broad categories based on the nature of the data. These are described as below and illustrated in Fig.1:

Supervised Learning: In this type of learning, there is a relationship between input and output and the output is known based on the given data set. The outcome variable i.e., dependent variable is predicted from a given set of independent variables [12]. A function is generated using the set of variables that map inputs to desired outputs. This training process is a continuous activity to achieve the desired accuracy. Different types of supervised learning are Classification, Regression, Random Forest model, Naive Bayesian model Support vector machines and Neural networks.

Unsupervised Learning: In this type of learning there is no relationship between input and output and there is no idea about the output. In this type, the outcome variable which has to be predicted is not known. The effect of the variables may be undesired. The main difference between supervised learning and unsupervised learning is that in a supervised learning input and output variables are known, while in an unsupervised learning, only input variables are known. Different types of unsupervised learning are Apriori algorithm, independent component analysis Principal component analysis Neural networks Anomaly detection Hierarchal clustering KNN (k-nearest neighbors) and K-means clustering.

Semi supervised Learning: In this type of learning, the input dataset is a blend of labeled and unlabeled data. This dataset has a very small amount of labeled data and a very large amount of unlabeled data. Unsupervised learning is applied to cluster similar data and after that the labeled data will be used to label the balance unlabeled data. Semi-supervised learning pre-assumes at least one of the assumptions from continuity assumption, cluster assumption and manifold assumption

Reinforcement Learning: This type of learning is used to train the machine to make certain decisions. The model trains itself based on trial-and-error basis. The learning takes place from past experience to build the knowledge base to take accurate decisions. Different types of reinforcement learning are: Positive, Negative and can be executed by three ways i.e., Model-Based, Policy-based and Value-Based.

Other various types of machine learning are Self-Supervised Learning, Multi- Instance Learning, Deductive Inference, Inductive Learning, Transudative Learning, Active Learning, Transfer Learning, Multi-Task Learning, Ensemble Learning and Online Learning etc.

C. Support Vector Machines

Support vector machines is a supervised machine learning technique which is mainly used for classification and regression problems. The SVM algorithm was pioneered by Vapnik and. Chervonenkis has roots in Statistical Learning Theory. It is about learning structure from given data and is capable to handle multiple continuous and categorical variables. SVM model represent different classes in a hyperplane in multidimensional space. The aim of SVM is to categorize a dataset into different classes to find out the maximum marginal hyperplane (MMH). Simple SVM model is illustrated in Fig.2

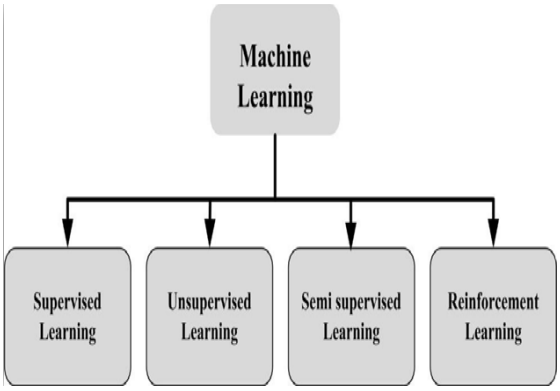

Fig. 1: Machine Learning Types

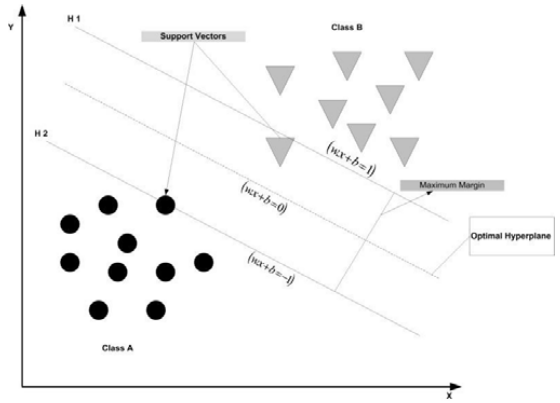

Fig..2: Simple SVM Model

SVM work with kernel functions to transform an input dataset into the desired form. The kernel trick technique is also used in SVM for non-linear problems with the help of slack variables and adding more dimension and transformed to a higher dimensional space. The following type of kernels are used in SVM as described in Table 1.

Table 1: Types of SVM Kernels

Kernel Type	Equation
Linear	$k(x, y) = x^T y + c$
Polynomial	$k(x,y) = (ax^T y + c)^2$
Radial Basis Function (RBF)	$k(x,y) = exp(-\frlVert x - y \rVert^2 / 2\sigma^2)$
Sigmoid	$k(x,y) = \tanh(ax^T y + c)$
Laplacian kernel	$k(x,y) = exp(-\lVert x - y \rVert / \sigma)$

SVM is preferred over other machine learning algorithms due to below reasons:

1. High Dimensionality
2. High Accuracy
3. Less Over fitting
4. Less Memory Requirement
5. Simple SVM model Training.
6. Low Classifier complexity with minimal error
7. Suitable for nonlinear data
8. Guaranteed Optimality
9. Useful in regression problems (Support Vector Regression)
10. Flexibility due to kernels

D. Recurrent Neural Network

Recurrent neural networks (RNNs) are acknowledged as recurrent because the algorithm function repeatedly with output dependent on previous computations. The computations are stored in memory as RNNs have memory feature. These computations are further used as input to the model. RNN are simply extension of artificial neural network (ANN) [5]. The main difference between RNN and feedforward neural networks (FFNN) is that in RNN the data moves in both directions. The output in RNN is depend on current input and previous prediction results. There are four topologies of RNN i.e., one to one, one to many, many to one and many to many. The input and output sequences of RNN are of four types i.e., sequence-to-sequence, sequence-to-vector network., vector-to-sequence network and vector-to-vector network.

The RNN suffers from vanishing gradients problems i.e., problem in learning long sequences. There are many techniques to address the vanishing gradient but the most capable and attracting solution for vanishing gradient problem is long short- term memory (LSTM).

Long Short-Term Memory (LSTM): LSTM pioneered by Schmidhuber and Hochreiter in 1997. It is a kind of RNN which is capable of learning long-term dependencies. LSTM networks are specially designed to remember the information for long periods to avoid the long-term dependency problem. LSTMs also have this chain like structure like RNN, but instead of having a single layer there are four layers which are working very carefully. The key element of LSTM is cell state. The cell state is like a conveyor belt system and managed by structures called gates. LSTM has three gates to control the cell state. Each LSTM module may have three gates named as forget gate, input gate, output gate. The complete RNN and LSTM model is illustrated in Fig 3.

Forget Gate: It is determined via the sigmoid function.

Input Gate: It is determined by the sigmoid and Tanh function.

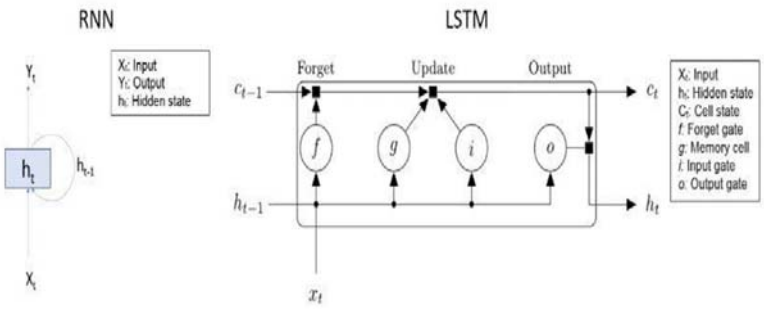

Fig..3: RNN and LSTM Model [5]

ISBN: 978-93-91355-11-1

Output Gate: It is also determined by the sigmoid and Tanh function.

1. Advantages of RNN-LSTM:

1. Handling input of any length
2. No impact on Model size with increasing size of input
3. Applicable to historical information and data
4. Weights are shared across time
5. LSTM can handle distributed representation and continuous values.
6. No need of parameter tuning
7. LSTM possess non-Markovian behavior

III. Literature Review

K. Aditya Shastry et al. proposed a cloud based agricultural framework for soil classification and crop yield prediction as a service. This designed cloud framework provides the two services i.e., soil classification and crop yield prediction as a service. The soil classification classifier used the hybrid support vector machine (SVM), the kernel parameters for SVM are obtained from Genetic algorithm (GA). The crop yield prediction used the customized artificial neural network (ANN) were established where hidden layers, neurons and learning are modified. Therefore, the authors have concluded that this framework will provide the reliable and precise services to the end users. For future, the authors have decided to extend this work by developing mobile agricultural application with sophisticated features [6].

Igor Oliveira et al., proposed a scalable machine learning system for pre- season agriculture yield forecast. In this research, the authors used the real time soil data and seasonal climate forecasting dataset to predict Soyabean and Maize yield for Brazil and USA. The dynamic weather dataset is handled by the recurrent long short-term memory (LSTM) layers although the static soil dataset is handled by fully connected layers. Deep neural network (DNN) is used in this work as a model and implemented in Keras. The model shows better prediction accuracy for Brazil Soyabean with RMSE of 385. They proposed a machine learning system that offers pre-season yield forecasting so that farmers can make easily farm management decisions [7].

Sonal Agarwal et al., proposed a hybrid approach for crop yield prediction using machine learning and deep learning algorithms to predict the best crop production. This approach will analyze the given data and help the farmers in forecasting a crop to increase the revenues. The authors have collected the data from kaggle.com and this dataset contains the various parameters such as temperature, rainfall, pH value, relative humidity and also an area. Firstly, the authors have applied artificial neural network (ANN) and random forest (RF) algorithms for a set of crops and find out the accuracy i.e., 93%. Secondly, they have applied long short-term memory (LSTM), support vector machine (SVM) and recurrent neural network (RNN) and find out the accuracy i.e., 97%. Therefore, the authors have concluded that the use of both machine learning algorithm and deep learning algorithm plays an essential role in predicting the improved crop yield with upgraded accuracy [8].

IV. Implementation

İn this section the analysis of SVM and RNN-LSTM is performed on crop datasets to predict the suitable crop for a particular region is ensuring high crop yield based on weather parameters such as temperature, humidity and rainfall. The study area selected for this research is Nashik District of Maharashtra, India. Nashik is known for agriculture and leading in production of various crops like Arhar, bajra, caster seed, cotton, groundnut, maize, moong, Niger seed, ragi, rice, soyabean, sunflower, safflower, gram, jowar, wheat, sugarcane, grapes, tomatoes, onions and raisins etc. The farmers of Nashik are experimenting with crops to maximize the output and this study may help to recommend the crops according to climatic conditions [1].

A. Dataset Description

The dataset used in the implementation is obtained from government web portal https://data.gov.in and also collected from agriculture research station situated at Pimpalgaon Baswant, Dist. Nashik, Maharashtra. The agricultural datasets contain information like production area, crop, year, yield, temperature, humidity and Rainfall. The dataset consists of 7 columns that represent the unique attributes. The dataset contains data for time frame of two years i.e., 2018,2019. The dataset parameters are normalized in the range of 0–1 for analysis of the machine learning techniques. It is preferable to evaluate machine learning techniques in context of labels, normalization and size etc.

B. SVM

SVM technique is implemented in python programming language using the scikit- learn library for machine learning [8]. The SVM implementation with crop dataset involves the four broad steps as follows:

1. Importing scikit-learn library and data set
2. Define Support Vector Classifier (SVC) method
3. Kernel selection and parameter setting
4. Accuracy prediction and comparison

The crop datasets are generally nonlinear in nature; hence the radial basis function (rbf) kernel is preferred over the other type of kernels. In various research works it has the best prediction accuracy when compared to linear, polynomial kernel functions and has the minimum

runtime. In rbf kernel the parameter C set to 1 and the gamma set to 0.1 to avoid the overfitting due to more curve decision. Prediction accuracy can be revamped by tuning the parameters using available function like GridSearchCV () in scikit-learn. The SVM will be implemented with a 10-fold cross validation.

C. RNN-LSTM

RNN Long Short-Term Memory (LSTM) also implemented in python programming language. The implementation involves the following steps:

1. Dataset preparation for LSTM:
   a. Prepare the data as per requirement
   b. Feature Scaling (Preprocessing of data)
   c. Split the dataset for train and test
2. LSTM model development
3. The LSTM model development layer consider 3 inputs i.e., Samples, Time steps and Features.
4. Compile the model i.e., Training
5. Accuracy prediction of model i.e., Test and compare with another model.

During implementation of the model the discussion about loss function and optimization algorithm is important. Mean Absolute Error (MAE) is taken as loss function and Adaptive Moment Estimation (ADAM) is taken as optimization algorithm. It will find out a suitable learning rate for the model to predict the output. Adam is a replacement optimization algorithm for stochastic gradient descent for training deep learning models [9]. The visualization of the implementation of the above two machine learning models is shown in Fig 3.

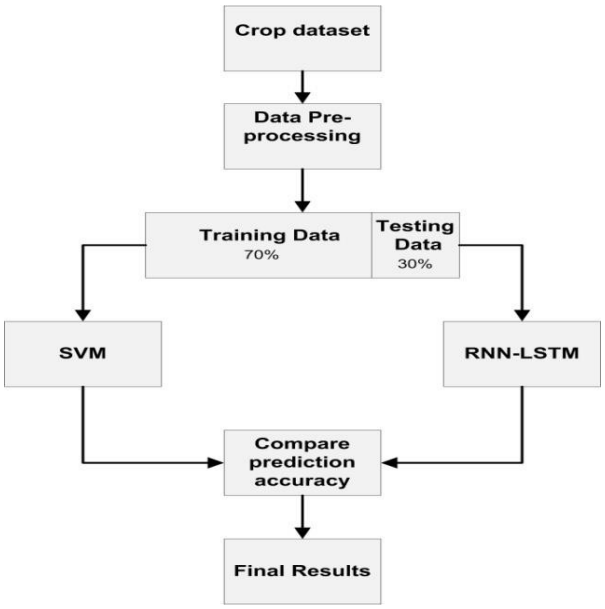

Fig. 3: Model Implementation Process

D. Evaluation parameters

The performance of the proposed model is measured to obtain the result and compare. There are certain equations that are used to get the accuracy of the model are given in Table 2.

Table 2. Evaluation Parameters

Metric	Equation
Average Prediction Accuracy (Avg)	$Avg = \frac{1}{n}\sum_{i=1}^{n} P.A_i$
Standard Deviation (SD)	$SD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P.A_i - Avg)^2}$
Standard Error (SEM)	$SEM = \frac{SD}{\sqrt{n}}$
Maximum Prediction Accuracy (MaxPA)	Highest value
Minimum Prediction Accuracy (MinPA)	Lowest value

V. Results and Discussion

The result is measured in terms of validation and prediction accuracy of both the machine learning techniques [2][3]. The average accuracy of the model is considered over the 10-fold cross validation and further standard deviation, Standard error of the mean of average prediction accuracy, maximum prediction accuracy and minimum prediction accuracy are calculated. The results are shown in the below mentioned in Table 3:

Table 3. Validation and Prediction Accuracy

Cross Fold	SVM		RNN-LSTM	
	Validation accuracy	Prediction accuracy	Validation accuracy	Prediction accuracy
1	0.756	0.775	0.655	0.678
2	0.751	0.791	0.643	0.667
3	0.743	0.765	0.687	0.732
4	0.721	0.742	0.665	0.689
5	0.71	0.755	0.675	0.745
6	0.767	0.783	0.689	0.721
7	0.773	0.792	0.624	0.678
8	0.734	0.769	0.648	0.682
9	0.765	0.783	0.677	0.712
10	0.789	0.805	0.679	0.718

The first column is showing the cross fold for each fold, second column showing the validation accuracy calculated during each training phase and the third column shows the prediction accuracy calculated during the testing phase of the analysis. The comparison of both

ISBN: 978-93-91355-11-1

models in terms of prediction accuracy are graphically shown as below in Fig. 4.

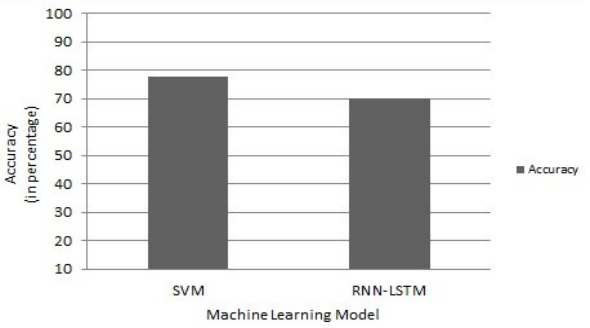

Fig. 4. Accuracy Comparison

The final results regarding prediction accuracy are shown in Table 4 mentioned below.

Table 4. Parameter Wise Results

Parameter	SVM	RNN-LSTM
APA	0.776	0.702
SD	0.017	0.025
SEM	0.005	0.008
MaxPA	0.805	0.745
MinPA	0.742	0.667

The results from above table depicts that there is a slight difference in precision accuracy between the SVM and RNN-LSTM. The SVM had a higher accuracy in average, maximum and minimum prediction accuracy of the models, as well as a lower SD and SEM. The results concludes that SVM is better when dealing with crop datasets. The accuracy results for both models may vary based on the dataset selection. This indicates the possibility of improvements for both the models.

VI. Conclusion

In this paper it is observed that we can use machine learning to crop datasets and further yield prediction. The parameters like rainfall, soil, temperature, humidity etc. are elements of datasets. The result shows that historical data has influence to predictions and results of SVM model is better when compared to RNN-LSTM in terms of average prediction accuracy. The accuracy of SVM models is recorded 78% approx. while the accuracy of RNN-LSTM is 70%. These machine learning approaches are capable and can be utilized according the nature of data. The outcome of this work will help the farmers for crop prediction and crop yield prediction in advance to plan the pre and post harvesting agricultural activities. In the future, a more specific crop prediction model will be developed using hybrid model consisting of SVM and RNN-LSTM, with graphical user interface to handle the complex and large datasets.

References

[07] Fegade, T.K.; Pawar, B.V. Crop Prediction Using Artificial Neural Network and Support Vector Machine. In Data Management, Analytics and Innovation; Sharma, N., Chakrabarti, A., Balas, V.E., Eds.; Advances in Intelligent Systems and Computing; Springer: Singapore, Volume 1016, pp. 311–324, 2020.

[08] An introduction to Machine Learning with SciKit Learn. https://scikit-learn.org/

[09] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg Scikit-learn: Machine learning in python. Journal of machine learning research, vol. 12, pp. 2825–2830, Oct 2011.

[10] Sujatha, R., & Isakki, P: A study on crop yield forecasting using classification techniques, In International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE) (pp. 1–4) 2016.

[11] Recurrent Neural Network (RNN). https://in.mathworks.com/discovery/rnn.html

[12] K. Aditya Shastry and H. A. Sanjay: Cloud-Based Agricultural Framework for Soil Classification and Crop Yield Prediction as a Service. Emerging Research in Computing, Information, Communication and Applications, Advances in Intelligent Systems and Computing 882.

[13] Igor Oliveira, Renato L. F. Cunha, Bruno Silva, Marco A. S. Netto: A Scalable Machine Learning System for Pre-Season Agriculture Yield Forecast, 2018.

[14] Sonal Agarwal and Sandhya Tarar: A hybrid approach for crop yield prediction using machine learning and deep learning algorithms, J. Phys.: Conf. Ser. 1714 012012, 2021.

[15] Dhivya B, Manjula, Bharathi S and Madhumathi March: A Survey on Crop Yield Prediction based on Agricultural Data International Conference in Modern Science and Engineering 2017.

[16] Gandhi, N., Petkar, O., & Armstrong, L. J: Rice crop yield prediction using artificial neural networks. In IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR) 2016.

[17] Subhadra Mishra, Debahuti Mishra, Gour Hari Santra: Applications of machine learning techniques in agricultural crop production: a review paper. Indian Journal of Science and Technology. 2016..

[18] Rakesh Kumar, M.P. Singh, Prabhat Kumar, J.P. Singh: Crop Selection Method to maximize crop yield rate using machine learning technique. International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM),27 August 2015.

[19] Aakunuri Manjula, Dr. G. Narsimha: XCYPF: A Flexible and Extensible Framework for Agricultural Crop Yield Prediction. Conference on Intelligent Systems and Control (ISCO), 2015.

[20] Renuka. Sujata Terdal: Evaluation of Machine Learning Algorithms for Crop Yield Prediction, 2019.