[1]D Jayanarayana Reddy

[2]**Dr** M Rudra Kumar

Crop Yield Estimation using Improved Salp Swarm Algorithm based Feature Selection

Abstract: **-** Crop yield estimation is the art of yield prediction before harvest and it is essential for planning and making conclusive agricultural policies. The forecasting of crop yield is essential in optimal nutrient management, crop insurance, crop market planning and harvest management. However, the crop yield estimation is considered as a challenging task because of huge amount of abundant information exists in the crop data. Therefore, an effective feature selection is required to be developed for removing the redundant attributes. In this research, an Improved Salp Swarm Algorithm (ISSA) based feature selection for an effective crop yield estimation. The Opposition Based Learning (OBL) and Local Search Algorithm (LSA) are incorporated in the ISSA's initialization and exploitation phase for selecting optimum feature subset. The selected features from the ISSA are used to enhance the classification using Modified Long Short Term Memory (MLSTM) classifier. The performance of the ISSA-MLSTM is analyzed using accuracy, precision, recall, F-score, Nash-Sutcliffe Efficiency Coefficient (NSEC), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The existing researches such as Ensemble approach and MLSTM are used to evaluate the ISSA-MLSTM. The accuracy of the ISSA-MLSTM is 99% that is high when compared to the MLSTM.

Keywords: Crop yield estimation, feature selection, improved salp swarm algorithm, local search algorithm, modified long short term memory, opposition based learning.

I. INTRODUCTION

Agriculture is the primary occupation of India and the economy of country is totally based on it for rural survival [1] [2]. The agriculture field is considered as a primary for country development because of the direct or indirect dependence on the huge amount of farmers, private companies, public sectors and middlemen. Agriculture is expected to be a gainful, only when there is a good crop year returns with high yield that generally offers remunerative prices [3]. The production of crop is sensitive to climate. Crop is vulnerable to different parameters such as inter-annual climate variability, average rainfall and temperature, dangerous weather events and shocks during specific phenological stages [4] [5] [6]. Therefore, a prior estimation of crop yield before the actual crop harvesting is mandatory to ensure the appropriate planning and policy making in agricultural field, mostly in the situation of climate change. This crop yield forecasting is used to discover the food crops distribution and additionally helps to take tactical decision in price fixing and import/ export of food stock of the country [7] [8] [9]. Crop yield prediction is essential but complex issue which is mandatory for supportable intensification and effective utilization of natural resources [10].

The crop yield data is progressively utilized for evaluating the carbon and nitrogen cycles, agricultural productivity potential, greenhouse gas emissions and the effect of changes in climate over the agricultural production [11]. The estimation of crop yield in various spatial levels is provided en extra value when it is accessible at small units or higher spatial resolutions. A consistent forecast in higher spatial resolution helps to detail the changes in yield in coarser levels as well as it offers the information for adapting the agricultural policies to certain areas [12] [13]. The agricultural crop is varied with time and requirement to discover the crop quantity and identify the shortage gains the supreme prominence [14]. An estimation of yield is difficult because of the complex relations among the crop growth and yield-influencing natural factors such as disease, soil conditions, weather, and anthropogenic factors includes rotation, tillage, irrigation, seed varieties and fertilizers [15]. Moreover, the estimation of crop yield is affected by abundant information. Hence, the feature selection is utilized for minimizing the data redundancy for obtaining higher reliable estimation. The feature selection approach is not only eliminate the dimensionality curse which also maintains the original data attributes for achieving the results highly interpretable [16].

The contributions are concise as follows:

- The ISSA based feature selection is developed for eliminating the irrelevant features from the overall data set. The incorporation of OBL and LSA improves the conventional SSA for achieving the improved population diversity and improved exploitation which helps to choose the optimum features.

[1] Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Anantapur,Ananthapuramu, india,djnreddy@gmail.com

[2] Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology (Autonomous), Hyderabad, India

- Next, the MLSTM used with Huber Loss Function (HLF) and Adam optimizer for achieving the better classification. The integration of HLF and adam optimizer used to minimize the MAE and MSE.

The remaining paper is sorted as follows: Section 2 provides the related work about crop yield estimation. The preliminaries used in the ISSA-MLSTM is given in section 3. The proposed ISSA-MLSTM method is detailed in section 4 whereas the results are provided in section 5. Finally, the conclusion is provided in section 6.

2. Related work:

This section provides the information about the related works about crop yield estimation with its advantages and limitations.

Elavarasan, D. and Vincent, P.D.R [17] presented the hybrid regression-based algorithm based on the random forest and reinforcement learning to predict the crop yield. This hybrid regression was operated the reinforcement learning for an each selection of splitting characteristic between the construction of trees. The variable significance measure was analyzed for choosing the significant variable for node splitting process during the development of model and helps an effective usage of training data. The issue of over fitting was avoided and less parameter tuning was obtained by using the internal cross-validation.

Verma, A.K et al. [18] developed the statistical models according to different weight values for forecasting the sugarcane yield. The development of statistical model was used different weighted and unweighted weather indices Similarly, the Nihar, A et al. [19] presented the machine learning regression approaches to discover the district wise sugarcane yield. Here, four machine learning approaches such as Gradient Boosting Regression (GBR), Random Forest (RF), Support Vector Regression (SVR), eXtreme gradient boosting regression (XGB) were utilized as ensemble approach for predicting the yield of sugarcane in district-wise.The processing of input data was mandatory for achieving a better prediction.

Gavahi, K et al. [20] analyzed the different wrapper feature selection methods in crop prediction. The different wrapper approaches such as Recursive Feature Elimination (RFE), Boruta, and Sequential Forward Feature Selection (SFFS) were considered for selecting the features. Here, the supervised learning was used during classification for handling the high-dimensional data.

Shafiee, S et al. [21] examined the efficiency of Support Vector Regression (SVR) with Sequential Forward Selection (SFS) in grain yield prediction. In SVR, an adequate kernel function was determined by using grid search. The SFS was a family of greedy search which used to choose the feature with higher accuracy. Therefore, the SFS was used to minimize the dimension of features. A different types of natural parameters was required to be considered for enhancing the prediction.

Iniyan, S. and Jebakumar, R [22] developed mutual information based enhanced ensemble regression for performing the crop yield prediction. The Mutual information based feature selection was the power statistical approach for discovering the feature relationship among the datasets. The classification or regression issue was supported by minimizing the input size of data.

Dwaram, J.R. and Madapuri, R.K [23] used the real time data for forecasting the crop yield and the min-max normalization was used to preserve the relationship among the collected data. The Long Short Term Memory (LSTM) with HLF and Adam optimizer was used to perform the classification using normalized data. This modified LSTM has adaptive learning rates for different parameters by discovering the 1st and 2nd-moment gradient evaluation for enhancing the efficiency. However, the processing of all features from the normalization was affected the classification.

3. Preliminaries

The approaches of SSA, Opposition Based Learning (OBL) and Local Search Algorithm (LSA) used in ISSA for feature selection are explained in the following sections.

3.1. Process of SSA

Generally, the SSA [24] is motivated by the salps actions that belongs to the family of Salpidae and it has a transparent barrel-shaped body. The SSA population is divided in to 2 types such as leader and followers for achieving the slap chain's mathematical model. The salp exist at salp chain front is denoted as leader and the rest is denoted as followers. The salp denotes that the leader guides the swarm and the followers.

The location of salp is denoted in the n-dimensional search space, where n is amount of variables. The y is matrix with 2 dimensions for saving the salp positions. The food source of salp at search space is F that is taken as swarm's target. Equation (1) expresses the location update for salp leader.

$$y_j^1 = \begin{cases} F_j + c_1\left(\left(ub_j - lb_j\right)c_2 + lb_j\right) & c_3 \geq 0.5 \\ F_j - c_1\left(\left(ub_j - lb_j\right)c_2 + lb_j\right) & c_3 < 0.5 \end{cases} \quad (1)$$

Where, the salp location 1 of jth dimension is denoted as y_j^1; F_j is the food source of jth dimension; the lower and upper bound of dimension j are denoted as lb_j and ub_j respectively; the generated random numbers are denoted as c_1, c_2 and c_3.

Equation (1) express that the leader updated the position based on food source. The c_1 of equation (2) is key value which used to balance the exploration and exploitation of the salp.

$$c_1 = 2e^{-\left(\frac{4r}{r_{max}}\right)^2} \quad (2)$$

Where, the current and maximum iteration are denoted as r and r_{max}. The values of c_2 and c_3 are random numbers which is uniformly created between $[0,1]$. Equation (3) expresses the location update for followers.

$$y_j^i = \frac{1}{2}\left(y_j^i + y_j^{i-1}\right) \quad (3)$$

Where, $i \geq 2$, y_j^i is the follower i's location at jth dimension. The process of SSA except initialization is continued until the execution reaches the maximum iteration.

3.2. Process of OBL

OBL denotes the optimizing that used to enhance the initialized populations quality by expanding the solutions. In the search space, the OBL is worked by searching in the both the directions. The two directions includes original solution and the remaining is opposite solution. Further, the OBL discovers the optimum solution from all solution.

- Opposite number: y is defined as the real value in the range of $y \in [lb, ub]$ whereas the opposite value is denoted as $\tilde{y}$ which is identified using equation (4).

$$\tilde{y} = lb + ub - y \quad (4)$$

The aforementioned equation (4) is generalized for applying it over the search space with multidimensions. This generalization is achieved in each search-agent location and its opposite position that is denoted by equations (5) and (6).

$$y = [y_1, y_2, \dots, y_D] \quad (5)$$

$$\tilde{y} = [\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_D] \quad (6)$$

Equation (7) is used to identify the values of overall elements in $\tilde{y}$.

$$\tilde{y} = lb_j + ub_j - x_j \qquad j = 1,2,3, \dots D \quad (7)$$

- Optimization using opposite population: Here, the $f(.)$ is considered as fitness function. Hence, if $f(\tilde{y})$ is greater than $f(y)$, then $y = \tilde{y}$; otherwise, $y = y$.
- Process of OBL incorporation in SSA is mentioned below:

1. The salp locations are initialized as Y as y_i where $(i = 1,2, \dots m)$
2. The opposite locations are determined as OY as $\tilde{y}_i$.
3. Choose the m optimum salps from $\{Y \cup OY\}$ and it is denoted as new population of SSA.

3.3. Process of LSA

For an each iteration completion of SSA, the LSA is called for improving the current F value. The F from SSA is stored as $Temp$ for each iteration completion and the LSA is iterated for number of times for enhancing the $Temp$. For each iteration of LSA, three random features are chosen from $Temp$. The setting or resetting the chosen features are done based on LSA values. Additionally, the LSA identifies the new solution's fitness value, if new solution's fitness is better than F, then F is fixed as $Temp$; Otherwise, F is unchanged in LSA.

4. ISSA-MLSTM method

In this research, the ISSA based feature selection is developed for removing the irrelevant features which helps to improve the classification. The important phases of ISSA-MLSTM method are given as follows: Dataset acquisition, MMN based pre-processing, ISSA based feature selection and classification using MSTM. Further, the incorporation of Adam optimizer and HLF is used to minimize the MAE in classification. Figure 1 shows the block diagram of ISSA-MLSTM method.

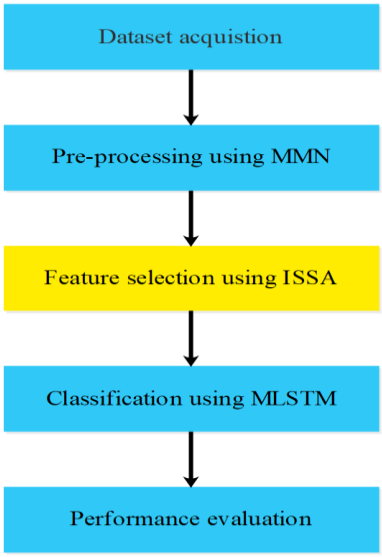

Figure 1. Block diagram of ISSA-MLSTM method

4.1. Dataset acquisition and preprocessing

At first, the data is obtained from the internet sources whereas 7 characteristics are considered such as minimum temperature, average temperature, maximum temperature, cloud cover. precipitation, rainfall and velocity potential to perform the prediction of crop yield. Next, the Min-Max Normalization (MMN) is used to pre-process the raw data which leads to understand the data characteristics [25]. This normalization is used to scale the agricultural data between the specific lower and upper bounds. Hence, the data is scaled between the range of -1 to 1, and 0 to 1 by using the equation (8).

$$x'_{i,n} = \frac{x_{i,n} - min(x_i)}{max(x_i) - min(x_i)}(nMax - nMin) + nMin \quad (8)$$

Where, the lower and upper bound to scale the collected raw data are denoted as $nMin$ and $nMax$; minimum and maximum value of attribute i are denoted as min and max. Here, the $[0, 1]$ $(MMN0)$ and $[-1, 1]$ $(MMN1)$ scales are considered for evaluating the classification.

4.2. Feature selection using ISSA

The preprocessed features are given as input to this ISSA for choosing best feature subset. There are two improvements are incorporated in the conventional SSA: 1) The population diversity is enhanced by using the OBL strategy in initialization, and 2) the exploitation is improved and SSA is avoided from stuck in local optima using LSA. In ISSA, the salps are generated by using OBL. A m amount of appropriate salps are taken from initial and opposite salps locations. Moreover, the optimum salp between the m appropriate salps is fixed as F. Next, the main loop is applied in m salps for updating the locations based in equation (1) or (3). Further, the LSA is used for verifying and discovering the best outcome, hence the ISSA provides the best feature set.

The developed ISSA is operated with KNN using wrapper mode to perform the feature selection. For an each iteration, the ISSA is used over the training data for discovering the features subset. A binary values are utilized to presented the selected and unselected features in feature selection issue, where 1 defines the selected features and 0 denotes the unselected features.

The steps of ISSA based feature selection are given as follows:

1. At first, the ISSA randomly creates the salps according to size of population. Additionally, an each generated solution has a feature subset that are randomly chosen from a overall feature set.
2. For each solution in step 1, the opposite solution is found by using the OBL. A m appropriate solutions selected by OBL creates the initial population set in step 1 and its opposite solutions discovered in step 2. The fitness value of ISSA is computed according to the accuracy error. Next, the F value assigned to the best solution from the solutions chosen by OBL that denotes the solution with less error.
3. Equation (1) and (3) are used to updated the location of each salp. Equation (1) used for location updated, when the current salp is leader; otherwise equation (3) is utilized for updating the location of follower.
4. The fitness of all salps is discovered an then the F value is updated, when there is a best solution.
5. An optimum solution is discovered by applying the LSA over the F. Next, the F is updated by LSA, when there is a best solution.
6. The steps 3, 4 and 5 are repeated until the ISSA reaches the maximum iteration count.
7. An optimum solution is provided by ISSA and it denotes the best feature subset selected by ISSA.

4.3. Classification using MLSTM

The selected features from ISSA is given as input to MLSTM for classification. The conventional LSTM is generally a development from RNN which altered the structure of memory cell by transforming $tanh$ layer. The LSTM has 3 inputs such as output in previous time step, input of a present network and unit state of previous time step. Subsequently, the LSTM has two outputs such as cell state and output. The parameters of MLSTM is shown in Table 1.

Table 1. Parameters of MLSTM

Parameter	Value
Batch size	150
Hidden layers	32
learning rate	0.0025
Lambda loss amount	0.0015
Display iteration	30000
Number of iterations	$data\ length \times 300$
Loop on dataset	300 times
Number of classes	3

The HLF is utilized to balance the MSE and MAE in LSTM. If the data has outliers and noises, the Huber function is considered as an effective loss function. According to the adaptive evaluation of lower order moments, the stochastic objective functions are optimized by using the Adam optimizer. The Adam optimizer has high computational efficacy, easy to design, effective gradient diagonal rescaling and restricted memory. The target

function is optimized by using the Adam optimizer that used to minimize the MSE. Adam optimizer operates with sparse gradients, so it's not required any stationary objective.

5. Results and discussion

The ISSA with MLSTM performance is evaluated using the Anaconda Navigator 3.5.2.0 with Python 3.7. Here, the system is configured with the Intel Core i9 processor, 128 GB RAM and Windows 10 operating system. The performance of ISSA-MLSTM is analyzed in terms of accuracy, precision, recall, F-score, NSEC, MAE and RMSE. NSEC is the ratio between the modelled time series-error difference and observed time series-error difference. Average error magnitude among the real and identified variables by avoiding the direction is referred as MAE. Next, the RMSE is utilized to discover the difference among the real and identified variables and mean. The NSEC, MAE and RMSE are expressed in equations (9) to (11).

$$NSEC = 1 - \frac{\sum_{t=1}^{T}(Q_m^t - Q_o^t)^2}{\sum_{t=1}^{T}(Q_o^t - \bar{Q}_o)^2} \quad (9)$$

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \quad (10)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \quad (11)$$

Where, the data dimension is denoted as n; an actual and predicted value are denoted as y_j and $\hat{y}_j$ respectively; time is denoted as t; observed discharge and modelled discharge are denoted as Q_o^t and Q_m respectively, and the discharge mean is denoted as Q_o.

The difference among the actual and predicted outcome is denoted using accuracy and precision denotes the amount of positive classes which belongs to positive class. Recall is ratio between the amount of positive classes discovered from all positive data in the input whereas the F-score computes the single score to balance the issues of recall and precision in one number. Equations (12) to (15) expresses the accuracy, precision, recall, F-score

$$Recall = \frac{TP}{FN+TP} \times 100 \quad (12)$$

$$Precision = \frac{TP}{FP+TP} \times 100 \quad (13)$$

$$F - score = \frac{2TP}{2TP+FP+FN} \times 100 \quad (14)$$

$$Accuracy = \frac{TP+TN}{FP+FN+TP+TN} \times 100 \quad (15)$$

Where, TN is true negative; TP is true positive; FN is false-negative and FP is false positive.

5.1. Performance analysis of ISSA-MLSTM method

The performance of the ISSA-MLSTM is analyzed with different classifiers such as MLP, RNN, LSTM-RMSPROP and LSTM-Adam for different k fold sizes. The k-fold sizes considered for evaluating the ISSA-MLSTM are 3, 5 and 10. Further, the performances are analyzed for all features and selected features from ISSA. The analysis of ISSA-MLSTM with different classifiers for all features is shown in the Tables 2 and 3. Similarly, the analysis of ISSA-MLSTM with different classifiers for selected features is shown in the Tables 4 and 5. Further, the graphical illustration of classification performances for different classifiers using all features and selected features with 10-folds are shown in Figures 2 and 3 respectively. From the tables, it is found that the MLSTM provides better performances than the MLP, RNN, LSTM-RMSPROP and LSTM-Adam. For example, the accuracy of MLSTM with selected feature for 10-fold is 99.00 % whereas MLP obtains as 74.92 %, RNN obtains as 78.29%, LSTM-RMSPROP obtains as 81.82 % and LSTM-Adam obtains as 81.74 %. The performances of MLSTM is improved by following reasons: 1) Utilization of Adam optimizer is used to optimize the target function that helps to minimize the error and 2) HLF is used to balance the MSE and MAE.

Table 2. Analysis of ISSA-MLSTM for different classifiers with all features using accuracy, precision, recall and F-score

Cross-folds	Measures	MLP	RNN	LSTM-RMSPROP	LSTM-Adam	ISSA-MLSTM
3-folds	Accuracy (%)	80.43	83.48	86.98	89.77	93.59
	Precision (%)	66.4	69.54	69.97	73.85	79.89
	Recall (%)	80.3	82.47	83.23	90.74	93.3
	F-score (%)	71.6	74.29	76.48	79.9	85.94
5-folds	Accuracy (%)	85.85	87.49	91.84	94.04	97.32
	Precision (%)	73.09	77.13	77.06	82.69	83.13
	Recall (%)	78.94	82.55	84.29	86	91.54
	F-score (%)	69.13	71.61	73.95	78.87	81.37
10-folds	Accuracy (%)	72.72	75.89	79.22	78.64	98.5
	Precision (%)	86.62	90.32	94.19	93.9	98.3
	Recall (%)	80.87	84.44	85.94	88.77	98.9
	F-score (%)	82.34	84.06	86.6	90.78	98.86

Table 3. Analysis of ISSA-MLSTM for different classifiers with all features using NSEC, MAE and RMSE

Cross-folds	Measures	MLP	RNN	LSTM-RMSPROP	LSTM-Adam	ISSA-MLSTM
3-folds	MAE	0.333	0.304	0.282	0.255	0.213
	RMSE	0.57	0.546	0.521	0.492	0.456
	NSEC	2.03	2.06	2.078	2.104	2.083
5-folds	MAE	0.189	0.162	0.143	0.117	0.066
	RMSE	0.433	0.408	0.379	0.362	0.316
	NSEC	2.026	2.05	2.078	2.102	2.089
10-folds	MAE	0.456	0.437	0.403	0.386	0.037
	RMSE	0.685	0.653	0.638	0.603	0.062
	NSEC	2.029	2.047	2.083	2.106	0.114

Table 4. Analysis of ISSA-MLSTM for different classifiers with selected features using accuracy, precision, recall and F-score

Cross-folds	Measures	MLP	RNN	LSTM-RMSPROP	LSTM-Adam	ISSA-MLSTM
3-folds	Accuracy (%)	82.93	86.48	89.88	93.07	94.29
	Precision (%)	69	72.74	73.37	76.45	81.29
	Recall (%)	83	85.67	85.63	93.34	94.20
	F-score (%)	75	77.69	79.58	82.90	86.94
5-folds	Accuracy (%)	88.35	90.19	94.24	97.14	98.12
	Precision (%)	76.19	79.93	80.06	85.49	83.73
	Recall (%)	82.14	85.75	87.29	89.20	92.14
	F-score (%)	71.43	73.91	75.95	81.67	82.57
10-folds	Accuracy (%)	74.92	78.29	81.82	81.74	99.00
	Precision (%)	89.52	93.02	96.79	96	99.00
	Recall (%)	83.87	86.94	88.64	92.07	99.60
	F-score (%)	85.74	87.16	89.40	93.58	99.96

Table 5. Analysis of ISSA-MLSTM for different classifiers with selected features using NSEC, MAE and RMSE

Cross-folds	Measures	MLP	RNN	LSTM-RMSPROP	LSTM-Adam	ISSA-MLSTM
3-folds	MAE	0.302	0.276	0.249	0.228	0.205
	RMSE	0.547	0.518	0.492	0.469	0.444
	NSEC	2	2.026	2.052	2.073	2.077
5-folds	MAE	0.161	0.136	0.113	0.084	0.056
	RMSE	0.401	0.381	0.352	0.330	0.307
	NSEC	2	2.024	2.054	2.078	2.081
10-folds	MAE	0.428	0.404	0.376	0.356	0.030
	RMSE	0.654	0.628	0.605	0.577	0.049
	NSEC	2	2.022	2.052	2.076	0.101

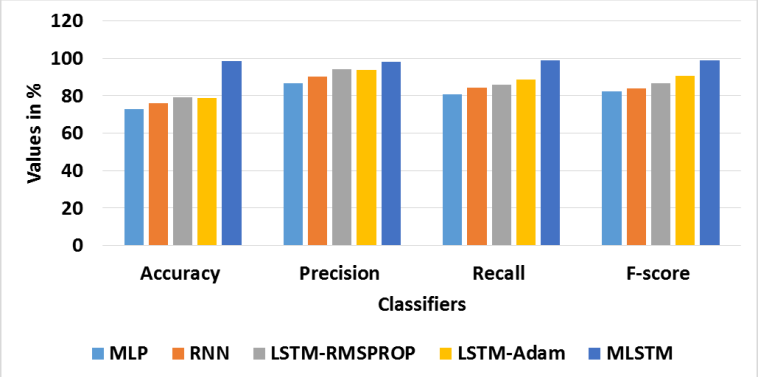

Figure 2. Graphical illustration of classification performances for different classifiers with all features and 10-folds

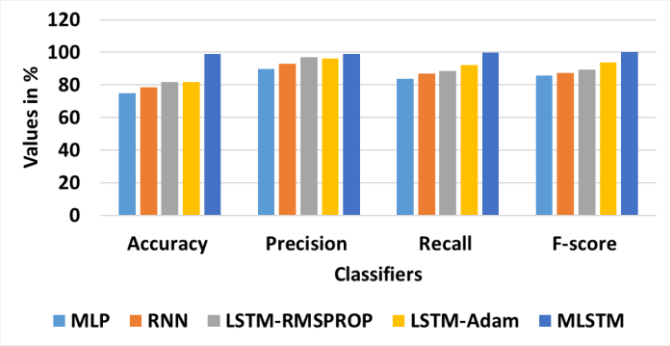

Figure 3. Graphical illustration of classification performances for different classifiers with selected features and 10-folds

The different feature selection approaches such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and SSA are considered for evaluating the ISSA. Tables 6 and 7 shows the performance analysis of ISSA with GA, PSO and SSA. Figure 4 shows the graph of classification performances for different feature selection approaches with 10 folds. This analysis shows that the ISSA achieves better performances than the GA, PSO and SSA. An enhanced population diversity using OBL and improved exploitation using LSA are used to improve the performances of ISSA.

Table 6. Analysis of ISSA-MLSTM for different feature selection approaches using accuracy, precision, recall and F-score

Cross-folds	Measures	GA	PSO	SSA	ISSA
3-folds	Accuracy (%)	82.68	86.08	89.47	94.29
	Precision (%)	69.14	69.47	73.05	81.29
	Recall (%)	81.97	81.73	90.24	94.2
	F-score (%)	73.79	76.08	79.7	86.94
5-folds	Accuracy (%)	86.69	90.54	93.74	98.12
	Precision (%)	76.33	76.36	82.29	83.73
	Recall (%)	82.15	83.69	85.5	92.14
	F-score (%)	70.81	72.45	78.17	82.57
10-folds	Accuracy (%)	74.69	78.12	77.84	99
	Precision (%)	89.82	93.19	92.7	99
	Recall (%)	83.54	84.94	88.57	99.6
	F-score (%)	83.86	85.6	89.88	99.96

Table 7. Analysis of ISSA-MLSTM for different feature selection approaches using NSEC, MAE and RMSE

Cross-folds	Measures	GA	PSO	SSA	ISSA
3-folds	MAE	0.311	0.287	0.263	0.205
	RMSE	0.557	0.529	0.505	0.444
	NSEC	2.063	2.086	2.11	2.077
5-folds	MAE	0.171	0.148	0.121	0.056
	RMSE	0.417	0.389	0.364	0.307
	NSEC	2.057	2.089	2.115	2.081
10-folds	MAE	0.444	0.413	0.39	0.03
	RMSE	0.658	0.64	0.614	0.049
	NSEC	2.058	2.088	2.109	0.101

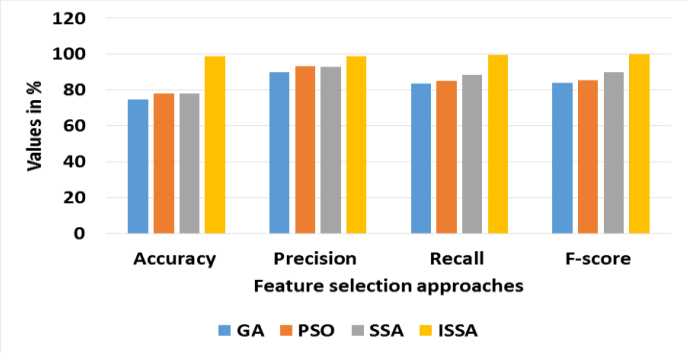

Figure 4. Graphical illustration of classification performances for different feature selection approaches with 10-folds

5.2. Comparative analysis

The existing researches such as Ensemble approach [19] and MLSTM [23] are used to compare the ISSA-MLSTM. Tables 8 and 9 shows the comparative analysis of ISSA-MLSTM with Ensemble approach [19] and MLSTM [23]. This comparison shows that the ISSA-MLSTM outperforms well than the Ensemble approach [19] and MLSTM [23]. Figure 5 shows the classification performance comparison for MLSTM [23] and ISSA-MLSTM. For example, the accuracy of the ISSA-MLSTM is 99% which is high when compared to the MLSTM [23]. An improved exploitation using LSA and enhanced population diversity using OBL of ISSA are used to choose the optimum feature subset which used to improve the crop yield estimation.

Table 8. Comparison in terms of accuracy, precision, recall and F-score

Methodology	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
MLSTM [23]	98.02	100	98.61	98.97
ISSA-MLSTM	99	99	99.6	99.96

Table 9. Comparison in terms of MAE and RMSE

Methodology	MAE	RMSE
Ensemble approach [19]	5.42	7.20
MLSTM [23]	0.030	0.049
ISSA-MLSTM	0.030	0.049

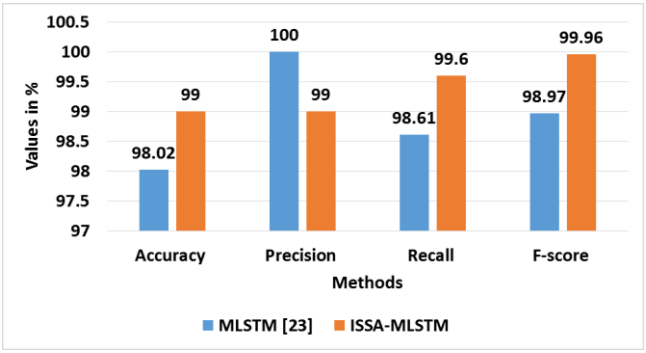

Figure 5. Classification performance comparison for MLSTM and ISSA-MLSTM

6. Conclusion

In this paper, the ISSA based feature selection is used for selecting the optimum feature subset for enhancing the crop yield prediction. Initially, the raw collected data is scaled by using the normalization approach which helps to know attributes of data. Next, the OBL used in the initialization phase and LSA used in the exploitation of ISSA are used to remove the redundant features from the overall feature set. This helps to choose the optimum feature subset using ISSA. Next, MLSTM classifier is used to perform the crop yield prediction where the Adam optimizer and HLF are used for minimizing the error rate. From the analysis, it is discovered that the ISSA-MLSTM outperforms well than the Ensemble approach and MLSTM. The accuracy of the ISSA-MLSTM is 99% that is high when compared to the MLSTM. In future, an effective hyper parameter tuning can be done for improving the performances of crop yield prediction.

References:

[1] Kamath, P., Patil, P., Shrilatha, S. and Sowmya, S., 2021. Crop yield forecasting using data mining. Global Transitions Proceedings, 2(2), pp.402-407.

[2] Banerjee, T., Sinha, S. and Choudhury, P., 2022. Long term and short term forecasting of horticultural produce based on the LSTM network model. Applied Intelligence, pp.1-31.
[3] Dharmaraja, S., Jain, V., Anjoy, P. and Chandra, H., 2020. Empirical analysis for crop yield forecasting in india. Agricultural Research, 9, pp.132-138.
[4] Shahid, M.R., Wakeel, A., Ishaque, W., Ali, S., Soomro, K.B. and Awais, M., 2021. Optimizing different adaptive strategies by using crop growth modeling under IPCC climate change scenarios for sustainable wheat production. Environment, Development and Sustainability, 23, pp.11310-11334.
[5] Satpathi, A., Setiya, P., Das, B., Nain, A.S., Jha, P.K., Singh, S. and Singh, S., 2023. Comparative Analysis of Statistical and Machine Learning Techniques for Rice Yield Forecasting for Chhattisgarh, India. Sustainability, 15(3), p.2786.
[6] Li, L., Wang, B., Feng, P., Wang, H., He, Q., Wang, Y., Li Liu, D., Li, Y., He, J., Feng, H. and Yang, G., 2021. Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across China. Agricultural and Forest Meteorology, 308, p.108558.
[7] Kakati, N., Deka, R.L., Das, P., Goswami, J., Khanikar, P.G. and Saikia, H., 2022. Forecasting yield of rapeseed and mustard using multiple linear regression and ANN techniques in the Brahmaputra valley of Assam, North East India. Theoretical and Applied Climatology, 150(3-4), pp.1201-1215.
[8] Feng, P., Wang, B., Li Liu, D., Waters, C., Xiao, D., Shi, L. and Yu, Q., 2020. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. Agricultural and Forest Meteorology, 285, p.107922.
[9] Hara, P., Piekutowska, M. and Niedbała, G., 2021. Selection of independent variables for crop yield prediction using artificial neural network models with remote sensing data. Land, 10(6), p.609.
[10] Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C. and Athanasiadis, I.N., 2021. Machine learning for large-scale crop yield forecasting. Agricultural Systems, 187, p.103016.
[11] Bojanowski, J.S., Sikora, S., Musiał, J.P., Woźniak, E., Dąbrowska-Zielińska, K., Slesiński, P., Milewski, T. and Łączyński, A., 2022. Integration of Sentinel-3 and MODIS vegetation indices with ERA-5 agro-meteorological indicators for operational crop yield forecasting. Remote Sensing, 14(5), p.1238.
[12] Paudel, D., Boogaard, H., de Wit, A., van der Velde, M., Claverie, M., Nisini, L., Janssen, S., Osinga, S. and Athanasiadis, I.N., 2022. Machine learning for regional crop yield forecasting in Europe. Field Crops Research, 276, p.108377.
[13] Nagy, A., Szabó, A., Adeniyi, O.D. and Tamás, J., 2021. Wheat yield forecasting for the Tisza River catchment using landsat 8 NDVI and SAVI time series and reported crop statistics. Agronomy, 11(4), p.652.
[14] Jayagopal, P., Muthukumaran, V., Koti, M.S., Kumar, S.S., Rajendran, S. and Mathivanan, S.K., 2022. Weather-based maize yield forecast in Saudi Arabia using statistical analysis and machine learning. Acta Geophysica, 70(6), pp.2901-2916.
[15] Pham, H.T., Awange, J., Kuhn, M., Nguyen, B.V. and Bui, L.K., 2022. Enhancing crop yield prediction utilizing machine learning on satellite-based vegetation health indices. Sensors, 22(3), p.719.
[16] Chen, Z., Jia, K., Xiao, C., Wei, D., Zhao, X., Lan, J., Wei, X., Yao, Y., Wang, B., Sun, Y. and Wang, L., 2020. Leaf area index estimation algorithm for GF-5 hyperspectral data based on different feature selection and machine learning methods. Remote Sensing, 12(13), p.2110.
[17] Elavarasan, D. and Vincent, P.D.R., 2021. A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. Journal of Ambient Intelligence and Humanized Computing, pp.1-14.
[18] Verma, A.K., Garg, P.K., Hari Prasad, K.S., Dadhwal, V.K., Dubey, S.K. and Kumar, A., 2021. Sugarcane yield forecasting model based on weather parameters. Sugar Tech, 23, pp.158-166.
[19] Nihar, A., Patel, N.R. and Danodia, A., 2022. Machine-Learning-Based Regional Yield Forecasting for Sugarcane Crop in Uttar Pradesh, India. Journal of the Indian Society of Remote Sensing, 50(8), pp.1519-1530.
[20] Gavahi, K., Abbaszadeh, P. and Moradkhani, H., 2021. DeepYield: A combined convolutional neural network with long short-term memory for crop yield forecasting. Expert Systems with Applications, 184, p.115511.
[21] Shafiee, S., Lied, L.M., Burud, I., Dieseth, J.A., Alsheikh, M. and Lillemo, M., 2021. Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery. Computers and Electronics in Agriculture, 183, p.106036.
[22] Iniyan, S. and Jebakumar, R., 2022. Mutual information feature selection (MIFS) based crop yield prediction on corn and soybean crops using multilayer stacked ensemble regression (MSER). Wireless Personal Communications, 126(3), pp.1935-1964.
[23] Dwaram, J.R. and Madapuri, R.K., 2022. Crop yield forecasting by long short‐term memory network with Adam optimizer and Huber loss function in Andhra Pradesh, India. Concurrency and Computation: Practice and Experience, 34(27), p.e7310.
[24] Mirjalili, S., Gandomi, A.H., Mirjalili, S.Z., Saremi, S., Faris, H. and Mirjalili, S.M., 2017. Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. Advances in engineering software, 114, pp.163-191.
[25] Singh, D. and Singh, B., 2020. Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97, p.105524.