Materials Today: Proceedings xxx (xxxx) xxx

Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Crop Price Prediction Using Random Forest and Decision Tree Regression:-A Review

Manik Rakhra [a,*], Priyansh Soniya [a], Dishant Tanwar [a], Piyush Singh [a], Dorothy Bordoloi [a], Prerit Agarwal [a], Sakshi Takkar [a], Kapil Jairath [b], Neha Verma [c]

[a] Department of Computer Science and Engineering, Lovely Professional University Phagwara, Punjab 14411, India
[b] Trinity College Jalandhar, India
[c] KRMDAV College Nakodar, India

ARTICLE INFO

Article history:
Received 20 February 2021
Accepted 10 March 2021
Available online xxxx

Keywords:
Machine learning
Neural network
Crop price
Agriculture
Deep learning

ABSTRACT

Machine Learning is the study in which we give the system the knowledge to think on its own using various techniques and algorithms. By thinking, it is meant that it adapts to new data without any human intervention. In today's time when almost everything has been digitalized, only the agriculture sector lacks technological advancements. India still shows a very slow pace to adapt more advanced technologies when it comes to farming. Machine Learning has emerged big data technologies and is expected to grow rapidly in the near future. With the help of machine learning we can improve the current situation in the agriculture industry and help the farmers with getting the best MSP for their crops. During the last two decades, great deals of papers have been published and a lot of different types of ANNs were investigated. Neural networks have been applied in diverse fields including aerospace, automotive, banking, defense, electronics, entertainment, financial, insurance, manufacturing, medical, oil and gas, speech, securities, telecommunications, transportation, and environment. With the use of algorithms, we can easily and accurately predict the weather conditions, soil conditions, demand in the market and correct price of the crops by comparing it with the past scenario. Keeping all the factors in mind, in this paper, we implement various machine learning techniques and create a platform to help the farmers and provide them with a more advanced way of farming. This will help them to get the right price of the crops from their buyers. This platform will act as a hub to let the farmers and the buyers explore multiple options and trade at their own convenience.

Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

1. Introduction

Farmers are the backbone of our society. They have great significance in the socio-economic structure of our country. Almost every food item of this world is produced by farmers and that is why we are hugely dependent on them. But somehow, they are the ones who are not able to feed themselves and their family due to lack of finances because farmers are not able to sell their crops in an optimum price to the buyers. One of the major reasons for this is the lack of a technological platform. Everything has gone online in recent years like education, bill payments, shopping, TV and cinema, but the only thing that lacks in technological advancement is farming and agriculture. The most recent and most severe problem in the history of farming, the country is facing the farmer's protest [1].

Keeping all the factors in mind we are creating a tech platform for the farmers with the help of Machine Learning to bring technology in Agriculture [2]. Unlike our traditions statistical methods where the focus was to work on a subset of population trying to predict the nature of whole population using that subset or sample, using Machine Learning we can also predict the nature of each and every single observation as well, rather than only getting values for the population, therefore we can now understand the behavior of individual observations, we give observations and their results to a machine learning model and the model gives us rules which can be applied to other observations whose results are unknown [3–6]. There are mainly three categories of machine learning models - supervised learning, unsupervised learning and reinforcement learning. These are some common machine learning algorithms:

* Corresponding author.
E-mail address: Rakhramanik786@gmail.com (M. Rakhra).

https://doi.org/10.1016/j.matpr.2021.03.261
2214-7853/
Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

- Linear Regression: It is an algorithm that is used for estimating the real values (cost of houses, number of calls, complete deals and so forth) in view of continuous variable(s). Here, we try to find a best fit line which can get us the relationship between independent and dependent variables. This best fit line is known as regression line and can be represented by the equation:
  - o $Y = m \times x + c$

where m is the slope of line and c is the intercept.

- Logistic Regression: It is a probabilistic model that predicts the probability of occurrence of an event by fitting data to a logistic function:
  - o $f(x) = \frac{1}{1+e^{-x}}$

Which gives values in a range of 0 to 1 being a probabilistic model and is a classification model where values above a certain threshold are on category 1 and below that threshold are of value 0. This is the building block of neural networks.

- Decision Tree: It is a tree based model and is a supervised learning algorithm which can be used for both classification and regression models here the nodes are decision points having conditions the results of which then extends the tree into more nodes.
- Support Vector Machine (SVM): In this algorithm, we plot each data item as a point in n-dimensional space (n: number of features) using the value as coordinates and then find a hyper plane which will divide dimensional space into two halve each repressing one class [7–10]. This can be used for both classification and regression problems but is mostly used for classification problems.
- Naive Bayes: It is a classification algorithm based on Bayes' Theorem and is a collection of many Bayes' Theorem based algorithms. It assumes that no two features being classified are dependent of each other or we can say are the every two features being classified are assumed to be independent of each other
- k- Nearest Neighbors (kNN): It is a simple distance-based algorithm which can be used for both classification and regression. Based on the nearest k number of observations to our target observation we can take the mode of those k-neighbors to get classifies the target observation and take mean of k-neighbors to get the regression value.
- K-Means: It is a type of unsupervised algorithm which is distances based and solves the clustering problem.
- Random Forest: Random Forest is a kind of democratic collection of many decision trees, where to tackle the problem of overfitting of a single Decision tree we now do voting and the most voted class wins and is the final result for your target observation.
- Dimensionality Reduction Algorithms: Dimensionality reduction is an unsupervised learning technique which can help in reducing the number of dimensions of multi-dimensional dataset to specified number of most important dimensions which explains the most variance in dataset.
- Gradient Boosting Algorithms:
  1. GBM: Gradient Boosting Machine (GBM) algorithm is used to deal with plenty of data to make predictions with higher accuracy. Boosting is an ensemble technique which combines the prediction given by several base estimators to improve the robustness over a single estimator here these combined estimators are multiple weak or average predictors which are used to a build strong predictor.
  2. XGBoost: The XGBoost is almost 10x faster than existing gradient booster techniques as it provide parallel tree boosting which results in more accuracy and speed and results in making it the one for the most popular tree based model.
  3. LightGBM: LightGBM is a gradient boosting framework designed to be efficient in memory usage, accuracy and large data handling along with supporting parallel processing with GPU as well.
  4. CatBoost: it the first machine learning technology to be classified as open-sourced by Russia. It's easy integration with deep learning frameworks like Google's TensorFlow and Apple's Core ML makes it a popular choice among data scientists.

This site will contain a Signup page where you will be able to make an account as a farmer or a buyer. Furthermore, there will be a trading lobby where the buyers will put out their price for the crops which they desire, based on their quality and quantity. The farmers will be able to choose the best price for their crops and the quantity they'll be supplying. This will terminate the "middleman" system and save the money of the farmers. A crop price predictor function will also be provided which will give farmers the idea of the MSP of the crop by comparing it with other regions [11–15]. A chat option will also be provided to make communication private and better between the users. This platform will act as a hub to let farmers and buyers explore multiple options and trade at their own convenience.

1.1. Time series prediction using LSTM

Time Series prediction is a supervised learning method which makes use of recursive link between iterations. It works on the ideology that the output from the parameters at a given time instance can be dependent on the previously known set of inputs and/or outputs. The time series prediction has three known entities which lead to determining the values of the weight matrix. These three entities are the input training data, target variable from the current instance of time, as well as previously known information from a past instance of time.

LSTM is a recursive neural network that is widely used for time series machine learning computations. LSTM stands for Long Short Term Memory. It is very useful to deal with the problem of vanishing gradients. LSTM uses different gates to read, write, delete, and update information inside itself. These gates are Input Gate, Forget Gate, and Output Gate. The input gate updates the data inside LSTM, forget gate determines which information should be retained and which to be deleted and the output gate decides the value of the next hidden unit. All of these gates have combinations of tanh and sinh functions. LSTM works really well with the predictions involving the analysis of trends. Such as Natural Language Processing and Stock Price Prediction. Both these problems require analysis of patterns from past instances of data and predicting the future value. LSTM hence is very useful in crop price prediction as well because as similar to stock prices, the price of crops also depends on the past instances of data [16–20].

2. Literature review

Applications of Machine Learning Techniques in Agricultural Crop Production describes about various Machine Learning applications that would prove to very useful in the agriculture sector. For these applications, a large amount of data available from many resources can be analyzed to find the hidden knowledge. This research field is growing day by day and will prove to be a great tool for the development of the agriculture sector in the future. The combination of Agriculture and Computer Science will provide

a great scope of development in the agriculture sector. The paper, Smart Farming: A Techno Agriculture Advancement Powered by Machine Learning's main aim is to make the agriculture sector aware about the modern technologies. Accurate predictions should be made with the help of machine learning instead of manual predictions so as to improve the commercial value of the crops. The main problem identified in the paper is that India is the only country which lacks in technological advancements in the agriculture sector, due to which manual predictions are done for everything. These results in crops being sold at a less price and sometimes even the crops are ruined due to wrong predictions about the weather. The solution to this is Machine Learning. ML is the technique to provide knowledge to the machine so that it can think on its own provided with the correct data. This will help to raise the standards of agriculture in India. Machine Learning in Agriculture. This paper, Machine Learning in Agriculture: A Review is based takes a practical approach and implements many learning models and algorithms in the field of Agriculture. It is evident from this paper that most of the studies use ANNs (Artificial Neural Networks) and SVMs (Support Vector Machines) models. Specifically, ANNs were for used for implementing the crops, soil and water management whereas, SVMs were used for livestock management. One of the interesting things about this paper is that it has explored the Weed Detection and management, which is another problem in the agriculture industry that should be emphasized on. The paper, Crop Price Prediction System using Machine Learning Algorithms' main objective is to estimate the crop price by analyzing the existing data using certain data analytics techniques. This paper shows a more of a practical approach towards the topic. Data from various sources have been collected and a system is created for the crop price prediction. The whole system has been implemented using the python programming language. The data is collected from reliable sources and stored in a storage where it is then used accessed, transfer and analyzed by an organization.

The data is then processed and transform the raw data into a more efficient format. Machine learning techniques like linear regression and neural networks are used here to determine important information and to increase the accuracy percentage of the price prediction. The final results are shown through visual elements like graphs and flow charts. This paper is very useful to understand the importance of machine learning techniques and how modern technologies can prove to be very useful in the development of agriculture sector. This paper, Crop Price Prediction using Random Forest and Decision Tree Regression's main objective is to predict the price of the crop and estimate the profit for the crops given in the system before sowing. The databases provide enough data for predicting the appropriate MSP for the crops and their demand in the market. Random Forest is an easy-to-use algorithm that provides great results even without hyper-parameter tuning in comparison to the Decision Tree regression. This paper shows that the Random Forest is a very effective technique in the prediction. As there are many algorithms in machine learning in agriculture, there are various ways to make predictions for the crops which is also one of the main things this paper has shown, that there are different algorithms for different crops and not a single one for all the crops. Thus, this paper shows how machine learning can prove to be beneficial for the farmers to take the right decisions in choosing the crops by analyzed results.

3. Machine learning

Machine learning is the branch of computer science that allows computers to learn finding the solutions to problems on their own. In other words, machine learning makes the computer able to find solutions without being explicitly programmed. A general computer program aims to take the input, process it over the provided instructions, and give the output. Machine Learning, on the other hand, focuses on the input provided and the solution to the problem in order to find the best suitable algorithm that led to the solution. The machine learning algorithms have three categories - supervised learning, unsupervised learning, and reinforcement learning. An Unsupervised Learning algorithm concerns the input data as well as the output. They learn by predicting the output value and minimizing the error between predicted and actual value over multiple iterations. Unsupervised learning algorithms are unknown to the output of the data provided. They make their own rules to give logical relations and patterns in the data. They are more machine-dependent than human-dependent [21–25]. The Reinforcement Learning algorithm works on giving rewards and punishment to the logic. It finds that logic that gives maximum reward or least punishment to find the solution. It learns by experience and finds the path of least resistance.

3.1. Machine learning in agriculture

Machine learning in the agricultural sector is a comparatively new concept. The implementation of machine learning in agriculture are being seen for a few decades but still, it has made a significant contribution to the field. The most helpful contribution in this field is the weather and rainfall prediction. It makes the farmers ready for the upcoming weather anomalies and protects their plants. The rainfall prediction helps the farmers to do the right amount of irrigation on their crops. It also makes them easier to prepare for droughts and atrocious conditions. The other applications of machine learning in farming include the study of soil and the prediction of the quantity of fertilizers and additional materials that are optimum for farming. Machine learning algorithms can also help in determining which type of crop would yield the best results and make the best use of the land by studying the geographical location, weather, atmospheric conditions, and soil type. The latest application is studying the air and its constituents and toxications that would affect crop health and yield. Today, the availability of technical devices and computation power has made it possible to apply high computational algorithms like deep learning very easily and effectively. Deep learning is a specialized domain of machine learning that aims to mimic human-like reasoning and decision making with the aid of deep network structures. These networks are built homologous to neurons in human brains. They form the neural network and learn progressively just like humans do. They are very powerful algorithms and can accomplish very complex and high computational tasks. Fig. 1 depicts the concept of precious agriculture in India. The most revolutionary concept of deep learning in computer vision. It allows the computers to study, analyse and differentiate images and videos. This capability is used in agriculture to detect anomalies and diseases in plants. Several countries are implementing computer vision-based analytical systems to monitor the health and development of the crops. We aim to make an algorithm that can predict the

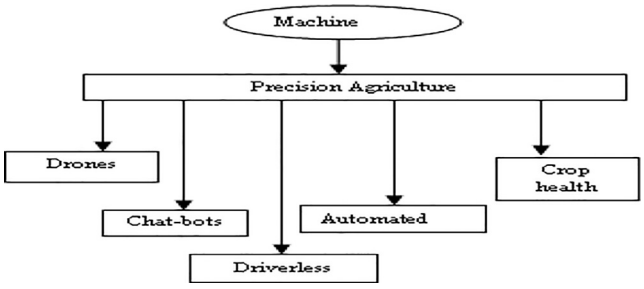

Fig. 1. Precious Agriculture in India.

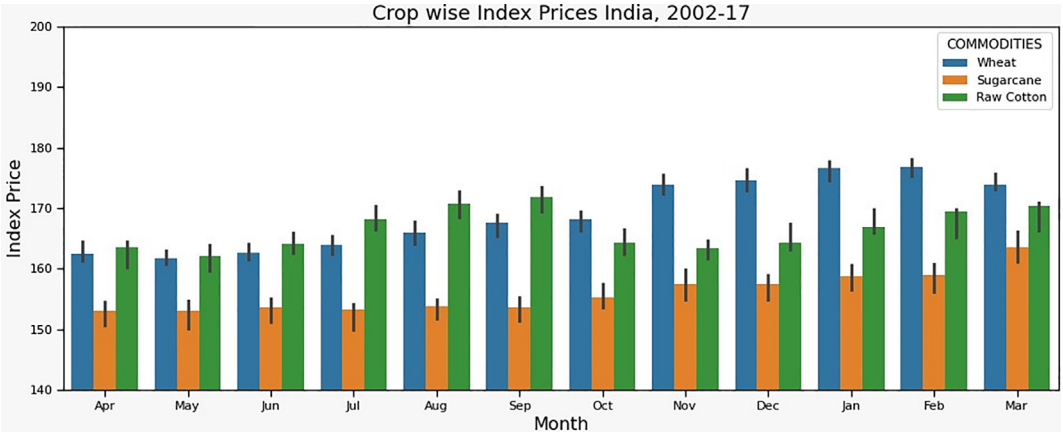

Fig. 2. Crop Wise Price index in India.

possible prices of the crops that farmers can get. There are studies going on in this field but no large-scale implementation is observed. This model can provide a solution to a very critical problem in India. MSP and crop prices have always been in a questionable status and an accurate and unbiased algorithm that predicts the closest possible price of crops on the basis of the type of grains, geographical location, and market analysis will be very beneficial for an agricultural country like India.

3.2. Machine learning algorithms for crop price prediction

There are a few algorithms that can predict the future price of crops. Both supervised and unsupervised learning algorithms have algorithms that can perform prediction tasks. The artificial neural network is a recently advanced field that performs human-like logic to predict the values from provided input.

3.3. Linear regression

Linear regression is a supervised learning algorithm. It aims to fit the problem on a linear hypothesis. When two or more input parameters are involved it is called multivariate linear regression. The algorithm assigns coefficients to the variable then predicts the output value. The difference between the actual and predicted value is called the cost. The cost is estimated by finding the distance between the actual data point and the regression line.

Cost Function- Cost Function calculates the average cost of all training examples over a single iteration. Then the new values of coefficients are calculated using the cost function.

Learning Rate- Learning rate determines the speed at which the algorithm descends towards minima that means the rate at which the cost function is minimized. If the learning rate is much higher, the algorithm might never reach the optimization value and fluctuate towards high errors. It leads to under fitting. If the learning rate is very small, the time number of iterations needed to each minima will increase, and also chances of over fitting. Assigning the right value to the learning rate is very important.

$$y = ax + b$$

$$\text{cost} = |y - \hat{y}|$$

$$\text{costfunction}\,(J) = \frac{1}{2m}\sum(\text{cost}(i))$$

Linear regression is very helpful in value prediction and forecasting. It is being used in many prediction systems like stock price prediction, house price prediction, crop production, and price estimation, etc.

3.3.1. Decision tree

The decision tree algorithm is based on recursive partitioning of input parameters to classify the data. It is an advanced algorithm that determines the importance and contribution value of input parameters. It learns on the dataset by calculating the Entropy and Information Gain of a decision. Different decisions over multiple iterations are made and the best fit is determined by that set of decisions that have the least entropy. It is also a part of the family of supervised learning algorithms where the dataset inputs have pre-hand knowledge of output. The decision tree algorithm has a recursive tree or graph network-like structure of decision calls. Each decision call contributes to a change in entropy. The successive decision calls leads to an output set of decision paths. On every decision call, the set of input parameters is divided into two or more subsets. It is called Splitting. There are different methods used to make the splitting decision like MARS, Chi-Squared, ID3, Cart, etc. Fig. 2 represents the crop wise price index in India. The algorithm begins with the root node which contains all training examples in the dataset. Then at each iteration of the algorithm calculates the Entropy and Information Gain. Using these values the attribute with the least entropy or highest information gain is selected as the splitting node and the rest attributes are split into subsets. The recursive splitting continues over each subset until the terminal nodes are reached. Entropy- Entropy of a decision is defined as the degree of randomness of the data. Higher value entropy means higher randomness in the data, which makes it harder to make logical analysis and conclusive study of data.

4. Conclusions

The main aim of this paper is to introduce modern technologies into the agriculture sector. To improve the commercial value of the crops of the state by accurate predictions. To provide a reasonable solution to our happening farmer crisis and any possible related crisis in near future. We know that ML models have been in use in multiple applications for crop management (61%), yield prediction (20%), and disease detection (22%). By applying machine learning to sensor data, farm management systems are evolving into real AI systems, providing richer recommendations and insights for the next decisions and actions with the ultimate scope of market improvement. For this scope, within the future, it's expected that the usage of ML models are going to be even more widespread, allowing the likelihood of integrated and applicable tools. This integration of automated data recording, data analysis, ML implementation, and decision-making will provide practical tools that come in line with the so-called knowledge-based agriculture for increasing trading levels, profit margins (not only to farmers but

also buyers).In this paper, certain Data Analytics techniques were adopted to estimate crop price analysis with existing data. Linear Regression is used for locating important information from the agricultural datasets. Neural Network is made for price prediction to extend the accuracy percentage. The root mean square error is calculated for each technique to accurately measure the accuracy of every system employed and therefore the most accurate system is then selected. Entropy is used to calculate the randomness and predict the logical analysis and conclusiveness of data. It is reliable to use yield/productions, and export/import profiles to predict the nearby close market price, which is of practical value for financial purposes. XGBoost anticipates the target better when compared to other algorithms. This could be a step to change from manual predictions to automated predictions and calculation by using machine learning, to boost the standards of the agriculture sector in India, to scale back the farmer suicides by making their crops obtainable and with added commercial value to them, or an ambition to form India the hub of techno-agricultural advancements.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Ahmed, 'Farm Mechanization in Bangladesh: Evidence from IFPRI National Household Survey', Rural Mechanization: Policy and Technology Lessons from Bangladesh and Other Asian Countries, pp. 7–8 (2013).
[2] P. Soni, Y. Ou. 'Agricultural Mechanization at a Glance Selected Country Studies in Asia on Agricultural Machinery Development', United Nations Asian and Pacific Centre for Agricultural Engineering and Machinery (2010).
[3] A. Bagheri, A. Ghorbani, Adoption and non-adoption of sprinkler irrigation technology in Ardabil Province of Iran, Afr. J. Agric. Res. 6 (5) (2011) 1085–1089.
[4] S. Biggs, S. Justice, D. Lewis, Patterns of rural mechanization, energy and employment in South Asia: reopening the debate, Econ. Polit. Weekly 46 (9) (2011) 78–82.
[5] S. Biggs, S. Justice 'rural and agricultural mechanization: a history of the spread of small engines in selected Asian countries', IFPRI -Discussion Papers. Available at: http://csisa.org/wp%0Acontenthwloads/sites/2/201412006/BiggsJustic (2015).
[6] Y. Bigot, H.P. Binswanger, Agricultural Mechanization and the Evolution of Fanning Systems in Sub-Saharan Africa, Johns Hopkins University Press, 1987.
[7] L.J. Clarke, Strategies for agricultural mechanization development: the roles of the private sector and the government, Agric. Eng. Int.: CIGR J. (2000) 2.
[8] L.J. Clarke. 'No Title', Agricultural Engineering Branch, Agricultural Support Systems Division 1997 FAO, Rome, Italy, (September).
[9] CSAM (2014) '2nd Regional forum on sustainable agricultural mechanization in Asia and the Pacific. Serpong, Indonesia', Centre for Sustainable Agricultural Mechanization (CSAM) United Nation Economic and Social commission for Asia and the Pacific (UNESCAP).
[10] X. Diao, F. Cossar, N. Houssou, S. Kolavalli, k. Jimah, P. Aboagye. 'Mechanization in Ghana searching for sustainable service supply models', FPRJ- Discussion Papers. Available at: http://ebrarv.ifuri.org/cdmfref/collectionlplS738colll57 (2012).
[11] X. Diao, F. Cossar, N. Houssou, S. Kolavalli, Mechanization in Ghana: emerging demand, and the search for alternative supply models, Food Policy 48 (2014) 168–181.
[12] M. Rakhra, D. Venkatesh, Agile adoption issues in large scale organizations: a review, Mater. Today:. Proc. (2020), https://doi.org/10.1016/j.matpr.2020.11.308.
[13] M. Rakhra, R. Singh, Smart data in innovative farming, Mater. Today:. Proc. (2020), https://doi.org/10.1016/j.matpr.2021.01.237.
[14] M. Rakhra, R. Singh. Internet Based Resource Sharing Platform development For Agriculture Machinery and Tools in Punjab, India" 978-1-7281-7016-9/20/$31.00 ©2020 IEEE (2020).
[15] M. Rakhra, R. Singh, A study of machinery and equipment used by farmers to develop an uberized model for renting and sharing, Mater. Today:. Proc. (2020), https://doi.org/10.1016/j.matpr.2020.11.784.
[16] M. Garg, G. Dhiman, Deep convolution neural network approach for defect inspection of textured surfaces, J. Inst. Electron. Comput. 2 (1) (2020) 28–38.
[17] M. Dehghani, Z. Montazeri, G. Dhiman, O.P. Malik, R. Morales-Menendez, R.A. Ramirez-Mendoza, L. Parra-Arroyo, A spring search algorithm applied to engineering optimization problems, Appl. Sci. 10 (18) (2020) 6173.
[18] K. Moorthi, G. Dhiman, P. Arulprakash, C. Suresh, K. Srihari, A survey on impact of data analytics techniques in E-commerce, Mater. Today:. Proc. (2021).
[19] A. Kaur, S. Kaur, G. Dhiman, A quantum method for dynamic nonlinear programming technique using Schrödinger equation and Monte Carlo approach, Mod. Phys. Lett. B 32 (30) (2018) 1850374, https://doi.org/10.1142/S0217984918503748.
[20] R.K. Chandrawat, R. Kumar, B.P. Garg, G. Dhiman, S. Kumar, An analysis of modeling and optimization production cost through fuzzy linear programming problem with symmetric and right angle triangular fuzzy number, in: Proceedings of Sixth International Conference on Soft Computing for Problem Solving, Springer, Singapore, 2017, pp. 197–211.
[21] G. Dhiman, MOSHEPO: a hybrid multi-objective approach to solve economic load dispatch and micro grid problems, Appl. Intelligence 50 (1) (2020) 119–137.
[22] A. Kaur, G. Dhiman, A review on search-based tools and techniques to identify bad code smells in object-oriented systems, in: Harmony Search and Nature Inspired Optimization Algorithms, Springer, Singapore, 2019, pp. 909–921.
[23] G. Dhiman, A. Kaur, Spotted hyena optimizer for solving engineering design problems, in: 2017 International Conference on Machine Learning and Data Science (MLDS), IEEE, 2017, pp. 114–119.
[24] P. Singh, G. Dhiman, A hybrid fuzzy time series forecasting model based on granular computing and bio-inspired optimization approaches, J. Comput. Sci. 27 (2018) 370–385.
[25] M. Garg, G. Dhiman, A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants, Neural Comput. Appl. (2020) 1–18.