AN EFFICIENT FEATURE SELECTION AND CLASSIFICATION FOR THE CROP FIELD IDENTIFICATION: A HYBRIDIZED WRAPPER BASED APPROACH

Mrs. K.A. Poornima, Assistant Professor in Computer Science, Gobi Arts & Science College (Autonomous), Gobichettipalayam, Tamilnadu, India.
Email: poornima191982@gmail.com

Dr. G. Dheepa, Assistant Professor of Computer Science, P.K.R. Arts College for Women(Autonomous), Gobichettipalayam, Tamilnadu, India. Email: dheepag@pkrarts.org

Abstract: Agricultural stakeholders are concerned about the anticipated crop production before the harvest. Many countries throughout the world employ computational technique's for predicting yield ahead of harvest to assess a country's food security and issue warnings about impending food shortages. This is a common method that aids strategy planners and decision-makers, particularly in rural economies. Crop statistical models have been used to track crop development and forecast production. The only inputs available at the field level will yield a prediction for a narrow region; remote sensing observations cover a broad area. They may be repeated at regular intervals, allowing for large-scale crop modelling. Crop yield prediction study necessitates a variety of production parameters and algorithms. Some algorithms are used to determine the optimum feature subset for improved prediction, while others are used to determine prediction. The proposed Correlation based Sequential Forward Feature Selection (CSFFS) is compared with the existing feature selection approaches. The classification with proposed feature selection attains effective accuracy in crop prediction.

Keywords: Feature selection, classification, accuracy, crop yield prediction, and wrapper approach.

1. Introduction

The progression in the agricultural industry has been undergoing an expansion from physical to digital at an unprecedented pace. As a result, it has become possible for people to collect more and more data conveniently in many ways. Data collection process is not significance without processing [1]. The collected data could not impact anything on the

agricultural industry or process without converting it into a machine-readable format. Data analysis leads us to a unique problem where data processing has to keep up with the pace of data accumulation [2].

One way to conquer this situation is to develop advanced data selection methods in pace with the rate of data. Finding unknown and interesting patterns from the datasets is known as data mining [3]. Recent years have witnessed extensive research in feature selection algorithms. Research starts from the conventional feature selection supervised to semi-supervised in between unsupervised feature selection for various types of features, namely structured and unstructured features [4].

In the dataset, some of the information on the features or variables may be redundant and it does not directly affect the prediction accuracy. The inclusion of all the attributes under such conditions does not lead to the gain of any desired outcomes. One method to reduce the cost and data processing is to devise methods that can identify more relevant data from the existing data. In data mining, this can be achieved by using feature selection tools. The feature selection models are convenient for detecting and eliminating the irrelevant features from the dataset [5].

Feature selection in predictive analytics can be defined as the process of identifying the essential variables. The usefulness of such methods is that they can progress the forecasting accuracy.As a result, the analyst could explain the outcome of the model easily. It also attains the cost reduction aspect by minimizing the count of features, which is necessary to gather. The process of removing independent features that are correlated with each other and ensuring the independent features highly correlated with dependent one is known as feature selection [6, 7].

There are several benefits in the analytical process of employing feature selection inclusion. The foremost one is less memory space requirement and less computational time than the full feature set since it focuses on feature subset instead of building the model with the entire feature set. The features selected from the set of features will improve yield prediction. In this context, the advantage of the feature selection algorithm turns out to be useful [7].

1.1.Motivation

The motivation behind the feature selection is the curse of dimensionality in dataset. It is difficult to predict unseen data by a set of rules that contains a limited number of training samples. A set of rules were set to predict classes based on the given training samplesIt was observed that the more features that were studied, required a broader rule set to be framed to reduce the noise. The process becomes worse when the number of features experiences an exponential increase in the hypothesis space from a linear pattern. Feature selection effectively reduces the hypothesis space by deleting redundant and irrelevant variables. Conversely, it has also been observed that when the hypothesis space is smaller, the correct hypothesis can then be easily found. The number of required training instances also gets significantly reduced when the dimensionality is lowered as a result of using a smaller sample from the entire population.

1.2. Contribution

The main contribution of the research are

- Redundant features can be considered as an irrelevant feature and were not considered for further processing. It was noted that the removal of irrelevant features did not affect learning performance. If a data set with n features were considered, the optimum application of the feature selection will give m relevant features and in all the cases.

- Informative features and it is further useful to feature set reduction, data reduction, data understanding, and performance improvement. Evaluation measures, models, and the search strategy were the crucial factors that are to be considered for feature selection.

The remainder of the article is organized as follows: recent works in feature selection is discussed in Section 2, the proposed feature section scheme is discussed in Section 3, acquired results are discussed in Section 4 and the article is concluded in Section 5.

2. Related Works

The feature selection approach is used to remove some irrelevant features in many applications, where the dataset has some irrelevant features and mainly focuses on finding the relevant features. Theoretically, more features provide better results, but in reality, more

features will fall off the learning process and redundant features may confuse the learning algorithm [8]. In statistical pattern recognition research, feature selection is an active and successful research field. Also, its robustness keeps more research on machine learning [9].

Feature selection theoretically and experimentally proves itself by improving learning efficiency, reducing result complexity, and increasing predictive accuracy. Feature selection approaches are categorized into three major typical models called filter, wrapper, and embedded approaches [10, 11]. The types of feature selection algorithms applied in this research work are classified. The wrapper approach is used to select feature subsets and the algorithms following this approach are computationally expensive if there are more features. Feature subset selection is independent of any learning algorithm in the filter approach [12].

Supervised and unsupervised feature selection approaches are used to select features from the dataset. The embedded or hybrid approach takes advantage of both supervised and unsupervised approaches. In an embedded approach, the feature selection is combined with the model-building to improve the learning performance [13]. Here, the feature selection has to be done in each branching node. Feature selection has attained success in many applications such as text categorization, image retrieval, genomic microarray analysis, and intrusion detection. There are various types and approaches of feature selection in crop yield prediction [14, 15].

The prominent feature selection algorithms namely forward feature selection algorithm (FFSS) [16], correlation-based feature selection algorithm (CBFS) [17], random forest variable Importance (RFVI) [18], and Variance inflation factor (VIF) [19] are reviewed. In these algorithms some of the significant features are missed due to lack of selection. By considering the drawback, a hybrid approach is framed that is Correlation based Sequential Forward Feature Selection (CSFFS).

3. Proposed Methodology: Hybridized Wrapper based Approach

3.1 Data Collection

Features or parameters or attributes that have been used in this research were the factors of the production of agricultural products. The agricultural production depends on these factors. The changes in these factors will have a meaningful impact on the selected areas yearly agricultural outcome. The attributes or parameters are mainly depended on the availability of the data. Two different sets of statistical data that were used for the study

where the statistical and agricultural data for paddy production and its weather data for the respective years. The collected two data sets were combined into a single data set. Table 1 describes the dataset.

Feature_ID	**Data Type**	**Feature Type**	**Category Type**	**Explanation**
CL	Integer	Predictor	Continuous	Length of canal that is utilized for irrigation in meters
TW	Integer	Predictor	Continuous	Whole count of tube wells utilized for irrigation
TK	Integer	Predictor	Continuous	Whole count of canal utilized for irrigation
OW	Integer	Predictor	Continuous	Whole count of open wells utilized for irrigation
AH	Integer	Predictor	Continuous	Land area utilized for the purpose of cultivation
KF	Numeric	Predictor	Continuous	Quantity of potash utilized for cultivation
NF	Numeric	Predictor	Continuous	Quantity of nitrogen utilized for cultivation
PF	Numeric	Predictor	Continuous	Quantity of phosphate utilized for cultivation
SD	Numeric	Predictor	Continuous	Quantity of seeds utilized for cultivation in kilogram
Rain	Numeric	Predictor	Continuous	Average amount of rainfall for the year in mm
AT	Numeric	Predictor	Continuous	Average temperature in mean for a year

Tmax	Numeric	Predictor	Continuous	Average of maximum temperature registered in a year
Tmin	Numeric	Predictor	Continuous	Average of minimum temperature registered in a year
SR	Numeric	Predictor	Continuous	Average of gathered radiation in a year
PD	Integer	Response/Target	Continuous	Whole production of the year in ton

3.2. Pre-Processing

Pre-processing converts all data into one type of unit system, finding missing values or entries, formatting entries, eliminating unnecessary values along with and cleaning were included since the data were not in the standard format. Unnecessary entries in the collected data were also filtered out during this process and the data needed to be converted into the same units of measurement from multiple formats. After getting through all the screening and cleaning processes, the data needed to be formatted and the CSV file was used in the model.

3.3. Feature Selection: Correlation based Sequential Forward Feature Selection (CSFFS)

Some heuristic search procedure is applied in the CSFFS algorithm to find an appropriate subset. The feature is useful if it is relevant to the class and is not redundant to any other relevant features. Two variable correlations were measured by using this algorithm. The filter approach is a multivariate one. Subsets of the feature are ranked based on their heuristic function. The evaluation function is designed toward subsets with attributes that are substantially associated with the class but unrelated to one another. Since they have a low association with the class, irrelevant characteristics are deleted. Traits that are significantly associated with one or more of the rest of the features are filtered out as well. A feature's worth will be determined by how well it forecasts classes in portions of the instance space where other features haven't yet predicted them. The score is calculated by using equation 1.

$$Feature_{Score} = \frac{FC\overline{avg}_{cf}}{\sqrt{FC + FC(FC-1)\overline{avg}_{ff}}}$$

where FC is the subset's feature count, $\overline{avg}_{cf}$ is the average correlation among every feature in S and the output variable C, and $\overline{avg}_{ff}$ is the average feature-to-feature pair-wise correlation among the features in S.The pair-wise feature correlation matrix is computed using m((n 2 n)/2) operations, where m is the count of occurrences and n is the starting count of features. The CSFFS approach has an O(2n) time complexity, where n is the count of features.

The collected agricultural data relevant to the current research were cleaned and subjected to the CSFFS algorithm. The CSFFS algorithm generates the correlation matrix and contains a correlation value between all the independent and dependent variables. After the matrix generation, it follows some evaluation function to calculate the score of each feature subset. It arranges the score in ascending order. The feature subset with the highest score was selected as the best subset and the features in the subset were considered necessary for better paddy crop yield prediction. In this research, by applying CSFFS, the feature subset {AH, OW, TK, Tmax, NF, PF, KF, SD} gave the highest score for the dependent feature PD. So, this feature subset was considered as the best subset and it was applied for further evaluation. features are reached. The SFFS algorithm is based on the Akaike Information Criterion (AIC) value for feature selection. The time complexity of the algorithm is O(n) for SFFS. The procedure for the CSFFS is given below.

Algorithm 1. Correlation based Sequential Forward Feature Selection (CSFFS)

F:={F_1, F_2, F_3,…., F_n}

Feature space correlation matrix generation

Estimation of feature score $Feature_{Score} = \frac{FC\overline{avg}_{cf}}{\sqrt{FC+FC(FC-1)\overline{avg}_{ff}}}$

Arranging features based on feature score

Estimation of AIC value for every feature

If F_i:=minimum AIC value then F_{subset}={ F_i }Repeat

{

i:=1

Estimation of AIC value for every feature subset

I=i+1

}

Feature subset return

In this procedure, the AIC value for feature selection is used to select features. Initially, it starts with the null set. The subsequent iterations keep on adding the new features until there is no further improvement in the prediction accuracy. Variable selection procedures generate variable subsets iteratively and evaluate the generated subset until a stopping criterion is met. The selection process is done by subset generation, subset evaluation, and stopping criteria. The feature subset selected by SFFS met the lowest AIC value. As the AIC value increases, the selection procedure was terminated, while adding one more feature on this set.

3.4. Classification using ANN

The use of artificial neural networks for data-driven learning was motivated by the discovery that human brains are very efficient at processing large quantities of input data from different sources. Neurons in the brain receive signals from other neurons, process them, and combine them into an output that is again passed on to other neurons. Like the brain, artificial neural networks (ANNs) are interconnected architectures of simple processing elements called nodes or neurons. ANNs, in their simplest form, can be regarded as data transformers, with the objective to relate elements in one set (input features) with elements in another set (output feature) [20, 21].

Neural networks involve backpropagation that encourages networks to change their invisible neuron layers in scenarios where the result does not conform the expected outcome. A multi-layer network's input layer select distinct feature before it is able to understand the process. By analyzing a significant number of input and output cases, ANN models discover associations to build a rule Complex relations can be modeled using Artificial Neural Networks (ANNs) since their potential to deal with numerous inputs and correlations is well established.

The model's accuracy was increased through feeding and testing different blends of inputs to the ANN models. Researchers used ANN models to predict weather systems by developing ensemble models. The ANN, widely used in yield predictions by the development of empirically-based agricultural models. Figure. 1 depicts the schematic of ANN with input, hidden and output layers.

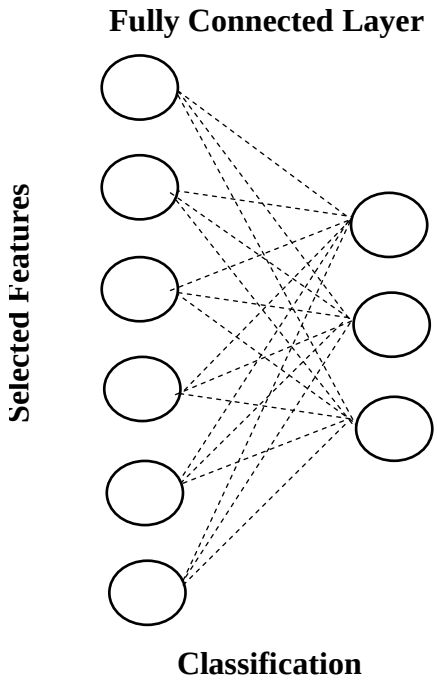

Figure 1. Schematic Representation of ANN

4. Result and Discussion

The implementation of the proposed scheme is done using Java with 4GB RAM and the performance investigation is given as follows,

4.1 Feature Selection

This research work analyses the agricultural data set with four existing feature selection algorithms, namely forward feature selection algorithm (FFSS) [16], correlation-based feature selection algorithm (CBFS) [17], random forest variable Importance (RFVI) [18], Variance inflation factor (VIF) [19] and the proposed CSFFS. By applying the different feature selection algorithms, different feature subsets are selected. The resultant feature subsets are shown in Table 2.

Table 2. Feature Selection by diverse Feature Selection Approaches

Feature_ID	Feature Selection Algorithm				
	SFFS	CBFS	RFVI	VIF	CSFFS
AH	Yes	Yes	Yes	Yes	Yes
CL	Yes	Yes	-	Yes	Yes
TK	Yes	Yes	Yes	Yes	Yes

TW	-	Yes	-	Yes	-
OW	Yes	Yes	Yes	Yes	Yes
SD	-	-	-	Yes	-
Rain	-	-	-	Yes	-
AT	-	Yes	-	Yes	-
Tmin	Yes	-	-	Yes	Yes
Tmax	-	Yes	-	Yes	Yes
SR	-	Yes	Yes	Yes	-
NF	Yes	-	Yes	Yes	Yes
PF	Yes	-	Yes	-	-
KF	Yes	-	Yes	-	-
PD	-	-	-	-	-

Every feature selection algorithm selected the best feature subset based on its unique selection procedure. The SFFS follows the AIC and feature score of each feature subset was considered for selecting features. In this procedure, once the feature selected as the best feature, the same features, along with other combinations, were calculated by following the iteration. Therefore, even if other combinations gave a better subset, it was not considered the best subset.

In CBFS, the features which had the highest correlations with dependent features produced the highest score and selected as the best feature subset. When the interaction among the features is strong, CBFS fails to select some of the relevant features. In VIF, co-linearity between independent features checked and collinear features filtered out from the entire feature set. The remaining features considered as the best features. However, in RFVI, each independent feature's importance concerning the dependent feature was calculated using the Gini Index. The redundant features removed from the entire data set to improve prediction accuracy.

In this research work, when AIC was calculated for an empty set, it was found to be –1566.6. In the subsequent step, each feature's AIC value was calculated individually and compared with each other to find the lowest AIC value feature. The lowest AIC value for AH was found to be –2464.22. In the next subsequent step, along with AH, the AIC value of all the other two variable feature subsets was calculated and the feature subset {AH, OW}

showed the lowest AIC value which was found to be –2517.4. The selection procedure is listed in Table 3.

Table 3. Feature Selection using AIC

Subset of Feature	**Rate of AIC**	**Procedure of Feature Selection**
{}	1566.6	
{AH}	–2464.22	↓
{AH,OW}	–2517.4	↓
{AH,OW,TK}	-2529.69	↓
{AH,OW,TK,CL}	-2533.63	↓
{AH,OW,TK,CL,Tmin}	-2534.73	↓
{AH,OW,TK,CL, Tmin,Tmax}	-2534.84	Stop
{AH,OW,TK,CL, Tmin,Tmax,NF}	-2534.8	↑

4.2. Classification

The process of classification is attained with full features and the selected features using ANN algorithm. The process of classification accuracy is compared in this section for the classification algorithm ANN with CSFFS and ANN without CSFFS. Classification accuracy is the accurate prediction of crop yield. The outcome of algorithms are given in Table 4 and the graphical illustration is given in Figure 2.

Table 4. Classification Accuracy with CSFFS and without CSFFS

Iteration	ANN without CSFFS	ANN with CSFFS
50	79	86
100	80	87
150	81	89
200	83	90
250	85	92

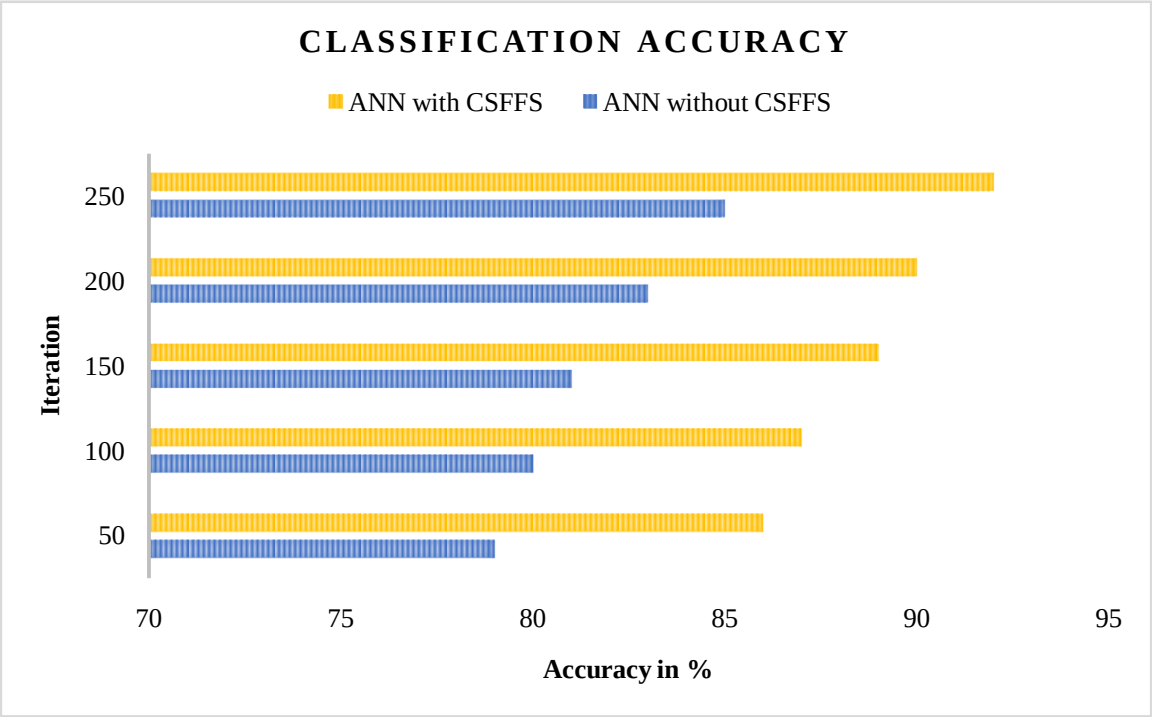

Figure 2. Comparison of Classification Accuracy

The performance of the classification accuracy is noted for diverse iterations and from the observation it is identified that the proposed approach is highly effective with feature selection scheme that attains 92% classification accuracy.

5. Conclusion

The current research used different types of feature selection models, which are most relevant for the yield prediction. The selection criteria for every feature selection algorithm were explained in detail and analyzed. These algorithms were used to identify the best feature subsets from the original agricultural data for the further analysis of the machine learning algorithms and regression models used to calculate the crop yield prediction.A range of production characteristics and algorithms are required for a crop yield prediction research. Some algorithms are used to find the best feature subset for better prediction, while others are used to figure out what to forecast. The suggested Correlation-based Sequential Forward Feature Selection (CSFFS) method is compared to current feature selection methods. Crop forecast accuracy is improved by using the proposed feature selection method.

Reference

1. Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709.

2. Elavarasan, D., & Vincent, P. D. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE Access*, *8*, 86886-86901.
3. Abbas, F., Afzaal, H., Farooque, A. A., & Tang, S. (2020). Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy*, *10*(7), 1046.
4. Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S., & Singh, A. K. (2021). Crop yield prediction integrating genotype and weather variables using deep learning. *Plos one*, *16*(6), e0252402.
5. Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S., & Singh, A. K. (2021). Crop yield prediction integrating genotype and weather variables using deep learning. *Plos one*, *16*(6), e0252402.
6. Qiao, M., He, X., Cheng, X., Li, P., Luo, H., Zhang, L., & Tian, Z. (2021). Crop yield prediction from multi-spectral, multi-temporal remotely sensed imagery using recurrent 3D convolutional neural networks. *International Journal of Applied Earth Observation and Geoinformation*, *102*, 102436.
7. Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, *1*(2), 56-70.
8. Liu, Z., Japkowicz, N., Wang, R., Cai, Y., Tang, D., & Cai, X. (2020). A statistical pattern based feature extraction method on system call traces for anomaly detection. *Information and Software Technology*, *126*, 106348.
9. Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2020). A review of unsupervised feature selection methods. *Artificial Intelligence Review*, *53*(2), 907-948.
10. Chen, C. W., Tsai, Y. H., Chang, F. R., & Lin, W. C. (2020). Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems*, *37*(5), e12553.
11. Hancer, E., Xue, B., & Zhang, M. (2020). A survey on feature selection approaches for clustering. *Artificial Intelligence Review*, *53*(6), 4519-4545.
12. Cekik, R., & Uysal, A. K. (2020). A novel filter feature selection method using rough set for short text data. *Expert Systems with Applications*, *160*, 113691.
13. Karasu, S., Altan, A., Bekiros, S., & Ahmad, W. (2020). A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series. *Energy*, *212*, 118750.
14. Zhou, Y., Cheng, G., Jiang, S., & Dai, M. (2020). Building an efficient intrusion detection system based on feature selection and ensemble classifier. *Computer networks*, *174*, 107247.
15. Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, *143*, 106839.
16. Ren, K., Fang, W., Qu, J., Zhang, X., & Shi, X. (2020). Comparison of eight filter-based feature selection methods for monthly streamflow forecasting–three case studies on CAMELS data sets. *Journal of Hydrology*, *586*, 124897.
17. Kurniawan, Y. I., Cahyono, T., Maryanto, E., Fadli, A., & Indraswari, N. R. (2020, December). Preprocessing Using Correlation Based Features Selection on Naive

Bayes Classification. In *IOP Conference Series: Materials Science and Engineering* (Vol. 982, No. 1, p. 012012). IOP Publishing.

18. Gholamnezhad, P., Broumandnia, A., & Seydi, V. (2020). An inverse model- based multiobjective estimation of distribution algorithm using Random- Forest variable importance methods. *Computational Intelligence*.
19. Folli, G. S., Nascimento, M. H., de Paulo, E. H., da Cunha, P. H., Romão, W., & Filgueiras, P. R. (2020). Variable selection in support vector regression using angular search algorithm and variance inflation factor. *Journal of Chemometrics*, *34*(12), e3282.
20. Bhardwaj, S., Khan, A. A., Muzammil, M., & Khan, S. M. (2020). ANN based classification of sit to stand transfer. *Materials Today: Proceedings*, *24*, 1029-1034.
21. Tiwary, S. K., Pal, J., & Chanda, C. K. (2020, February). Multi-dimensional ANN application for active power flow state classification on a utility system. In *2020 IEEE Calcutta Conference (CALCON)* (pp. 64-68). IEEE.