International Journal of Data Informatics and Intelligent Computing (IJDIIC)
Vol. 3, No. 2, June 2024, pp. 52~58
DOI: 10.59461/ijdiic.v3i2.120

Optimizing Crop Yield Prediction through Multiple Models: An Ensemble Stacking Approach

Renju K[1], Brunda V[1]
[1]Department of Computer Science, Mount Carmel College Autonomous, Bengaluru, Karnataka, India

Article Info

Article history:

Received April 16, 2024
Revised June 08, 2024
Accepted June 17, 2024

Keywords:

AdaBoost Regressor
Decision Tree
Ensemble Learning
Linear Regressor
Stacking Regressor

ABSTRACT

Agriculture plays a pivotal role in enhancing India's economy, providing employment opportunities, supporting various industries, and contributing significantly to livelihoods and rural development. Accurate crop yield prediction is essential for effective crop management, productivity enhancement, and ensuring a balance between supply and demand. Leveraging machine-learning techniques, particularly stacking regressors, can offer improved predictive accuracy by capturing complex relationships between agricultural variables. This research paper conducts a comparative analysis of various machine learning techniques and introduces stacking ensemble learning for predicting crop yields in Indian agriculture. Each ML model, including Decision Tree, AdaBoost Regressor, and Linear Regressor, underwent individual training, testing, and prediction with hyperparameter tuning. Furthermore, the study implemented stacking ensemble learning using Linear Regressor, Decision Tree, and AdaBoost Regressor as base learners, with Linear Regressor serving as the meta-learner. The experimental results demonstrated that the stacking ensemble learning model outperformed all individual ML models, showcasing an impressive R2 value of 98.92%. These findings underscore the efficacy of stacking regressors in enhancing predictive accuracy for crop yield prediction, offering valuable insights for agricultural decision-making and resource allocation in the Indian agricultural sector.

Corresponding Author: Renju K (e-mail: renju.k@mccblr.edu.in)

1. INTRODUCTION

Agriculture plays a pivotal role in society by ensuring food security, contributing significantly to the economy, providing employment opportunities, and supporting various industries through the production of raw materials. In countries like India, where a large portion of the population is engaged in agriculture, agriculture serves as a crucial economic sector, contributing to the GDP and shaping the socio-economic fabric of the nation. The agricultural domain not only sustains livelihoods but also influences income distribution, rural development, and overall societal well-being. Additionally, agriculture is essential for meeting the basic food needs of the population and plays a vital role in ensuring economic stability and growth [1]. Yield prediction is one of the significant topics in precision agriculture, which is crucial for crop management to enhance productivity, yield mapping, yield estimating, and match crop supply and demand. The study's objective was to give growers information relevant to yield so they could maximize their grove's potential for profit and higher output [2].

A variety of machine-learning techniques have been applied to support crop prediction research. Machine learning, with its data-driven approach, can leverage larger amounts of data and capture nonlinear relationships between predictors and returns at a regional level. This model has proposed ensemble stacking

regressor based on three heterogenous machine learning models: multivariate logistic regression, decision tree, and Adaboost regressor [3].

Measuring crop yields, in particular, through agricultural monitoring, is crucial to assessing the level of food security in an area. Crop production is primarily influenced by the climate, the quality of the soil, the landscape, pest infestation, the availability and quality of water, genotype, and crop activity planning. Time-varying and highly nonlinear in nature, crop yield processes and techniques are complex because they incorporate a large number of associated components that are influenced by external factors and non-arbitrate runs. The methods used in machine learning agriculture frameworks come from the process of learning [4].

2. LITERATURE REVIEW

Predicting crop yields is a crucial task for national and regional decision-makers to make quick decisions. Farmers can decide what to grow and when using an accurate yield prediction model and they are conducted from different perspectives. Several research papers were reviewed that demonstrated the potential for using machine learning in crop yield prediction in the literature [5][6].

Pandith et al. [7] predicted mustard plant yield from soil analysis using diverse supervised machine learning techniques, namely K-Nearest Neighbor (KNN), Naive Bayes, Multinomial Logistic Regression, Artificial neural network (ANN) and Random Forest. The result suggests that ANN (artificial neural network) is the most accurate mustard yield prediction technique to help farmers select the right quantity of fertilizers [8]. However, they focused on predicting performance based on a small dataset of about 5000 instances and suggested that crop yield prediction with a huge soil data set can be implemented in a Big Data environment.

Rashid et al. [9] suggested a machine learning based prediction system to anticipate the yield of six crops at the national level in West African countries throughout the course of the year. The climatic, meteorological, agricultural yield and chemical data were merged to help farmers and decision-makers predict the annual crop yields in their respective nations. Decision tree, multivariate logistic regression, and k-nearest neighbour models were utilized to develop their system; along with it, they employed a hyper-parameter tuning technique during cross-validation to improve the model and avoid overfitting issues where decision tree reported the promising outcome of R^2 of 95.3%. Moreover, the best feature set is needed to improve the prediction system because the palm oil yield prediction used very few feature sets, which culminated in significant discrepancies between the expected and real palm oil yield [10].

Nosratabadi et al. [11] conducted research to provide models for predicting agricultural yield contingent on hybrid machine learning techniques. This study focused on the farms located near the city of "Kerman" in Iran, which evaluated the effectiveness of the Artificial Neural Network-Gray Wolf Optimizer (ANN-GWO) and Artificial Neural Network-Imperialist Competitive Algorithm (ANN-ICA) models for crop yield prediction. According to the results, ANN-GWO outperformed the ANN-ICA model in crop yield prediction with an R^2 value of 0.48, RMSE of 3.19 and MEA of 26.65. Moreover, it did not address the usage of ensemble learning methods to identify models with higher efficiency.

Paudel et al. [12] conducted a study in order to establish a machine-learning baseline for large-scale agricultural production forecasting. The project blended machine learning with agronomic concepts of crop modelling. A workflow that gave preference accuracy, decomposition, and reuse was the baseline. It utilized soil, meteorological, and remote sensing data from the MARS Crop Yield Forecasting System (MCYFS), together with crop simulation results. However, by including other data sources, creating greater likelihood features, and testing various machine learning algorithms, the baseline could be enhanced.

3. METHOD

3.1. Data collection and pre-processing

This study was based on agricultural data from different states in India from 1997 to 2020. The data was collected from various sources, including government websites and the Open Government Data (OGD) Platform India. The target variable of the models was nationwide crop yield data consisting of 20000 instances. The characteristic parameters that were chosen for the prediction models had a strong influence on most types of agriculture in the region, which included nominal features such as types of crops that are grown, seasons, and states in India and numeric features that are described in table 1 below.

Following data collection, pre-processing culminated in imputing missing data and duplicate data, using statistical techniques to identify any outliers in the data and scaling using a min-max scalar from the sci-kit learn library. We performed a multivariate analysis to examine the multi-correlation between the parameters using the Pearson formulae. Figure 1 shows the graph of the correlation between parameters. We remark that the yields are more correlated with production than with fertilizer and pesticide parameters. This shows that the area has a direct correlation with fertilizer and pesticide of 0.97 and has a negative correlation with annual rainfall [13].

Table 1. Description of numeric features in the dataset

Sl. No	Numeric Parameters	Description	Units
1.	Area	Total land area under cultivation for the specific crop	hectares
2.	Production	Quantity of crop production	metric tons
3.	Annual rainfall	Annual rainfall received in the crop-growing region	mm
4.	Fertilizer	Total amount of fertilizer utilized for the crop	kilograms
5.	Pesticide	Total amount of pesticide applied for the crop	kilograms

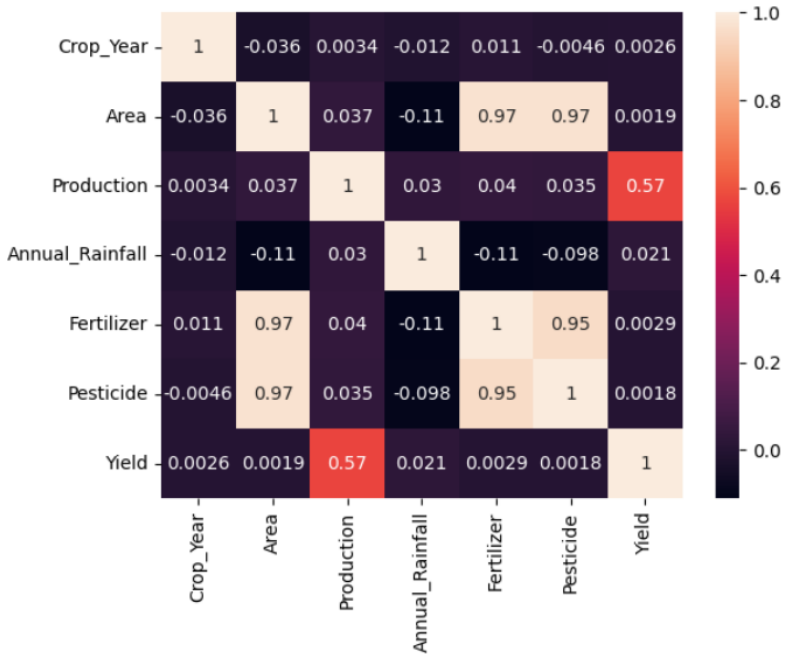


Figure 1. Heat map representing the co-relation of the features

During our data exploration and analysis process, there were to categorical columns which was necessary to them into numerical form for further model building. Categorical features which include type of crop, seasons and states were transformed into binary (0's and 1's) with columns in the output are each named after the value using get dummies method from pandas library.

3.2. Model Building

3.2.1. Linear Regression

LR is a modelling technique used in statistics to establish connections between an independent variable using one specific or heavily reliant variable. In LR, predefined formulas are used to calculate features like slope, Y-intercept, and regression coefficient. However, the way the LR method functions in machine learning differs from that of traditional statistics. When it comes to machine learning, LR employs data and techniques like gradient descent to minimize losses (also known as RMSE or MSE). The gradient descent technique fits the models with minimized loss functions, depending on the type of data, improving the model's prediction accuracy. This formula is commonly used to define LR:

$$y = a + bx \quad (1)$$

Where a is intercept and b is slope of a regression line. The main idea is to obtain a line that best fits the data. The best-fit line has the least total predictor error [14].

3.2.2. Decision Tree

A multi-level, hierarchical decision system or a structure like a tree is a cornerstone of a decision tree. The nodes that hold up the tree are a root node, which constitutes all of the data, a multitude of internal nodes, also known as splits, and numerous end nodes, also known as leaves. Every decision node in the decision tree makes a choice in a binary fashion that divides a class or a subset of classes from the other classes. The implicit

premise of the decision tree regression approach is that either linear or nonlinear correlations exist between targets (i.e., yield) and attributes (i.e., crop data) (e.g., in logistic regression). This allows complex nonlinear relationships to be managed. Decision tree regression is used to approximate real-valued functions such as class proportions. It uses binary recursive partitioning to split data into partitions, with each new branch applying the splitting process until it reaches a user-specified minimum node size and forms a terminal node. This iterative process minimizes squared deviations from the mean [15] [16].

3.2.3. AdaBoost Regressor

Adaptive Boosting Regression (ABR) is a machine learning technique that randomly combines weak learners from a dataset to create a strong learner. It is a sequential technique where weight is assigned to all the training points. After that, choosing the weak learner and assigning the higher weight continues to get the best prediction. It assigns weights to each sample observation, identifying false predictions and assigning them to the next base learner. This course of action occurs over and over by the algorithm until the output is correctly classified [16].

$$C_m(x) = \sum_{i=1}^{m}. \alpha_m K_m(x) \quad (2)$$

Where $K_m(x)$ is the prediction made by the stump we trained at iteration m, *and* α_m is the confidence we place on the predictive power of stump m.

3.2.4. Stacking Regressor

Ensemble models combine different learning models to improve the results of each individual model. Ensemble learning involves applying multiple learning modules to a data set to extract multiple predictions, which are followed by carrying out a composite prediction. This process typically involves two phases: extracting basic learners from training data and combining them to create a unified predictive model. Ensemble learning is successful in machine learning for three main reasons: statistical, computational, and representational. Statistically speaking, ensemble methods help avoid selecting the best hypothesis from limited data sets, while ensemble methods provide computationally better approximations of the true unknown function. Figure 2 explains the implementation of a stacking regressor, which uses various learning algorithms to build its models and then trains a combination algorithm to produce the most accurate predictions by leveraging the predictions made by the base algorithms [17].

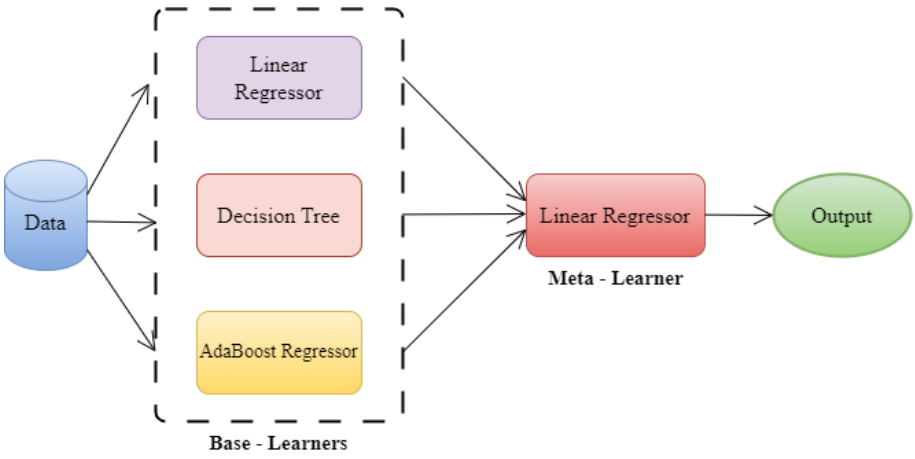


Figure 2. Overview of stacking regressor

3.3. Implementation

The data samples were divided into two sets, 20% and 80%, to create the training and testing sets. Linear regressor, decision tree and Adaboost regressor were trained as base learners. Combining multiple base learners reduces overfitting and handles nonlinearity and complexity, improving predictive performance by leveraging their complementary strengths. Stacking enables a flexible selection of base learners based on problem domain, data properties, and modelling requirements. It is critical in stacking for their ability to provide improved generalization and stability.

Grid search was used to determine the hyperparameters of the machine learning algorithms. It is essential to test various hyperparameters for various datasets because they behave differently depending on the dataset. Grid search uses hyperparameters to find the best combination for a model. It involves defining

parameter grids as highlighted in table 2, performing k-fold cross-validation on training data, training the model for each hyperparameter combination, evaluating the model's performance using the validation set of each fold, and selecting the best model.

Table 2. Parameter grids for hyper tuning

Algorithm	Hyperparameter	Parameter grids
Linear Regressor	Alpha	[0.1, 0.5, 1.0, 5.0, 10.0]
Decision Tree	Maximum depth	[None, 5, 10, 15]
	Minimum samples Split	[2, 5, 10]
	Minimum samples leaf	[1, 2, 4]
AdaBoost Regressor	Estimator	Decision Tree Regressor
	N estimators	[50, 100]
	Learning rate	[0.5, 1]

The testing process was further improved by implementing the k-fold cross-validation method, which evaluates the ability of an ML algorithm to handle novel and unknown data. With this method, the data set is randomly divided into k groups of approximately equal size. The data is trained on k-1 folds, with the first fold serving as the test set. For each data set in this study, three folds (k = 5) were examined [14].

The meta-regressor is a crucial component in stacking ensemble techniques because it effectively combines the predictions of base regressors, resulting in improved prediction performance. This is achieved through model aggregation, a balance between bias and variance, and the ability to deal with model errors. The meta-regressor also improves generalization performance by learning from the patterns and relationships in base regressor predictions. Stacking, a form of ensemble learning, leverages the power of multiple models to improve performance. The linear regressor was chosen as the meta-regressor, although the choice of the meta-regressor is flexible and depends on the problem and the data properties. Overall, the meta-regressor improves the effectiveness and performance gains achieved by the stacking ensemble technique.

4. RESULTS AND DISCUSSION

Each model was trained, tested and predicted individually using the hyperparameter tuning. The decision tree and AdaBoost regressor predicted the highest R2 score of 98.92% and 96.47%, respectively, whereas the linear regressor predicted the lowest accuracy of 78.42. %. Table 3 presents the comparative analysis for all ML techniques under study.

Table 3. Performance evaluation of Stacking Regressor

Algorithm	R2	MAE	RMSE
Linear regression	0.7842	65.64	409.16
Decision Tree	0.9647	9.19	165.43
Adaboost regression	0.9782	6.23	165.43

These techniques have been evaluated on the basis of R2 (Coefficient of Determination), MAE (Mean Absolute Error), and RMSE (Root Mean Squared Error). Among all the undertaken ML techniques for the study, Decision tree and AdaBoost regressor predicted best performance when trained individually. Figure 3 illustrates the summary of comparative results of all ML techniques under study.

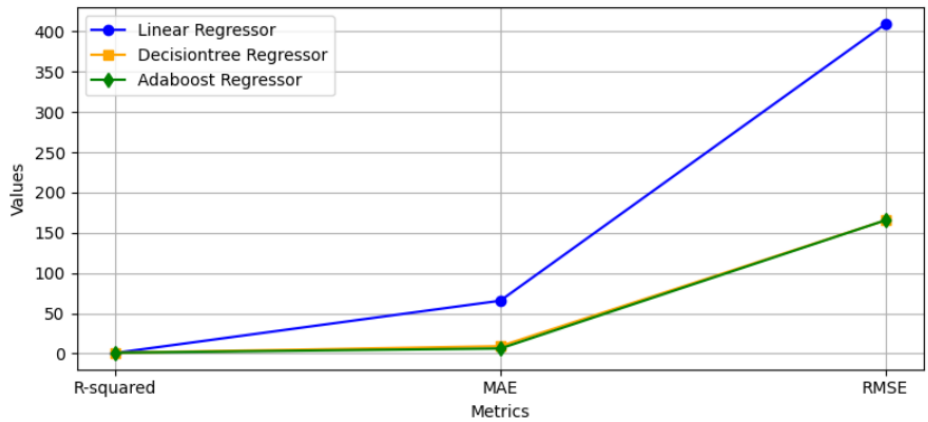


Figure 3. Comparison of results using metrics

Experimental results after using stacking ensemble learning showed that it clearly outperformed compared to all the machine learning models when tested individually. The R2 (Co-efficient of Determination) value was found to be 98.92% with MAE (Mean Absolute Error of 6.18 and RMSE (Root Mean Squared Error) of 91.21.

5. CONCLUSION

Research on the effectiveness of stacking regressors in crop yield prediction compared to traditional methods, exploring their ability to capture complex relationships between agricultural variables, and assessing their robustness across different crops and regions is needed to improve accuracy. Stacking regressor combines predictions from diverse base models using a meta-learner to achieve potentially improved predictive performance. Statistical characteristics were utilized in modelling procedures to produce yield predictions with regard to the gathered data. The stacking regressor model consistently outperforms traditional regression techniques and individual base models in terms of predictive accuracy.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest in this work.

REFERENCES

[1] B. M. Sagar and N. K. Cauvery, "Agriculture data analytics in crop yield estmation: A critical review," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 3, pp. 1087–1093, Dec. 2018, doi: 10.11591/ijeecs.v12.i3.pp1087-1093.

[2] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors (Switzerland)*, vol. 18, no. 8. MDPI AG, Aug. 14, 2018. doi: 10.3390/s18082674.

[3] D. Paudel *et al.*, "Machine learning for regional crop yield forecasting in Europe," *Field Crops Res*, vol. 276, Feb. 2022, doi: 10.1016/j.fcr.2021.108377.

[4] D. Elavarasan and P. M. Durairaj Vincent, "Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications," *IEEE Access*, vol. 8, pp. 86886–86901, 2020, doi: 10.1109/ACCESS.2020.2992480.

[5] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177. Elsevier B.V., Oct. 01, 2020. doi: 10.1016/j.compag.2020.105709.

[6] G. Lischeid, H. Webber, M. Sommer, C. Nendel, and F. Ewert, "Machine learning in crop yield modelling: A powerful tool, but no surrogate for science," *Agric For Meteorol*, vol. 312, Jan. 2022, doi: 10.1016/j.agrformet.2021.108698.

[7] V. Pandith, H. Kour, S. Singh, J. Manhas, and V. Sharma, "Performance Evaluation of Machine Learning Techniques for Mustard Crop Yield Prediction from Soil Analysis," *Journal of scientific research*, vol. 64, no. 02, pp. 394–398, 2020, doi: 10.37398/jsr.2020.640254.

[8] D. A. Bondre and M. Santosh Mahagaonkar, "PREDICTION OF CROP YIELD AND FERTILIZER RECOMMENDATION USING MACHINE LEARNING ALGORITHMS," 2019. [Online]. Available: http://www.ijeast.com

[9] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin, and N. Khan, "A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches with Special Emphasis on Palm Oil Yield Prediction," *IEEE Access*, vol. 9. Institute of Electrical and Electronics Engineers Inc., pp. 63406–63439, 2021. doi: 10.1109/ACCESS.2021.3075159.

[10] L. S. Cedric *et al.*, "Crops yield prediction based on machine learning models: Case of West African countries," *Smart Agricultural Technology*, vol. 2. Elsevier B.V., Dec. 01, 2022. doi: 10.1016/j.atech.2022.100049.

[11] S. Nosratabadi, K. Szell, S. Ardabili, B. Beszedes, and A. Mosavi, "Hybrid Machine Learning Models for Crop Yield Prediction."

[12] D. Paudel *et al.*, "Machine learning for large-scale crop yield forecasting," *Agric Syst*, vol. 187, Feb. 2021, doi: 10.1016/j.agsy.2020.103016.

[13] L. S. Cedric *et al.*, "Crops yield prediction based on machine learning models: Case of West African countries," *Smart Agricultural Technology*, vol. 2. Elsevier B.V., Dec. 01, 2022. doi: 10.1016/j.atech.2022.100049.

[14] F. Abbas, H. Afzaal, A. A. Farooque, and S. Tang, "Crop yield prediction through proximal sensing and machine learning algorithms," *Agronomy*, vol. 10, no. 7, Jul. 2020, doi: 10.3390/AGRONOMY10071046.

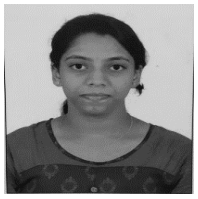
[15] M. Xu, P. Watanachaturaporn, P. K. Varshney, and M. K. Arora, "Decision tree regression for soft classification of remote sensing data," *Remote Sens Environ*, vol. 97, no. 3, pp. 322–336, Aug. 2005, doi: 10.1016/j.rse.2005.05.008.

[16] G. Shanmugasundar, M. Vanitha, R. Čep, V. Kumar, K. Kalita, and M. Ramachandran, "A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining," *Processes*, vol. 9, no. 11, Nov. 2021, doi: 10.3390/pr9112015.

[17] F. Divina, A. Gilson, F. Goméz-Vela, M. G. Torres, and J. F. Torres, "Stacking ensemble learning for short-term electricity consumption forecasting," *Energies (Basel)*, vol. 11, no. 4, Apr. 2018, doi: 10.3390/en11040949.

BIOGRAPHIES OF AUTHORS

Renju K received a master's degree in computer application and a master's degree in philosophy. She has worked for 17 years as an Assistant Professor at Mount Carmel College Autonomous, Bengaluru, Karnataka, India. She works as a Head of the Department, Department of Computer Science at Mount Carmel College Autonomous, Bengaluru, Karnataka, India. Her research interests include deep learning and machine learning. Published many papers in various national and international conferences and Journals. Her research includes audio and signal processing using deep learning. She can be contacted at email: renju.k@mccblr.edu.in.



Brunda V, Research Scholar, Department of Computer Science, Mount Carmel College Autonomous, Bengaluru, Karnataka, India. Received a B.Sc. (Computer Science, Mathematics, and Electronics) from St Joseph's University, Bengaluru, Karnataka, India. Her research interests include machine learning, data mining, cloud computing, and big data analytics. Her previous research includes Automated Systems Operations in control systems. She can be contacted at email: brundavadiga@gmail.com