SEUSL

Sri Lankan Journal of Technology

Research

SLJoT
2022 Sp Issue 19 - 24
ISSN 2773-6970

An Ensemble Approach to Predict Construction Delays in Road Projects of Sri Lanka

S.M.Safnas[1] and A.R.F.Shafana[2,3*]

[1]Department of Civil Engineering, University of Peradeniya, Sri Lanka
[2]Department of Information and Communication Technology, South Eastern University of Sri Lanka, Sri Lanka
[3]School of Engineering & Technology, Asian Institute of Technology, Thailand.

[*]Corresponding Author: arfshafana@seu.ac.lk || ORCID: 0000-0001-6926-9196

Received: 05-10-2022 * Accepted: 02-11-2022 * Published Online: 30-11-2022

***Abstract*—**The construction industry is often dynamic and quite complex. Due to its dynamic nature, delays have been witnessed in many construction projects across the world. Despite having several mitigation techniques, the delay has not been well-addressed, thus leading to various issues for the contractors in obtaining the construction claims that arise from cost overruns. This study utilizes the machine learning approach to predict the construction delays in road overlaying projects in Sri Lanka. In particular, the methodology adopted is an ensemble approach that constitutes the benefits of multiple machine learning algorithms and provides a collective advantage for the prediction. The study has been successful in obtaining a predictive model based on a stacking approach that is prone to lesser error than the individual algorithms concerned.

***Keywords*—**Construction Delay, Stacking, Boosting, Bagging, MLens, Predictive Analytics

I. Introduction

The construction industry is quite dynamic and complex. Owing to its complexity and uncertainty, the delays have become a common phenomenon in any construction projects (Yaseen *et al*., 2020). The road construction industry is not an exception in this case. In specific, for a developing country like Sri Lanka, road construction and overlaying projects occupy a prominent place in the construction sector since a major portion of the national budget is channeled to them (Kaliba, Muya and Mumba, 2009). Despite being an important industry, this suffers from cost overruns mainly due to the delays in such projects which in turn brings an overall dissatisfaction towards the overall performance. Delays in projects are well described as the additional time required beyond the scheduled timeline of project completion and is the key cause of many cost overruns (Haq, Rashid and Aslam, 2014).

Past studies revealed that nearly 9 out of 10 large-scale projects suffer from project delays (Flyvbjerg, 2014; Rhodes, 2015). In the context of Sri Lanka, this is even worse where between 78% to 90% of construction projects encounter delays (Ramachandra, Rotimi and Gunaratne, 2014). Road construction projects also suffer from delays greatly to a percentage around 56% to 88% (Pathiranage, 2010) while Jahankeer, (2016) emphasized that the on-going projects would still face similar issues. This effect is well pronounced for contractors as they get affected by the overhead costs directly. This increases the importance of the construction claims in the contractors' side through which they could request for additional payment by requesting an Extension of Time (EOT) (Alnaas, Khalil and Nassar, 2014). At times, the EOT is rejected mainly being not adhering to the time schedule to which the delay notification should be sent (Perera *et al*., 2019). Thus, it is always wise for the contractors to predict the delay well in advance, so they do not get affected adversely with the rejection of construction claims.

In response to the emerging needs of predicting construction delays, several studies investigated the factors influencing the construction delays. Jayalath (2019) investigated the factors contributing to the cost overruns in Sri Lankan Road projects and listed out the following 10 factors as the key causes such as market vacillations, corruptions, scope creep, shortage of materials, delays from sub-contractors, lack of pre-contract project coordination, labour scarcity, changes in specification, labour disputes and insufficient quantity of skilled labourers. The identified factors were reliable and consistent as supported by a higher Cronbach's alpha (0.96). In Sri Lankan contexts, it appears that the factors contributing to the cost overruns changes radically with the passage of years, since the factors outlined by Pathiranage (2010) a decade ago is well delineated from that of Jayalath (2019). The former study identified that the financial problems from the contractor and the client is the most influential towards delays. Other features identified as critical are but not limited to the lack of proper site management by the contractor, inclement weather conditions, client's modifications in the

Copyright ©2022 belongs to Faculty of Technology, South Eastern University of Sri Lanka, University Park, Oluvil, #32360, Sri Lanka

contract, incomplete documents, and the slow response towards decision making. Pathiranage (2010) also enlisted that the lack of site labourers and materials, poor skills among sub-contractors, mistakes in design or work, less skilled or inexperienced labourers and the delay caused in delivery of materials to the working site are other causes. Thus, we postulate that the factors contributing to the cost overruns resulting from construction delays are susceptible to changes and the timely assessment is crucial.

Contemporary studies have pointed out various factors towards the construction delays. Kesavan *et al*. (2015) identified that delays are caused under several categories such as those caused by contractor, consultant, client, labourer, material, equipment and external causes. Conflicts occurred in the schedule of sub-contractor, delay in approval of the major work revisions, delay in progress payments, lack of labourers, delay in the delivery of material, lack of technology in equipment and the inclement weather condition were pointed as the major factors respectively in each of the category. The study used the Relative Importance Index for identifying the most important factor. Wijekoon & Attanayake (2012) also made a significant contribution in identifying the potential causes of delays in road construction projects in Sri Lanka. The identified causes were very much similar to that listed by Pathiranage(2010) and Kesavan *et al*. (2015). Thus, the assessment of all the features identified across the decade in a more comprehensive way could yield a better model that can precisely forecast the road project delays, which is lacking, to date.

Predictive Models have been a wise choice of predicting the delays in the construction industry as a measure of risk mitigation which in turn leads to the timely payment of construction claims (Gondia *et al*., 2020; Yaseen *et al*., 2020). Gondia *et al*. (2020) was able to improve project delay risk assessments precisely and forecast them in building constructions of Egypt. The study utilized Decision Tree and Naïve Bayesian Classifiers and obtained an accuracy of 74.5% and 78.4% respectively. In another study by Asadi *et al*. (2015), the above algorithms were used to assess the delays in construction logistics in Qatar and obtained an accuracy of 79.41% for Decision Tree and 73.52% for Naïve Bayes. However, a combined approach by Yaseen *et al*. (2020) yielded relatively a better model with an accuracy of 91.67%. In the latter study, Genetic Algorithm and Random Forest algorithms have been ensembled to obtain such a better model for predicting construction delays in Iraq. Furthermore, the work of Egwim *et al*. (2021) is very significant in this series as the study utilized the ensemble of ensemble predictive models in construction projects in Nigeria. The study claims that this approach is quite successful as the accuracy of this model was almost around 80% while the accuracy of any single algorithm used were significantly lower. Thus, it could be noted that the ensemble approach possesses the combined advantage of single algorithms in terms of performance and robustness.

Despite much research have attempted to identify the potential causes of road construction delays in many countries using machine learning approach, not many efforts have been attempted in the context of Sri Lanka. Moreover, the review of the existing literature suggests that the ensembling technique could yield a robust predictive model. In this ground, this paper proposes a robust predictive model that is woven around the Machine Learning algorithms, an evolving field that has been adopted in several industries as a strong instrument for prediction (Blanco *et al*., 2018). This study in particular utilizes the ensemble approach to develop a multilayer robust predictive model to predict the construction delays in the road projects of Sri Lanka which is based on 24 features that are considered as strong predictors in the literature and from expert reviews. Multiple approaches of ensembling regressors were investigated and stacking has been identified as the best approach. Thus, the predictive model is built upon stacking of the robust regressors identified.

The contribution of this study is two-fold. Firstly, the computational benefit from machine learning algorithms is applied in Sri Lankan Road construction context, which is still lacking to date. The study has identified 24 critical features pinpointed as major causes of construction delays in Sri Lankan scholarly works and was able to identify the most important features using a robust predictive model. Secondly, the study has applied most of the important regressors to date, and forwards that the application of ensemble approach is the most suitable for prediction. Thus, this study further corroborates the combined advantage from ensemble algorithms and suggests its usage in various other contexts.

II. Methodology

A. Data Collection and Preprocessing

From the review of literature on important predictors of the construction delays in Sri Lanka, a set of 24 factors as shown in Table I were identified. A questionnaire adopted from Egwim *et al*. (2021) was used to obtain the answers for the 24 identified features from the potential respondents in the construction industry. The target variable for the above study is the construction delay. Convenient Sampling was used to collect the data due to its accessibility and geographic proximity. The 75 respondents to the survey were stakeholders from the road construction projects in Sri Lanka who have at least 3 years of experience in working in major projects. The respondents were instructed to answer the questionnaire based on any single road project that they have worked in the past. The responses were collected using the Google Form as well as by distributing the questionnaire in person.

The data thus obtained have been pre-processed and converted into an analyzable form. The resulting dataset is a 2D array with 25 columns and 75 rows. The columns ranging from first to 24th are of the factors and the target is listed in the final column. As the questionnaire had the percentage of delay as the response, encoding was done on the 25th column such that the values that are less than or equal to 3

Copyright ©2022 belongs to Faculty of Technology, South Eastern University of Sri Lanka, University Park, Oluvil, #32360, Sri Lanka

Table I: Features and Target Variable

Feature ID	Features
F1	Project Size
F2	Equipment
F3	Inflation
F4	Labour Disputes or Strikes
F5	Poor Communication
F6	Inclement or Bad weather
F7	Contractor's Financial Difficulties
F8	Design Variations
F9	Late deliveries
F10	Changed Orders
F11	Price Fluctuation
F12	Project Management
F13	Slow Decision Making
F14	Cash Flow
F15	Government Regulations
F16	Material Procurement
F17	Site Condition
F18	Political Influence
F19	Project Schedule
F20	Site Accident
F21	Project Quality Control
F22	Late Payment
F23	Proportion of unskilled laborer
F24	Late Delivery
F25(Target)	**Delay**

have been considered as no delay and were given a score of zero while for the values with 4 and 5 have been identified as delayed projects and were coded as one.

B. Correlation Analysis

The goal of the correlation analysis was to assess the multicollinearity among predictors Egwim *et al*. (2021). The correlation matrix is as presented in Figure 1. The results suggest that the cross correlation exist between each of the following features with the target.

C. Reliability Analysis

Reliability analysis was primarily done to identify the accuracy of the data obtained through questionnaire while assessing if there exists an internal consistency coefficient among the data. The analysis also helps to identify if the consolidated features could predict the delay better. The study used the Alpha Test of Cronbach for this purpose and obtained a coefficient of 0.92 which suggests that the responses were accurate towards the factor as it is considered that this value should surpass 0.8 for having an internal consistency (Nunnally JC, 1994)

D. Standardization and Feature Selection

The preprocessed data was standardized to have a zero mean and unit variance such that the data was normally distributed (Pedregosa *et al*., 2011). Chi-Squared Test, a multivariate filter-based feature selection, was used to improve the model accuracy by removing the features that are redundant and noisy for the model. The Figure 2 shows the association of target and the features. Chi-Squared Test was done with 95% confidence interval while $p < 0.05$ were removed to retain only the limited yet significant features.

Table II: Algorithms used in the study

Technique	Algorithms
Base estimator	Decision Tree
Bagging	Lasso
	ElasticNet
	KernelRidge
	LinearRegression
	RandomForest
Boosting	AdaBoost
	XGBoost
	LGBM Regressor
	Gradient Boosting
Naive Bayes	Bernoulli NB
	Gaussian NB
	Multinomial NB
Stacking	Stacking
Ensemble of Ensemble	MLens

This resulted with only 09 features such as 'F8', 'F9', 'F10', 'F11', 'F14', 'F18', 'F21', 'F23' and 'F24' which were retained in the final model prediction. This resulting dataset was split for the training set and the validation set as a 60% and 40% ratio.

E. Ensemble Machine Learning Algorithms

Ensembling is a technique to use multiple algorithms that result in an overall better performing model since the integrated outcome from ensembling is always greater in terms of predictive accuracy while using a single algorithm. The Root Mean Squared Error (RMSE) metric has been used to validate and compare the performance of the models. The machine learning algorithms that have been used for ensembling are presented in the following table (Table II).

1) Bagging: Bagging is the process of combining the results of multiple models in order to get a generalized result. The basic regressors such as linear regression, lasso regression, elastic net and kernel ridge were performed and finally using the BaggingRegressor from sci-kit learn was used to bag while having the Decision tree as the base estimator.

2) Boosting: Boosting is another ensembling technique that uses a sequential process where subsequent model attempts to correct the errors from the previous model. XGBoost, LightGBM, AdaBoost Regressor and Gradient boosting were used for this choice of technique.

3) Naive Bayes: It is yet another technique and one of the most effective and efficient inductive ensemble machine learning algorithms. The study checks the Gaussian, Bernoulli and the Multinomial regression from sci-kit Learn package.

4) Stacking: Stacking as the name suggests stacks the predictions from multiple models and build a new robust model. The built model is tested on the test data for its accuracy.

5) Ensemble of Ensemble: The MLens (Machine Learning Ensemble) is a memory-efficient parallel ensemble learning technique. This is similar to a neural network that ensembles several layers to support many different ensemble architectures.

Copyright ©2022 belongs to Faculty of Technology, South Eastern University of Sri Lanka, University Park, Oluvil, #32360, Sri Lanka

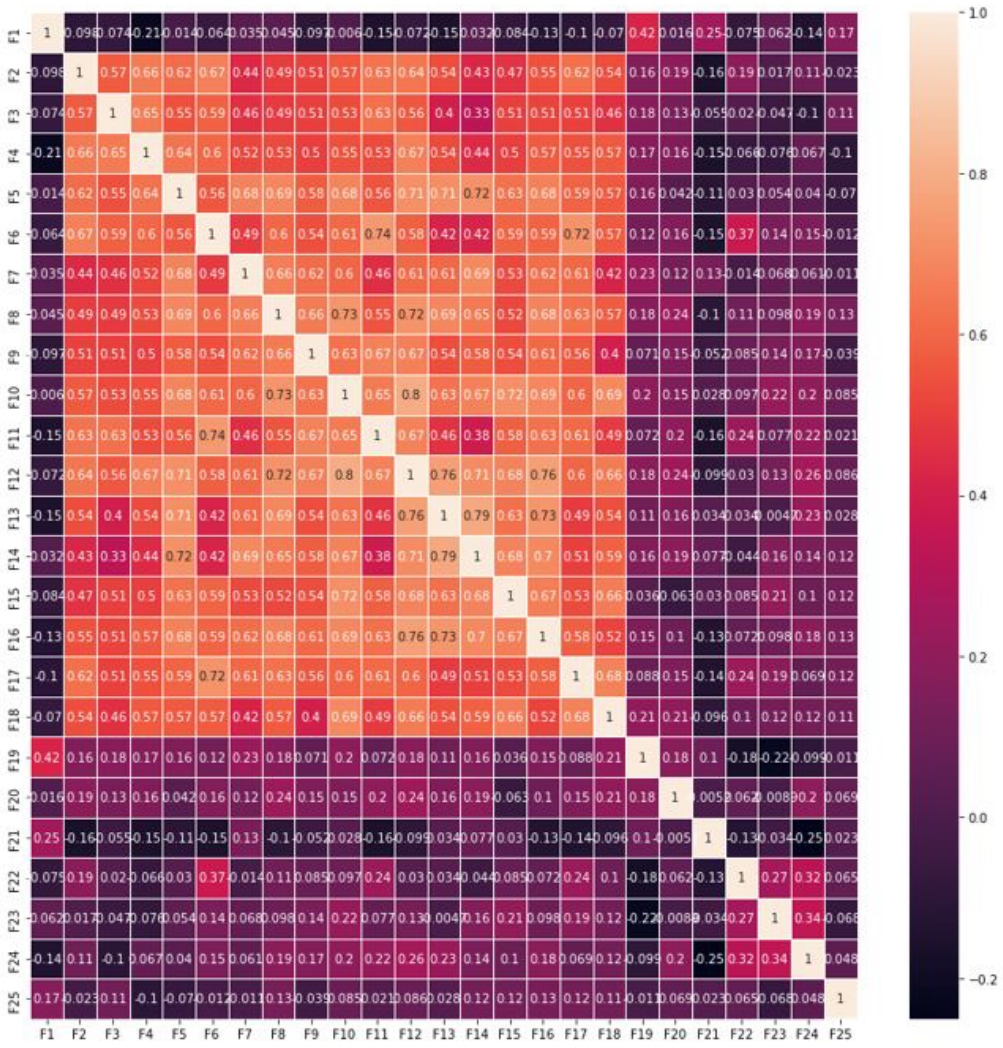

Figure 1: Correlation Matrix

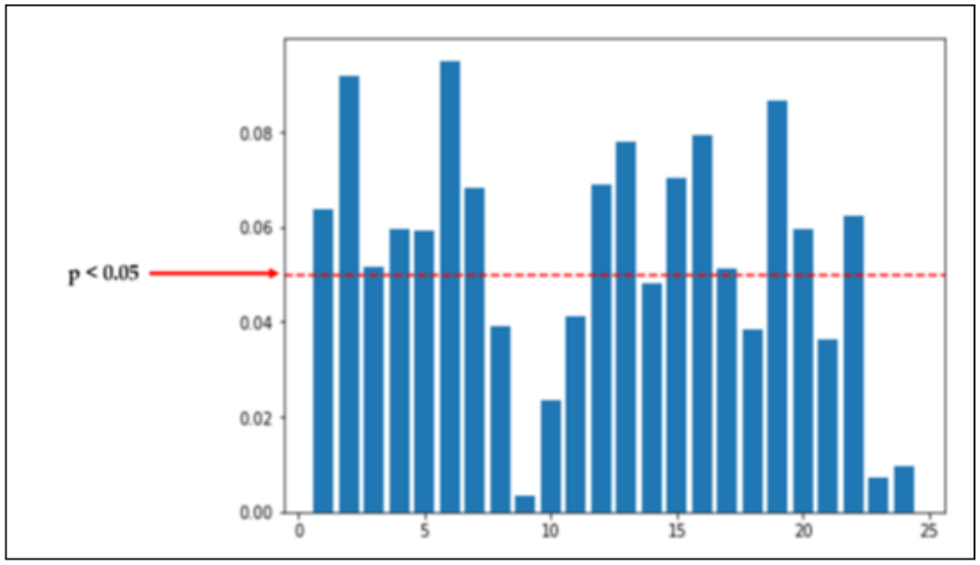

Figure 2: Chi-Squared Test

Copyright ©2022 belongs to Faculty of Technology, South Eastern University of Sri Lanka, University Park, Oluvil, #32360, Sri Lanka

Table III: Ranking based on RMSE score

Rank	Regressor	RMSE score
1	**Stacking**	**0.498860**
2	LGBMRegressor	0.518934
3	RandomForestRegressor	0.566105
4	AdaBoostRegressor	0.574436
5	Ensemble of Ensemble	0.575317
6	GradientBoostingRegressor	0.580583
7	Bagging	0.581951
8	XGBoost	0.606383
9	Lasso	0.616841
10	ElasticNet	0.617136
11	KernelRidge	0.617185
12	LinearRegression	0.621085
13	GaussianNB	0.683130
14	BernoulliNB	0.683130
15	MultinomialNB	0.683130
16	DecisionTree	0.707107

III. Results and Discussion

Root Mean Squared Error was calculated for each of the ensembling model and the model with the least mean squared error was chosen as the best predictive model for predicting construction delays in Sri Lanka. Table III provides the performance of each of the model in increasing order of their RMSE. It is evident from the below table that the Stacking performs as a robust regressor.

Although Bagging and Boosting provide an overall improvement with respect to their individual algorithms concerned, the power of Stacking is well witnessed here as the stacking yield a relatively lower RMSE. Although weak regressor like Decision Tree was also included in the stack, the results prove that the stacking is one of the best techniques to develop an ensemble of classifiers.

IV. Conclusion and Future Works

This study has been successful in its primary objective of identifying a robust model for predicting the road construction delays in Sri Lanka using an ensemble of regressors. Since the notification of the delays in well-advance is an optimal solution to overcome the issue of the rejection of construction claims, this model could be used to assess and forecast road construction delays, whereby the risk could be mitigated to a certain extent. Moreover, another technical contribution of this study is the discernment of the combined advantage of ensemble approaches in machine learning models, suggesting the usage of ensemble approaches in various prediction contexts in place of individual algorithms.

Despite being a robust model thus far developed to predict the road construction delays in Sri Lanka, due to the limited computational power and the time, hyperparameter optimization has not been experimentally done and the values were adopted from existing literature alone. Thus, this is a potential limitation of this work, and the work can be further improved by applying hyperparameter optimization techniques like Grid Search and the results could be further improved.

References

Alnaas, K. A. A., Khalil, A. H. H. and Nassar, G. E. (2014) "Guideline for preparing comprehensive extension of time (EoT) claim," *HBRC Journal, 10*(3), pp. 308–316. doi: 10.1016/J.HBRCJ.2014.01.005.

Asadi, A., Alsubaey, M. and Makatsoris, C. (2015) "A machine learning approach for predicting delays in construction logistics," *International Journal of Advanced Logistics, 4*(2), pp. 115–130. doi: 10.1080/2287108X.2015.1059920.

Blanco, J. *et al*. (2018) "Artificial intelligence: Construction technology's next frontier," *Building Economist*, pp. 7–13.

Egwim, C. N. *et al*. (2021) "Applied artificial intelligence for predicting construction projects delay," *Machine Learning with Applications, 6*, p. 100166. doi: 10.1016/J.MLWA.2021.100166.

Flyvbjerg, B. (2014) "What you Should Know about Megaprojects and Why: An Overview:," https://doi.org/10.1002/pmj.21409, 45(2), pp. 6–19. doi: 10.1002/PMJ.21409.

Gondia, A. *et al*. (2020) "Machine Learning Algorithms for Construction Projects Delay Risk Prediction," *Journal of Construction Engineering and Management, 146*(1). doi: 10.1061/(ASCE)CO.1943-7862.0001736.

Haq, S., Rashid, Y. and Aslam, M. S. (2014) "(PDF) Effects of Delay in construction Projects of Punjab-Pakistan: An Empirical Study," *Journal of Basic and Applied Scientific Research, 4*(4).

Jahankeer, M. Z. M. (2016) Factors Causing Delay in Road Construction Project in Sri Lanka Jayalath, C. (2019) "Severity Analysis Of The Factors Influencing Cost Overrun In Road Projects," *International Journal of Advanced Research and Publications, 3*(7). Available at: www.ijarp.org (Accessed: August 16, 2022).

Kaliba, C., Muya, M. and Mumba, K. (2009) "Cost escalation and schedule delays in road construction projects in Zambia," *International Journal of Project Management, 27*(5), pp. 522–531. doi: 10.1016/J.IJPROMAN.2008.07.003.

Kesavan, M., Gobidan, N. and Dissanayake, P. (2015) "Analysis of Factors Contributing Civil Engineering Project Delays in Sri Lanka." Available at: http://dl.lib.uom.lk/handle/123/11597 (Accessed: August 12, 2022).

Nunnally JC (1994) Nunnally: *Psychometric theory 3E* . Tata McGraw-hill Education. Pathiranage, Y. (2010) "Factors influencing the duration of road construction projects in Sri Lanka." Available at: http://dl.lib.uom.lk/handle/123/1934 (Accessed: August 12, 2022).

Copyright ©2022 belongs to Faculty of Technology, South Eastern University of Sri Lanka, University Park, Oluvil, #32360, Sri Lanka

Pedregosa, F. *et al*. (2011) "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research, 12*, pp. 2825–2830.

Perera, B. A. K. S. *et al*. (2019) "Improving the efficacy of delay notification process of construction projects in Sri Lanka," 21(7), pp. 755–768. doi: 10.1080/15623599.2019.1581593.

Ramachandra, T., Rotimi, J. and Gunaratne, S. (2014) "Reasons for contractors' delay claims failures in Sri Lanka," in *Proceedings of the 30th Annual Association of Researchers in Construction Management Conference*.

Rhodes, C. (2015) "The construction industry: statistics and policy." Available at: https://commonslibrary.parliament.uk/research-briefings/sn01432/ (Accessed: August 12, 2022).

Wijekoon, S. and Attanayake, A. (2012) "Study on the cost overruns in road construction projects in Sri Lanka." Available at: http://dl.lib.uom.lk/handle/123/8969 (Accessed: August 12, 2022).

Yaseen, Z. M. *et al*. (2020) "Prediction of Risk Delay in Construction Projects Using a Hybrid Artificial Intelligence Model," *Sustainability 2020, Vol. 12*, Page 1514, 12(4), p. 1514. doi: 10.3390/SU12041514.

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. Te images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Copyright ©2022 belongs to Faculty of Technology, South Eastern University of Sri Lanka, University Park, Oluvil, #32360, Sri Lanka