A Stacking Ensemble Learning Model for Waste Prediction in Offset Printing

Alexandros S. Kalafatelis*
National and Kapodistrian University of Athens, Psachna, Evia, 34400, Greece
akalafat@core.uoa.gr

Chris Trochoutsos
Pressious Arvanitidis, Chalandri, Athens, 15232, Greece
chtrox@pressious.com

Anastasios E. Giannopoulos
National and Kapodistrian University of Athens, Psachna, Evia, 34400, Greece
angianno@uoa.gr

Angelos Angelopoulos
National and Kapodistrian University of Athens, Psachna, Evia, 34400, Greece,
a.angelopoulos@uoa.gr

Panagiotis Trakadas
National and Kapodistrian University of Athens, Psachna, Evia, 34400, Greece
ptrakadas@pms.uoa.gr

ABSTRACT

The production of quality printing products requires a highly complex and uncertain process, which leads to the unavoidable generation of printing defects. This common phenomenon has severe impacts on many levels for Offset Printing manufacturers, ranging from a direct economic loss to the environmental impact of wasted resources. Therefore, the accurate estimation of the amount of paper waste expected during each press run, will minimize the paper consumption while promoting environmentally sustainable principles. This work proposed a Machine Leaning (ML) framework for proactively predicting paper waste for each printing order. Based on a historical dataset extracted by an Offset Printing manufacturer, a two-level stacking ensemble learning model combining Support Vector Machine (SVM), Kernel Ridge Regression (KRR) and Extreme Gradient Boosting (XGBoost) as base learners, and Elastic Net as a meta-learner, was trained and evaluated using cross-validation. The evaluation outcomes demonstrated the ability of the proposed framework to accurately estimate the amount of waste expected to be generated for each printing run, by significantly outperforming the rest of the benchmarking models.

CCS CONCEPTS

• **Computing methodologies**; • **Machine Learning**; • **Machine Learning Algorithms**; • **Ensemble methods**;

*Corresponding Author.

ICIEA-EU 2023, January 09–11, 2023, Rome, Italy

ACM ISBN 978-1-4503-9852-7/23/01.
https://doi.org/10.1145/3587889.3588210

KEYWORDS

Machine Learning, Waste Prediction, Offset Printing, Stacking Ensemble Learning

ACM Reference Format:
Alexandros S. Kalafatelis, Chris Trochoutsos, Anastasios E. Giannopoulos, Angelos Angelopoulos, National and Kapodistrian University of Athens, Psachna, Evia, 34400, Greece, a.angelopoulos@uoa.gr, and Panagiotis Trakadas. 2023. A Stacking Ensemble Learning Model for Waste Prediction in Offset Printing. In *2023 The 10th International Conference on Industrial Engineering and Applications (ICIEA-EU 2023), January 09–11, 2023, Rome, Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3587889.3588210

1 INTRODUCTION

The printing industry is one of the largest manufacturing industries in the world, with products ranging from packaging, flyers, books, magazines and newspapers. The sector is a vital part of the European economy, generating an annual turnover of around € 88 billion in EU GDP alone, while employing more than 770 thousand workers [1].

Nevertheless, the industry is currently facing environmental and economic challenges, which may have a negative impact on demand and consequently to the sectors economy. Particularly, the process of producing quality printing products, is highly complex and has extensive environmental impacts, since it requires the extensive use of raw materials (i.e., water, paper, ink and aluminum) and chemicals that are causing major environmental degradation [2].

A widespread lithographic printing technique accommodating many different types of printing jobs, is offset printing. Offset printing utilizes rotating plates to repeatedly transfer ink onto a printing substrate. The technique enables the production of large quantities, as the variable production costs are deemed small compared to the setup costs of the printing plates [3]. However, this method leads to substantial paper waste, especially when the demand for a product requires heterogeneous images on the same plate. In detail, paper is almost always wasted during the press setup or in the final quality control procedures, leaving thus no space for corrective actions [4] [5].

To account for the waste production, companies utilize additional paper for every printing process to successfully produce the requested final product. Specifically, companies employ various industry-standard mathematical formulae, to quantify the overall paper waste. These formulae, calculate the amount of paper that is going to be wasted during the production of a particular job, with respect to the run length (i.e., the requested number of products) and the printing characteristics (e.g., type, colors, etc.) [6]. Nevertheless, these methods have been found to be too pessimistic than necessary, increasing the environmental footprint of the companies, while also not taking into account, variables such as the temperature or the relative humidity that is found in the printing processes and can greatly affect the quality of the final product. Moreover, due to their pessimistic nature, a surplus quantity of paper is used, which in the majority of the cases cannot be reused when returning from production, thus draining a company's both material and financial resources. Therefore, a clear challenge lies in devising an effective method, that accurately informs the manufacturing personnel of the expected printing defects.

Nowadays, factory environments use IoT devices to extensively monitor the production chain, creating massive amounts of data [7]. These data can be utilized to increase the efficiency of the printing processes, by Artificial Intelligence (AI) and Machine Learning (ML) techniques, facilitating the evolution of the industry [8]. Furthermore, the integration of AI/ML techniques in the manufacturing environments, has already proven to play a key role in enabling optimization of processes in terms of automation, waste reduction/prediction and product quality, towards Zero-Defect-Manufacturing (ZDM) [9].

In this paper, we propose a stacking ensemble ML model for waste prediction in Offset Printing, which aims to minimize the unnecessary paper surplus methodology currently utilized by the industry, making thus the printing manufacturing environment more sustainable. In summary, the contributions of this paper include: i) a detailed explanation of existing state-of-the-art waste prediction solutions and their drawbacks, ii) the exploitation of historical knowledge extracted by an Offset printing environment to obtain accurate prediction models and (iii) the proposition of a stacking ensemble ML model, targeting at accurately proactively predicting paper waste before printing jobs.

This paper is structured as follows. In Section 2, we describe the related work of waste prediction and the use of ensemble ML models. In Section 3, we suggest the details of the utilized dataset and of the proposed stacking ensemble model. In Section 4, the experimental results used to assess the performance of both the base and the proposed models are presented. Finally, in Section 5 the results are summarized and discussed.

2 RELATED WORK

Literature investigation reveals that parallel to the experimental practices, mathematical models and different AI/ML approaches have been developed to predict the generation of different types of waste material. For instance, Samarin [6] developed the following mathematical model to estimate paper waste in the sheetfed printing process:

$$n_w = k(N_m + p_p \times n)$$

where n_w is the estimated number of waste sheets, k is the number of job colors, n is the required quantity of printed sheets, N_m it the waste sheet quantity required for makeready, and p_p is the percentage of waste expected during the press run.

In detail, the work of Samarin, is considered as one of the first applications of waste prediction in the domain of Offset Printing. However, a key issue of this proposed model is that the predicted value, is directly proportional to the number of printed colors, leading to a significant overestimation of the waste quantity. Towards the same aim, Hamerliński and Pyryev [6], developed the following mathematical model, aiming to improve accuracy and reduce the quantity of paper required for a particular print run. The proposed model showcased a better accuracy compared to the model proposed by Samarin, however it also leads to overestimation in terms of the waste quantity, while it doesn't take into consideration the potential impact of temperature or humidity in during the press run:

$$\delta = C_1 \times n^A \times N_S^B \times \left(\frac{S_S}{S}\right)^C \times k^E \times \left(\frac{q}{q_S}\right)^H \times \left(\frac{F_T}{S q_S d^2}\right)^I$$

where δ is the waste sheet coefficient (dimensionless) expressed as a ratio of waste sheet quantity to the number of copies printed, C_1, A, B, ... $\mathcal{J}$ are constants, n is number of copies (print run length), N_S is number of pages in job, S_S is page size, k is the number of colors, d is the print speed, q is the paper grammage, S is the sheets size, F_T is the ink tank and q_S is the ink consumption per area.

Subsequently, various methods based on statistical learning theory have been also introduced. The key benefit of ML, is that it promotes low-cost computing through algorithmic learning, without the need of physical-based equations [10]. Based on these advantages, different methods have been developed for different purposes in the field of waste estimation. For example, in [11] the authors used a Gradient Boosting model to accurately forecast the weekly solid waste generation in New York City, while in [12], the authors used Support Vector Machines (SVM) and Random Forest (RF), to predict the municipal solid waste generation in different areas. Furthermore, towards that direction, the authors in [13], trained several ML models, including RF, Decision Trees (DT), SVM and Logistic Regression (LR) to design an intelligent waste management system, that enables the prediction of different types of wastes for smart cities.

Another technique utilized by multiple authors currently, is Ensemble learning, which is defined as a technique of combining several weak base models instead of using a single "powerful" model, in order to make accurate predictions. In [14], the authors proposed a two-level stacking heterogeneous ensemble algorithm, combining RF, SVM and CatBoosting. The final two-level stacking ensemble model showed significant improvements in terms of accuracy against the individual base models, while it also reduced biases.

As today's problems become more and more challenging, different approaches have proved their capabilities in several domains. During the last decades, several attempts have been made to accurately estimate the generation of waste, mainly focusing on SW forecasting, however, to the best of our knowledge, there has been no approach that distills the knowledge coming from the printing industry to proactively predict the wasted resources before a press run using ML.

Table 1: Parameters used to train and test the ML models

Parameter	Description
Unique Order ID (UO_ID)	Unique identifier ranging from 1 to 10000
Quality (Qual)	Paper type requested in a particular order. Quality is a categorical variable that takes values 'Velvet', 'Uncoated' or 'Illustration/Gloss'
Quantity (Quan)	Number of pieces requested in a particular order
Type (T)	The outcome type of a particular order. It is a categorical variable that takes values 'Book', 'Poster' or 'Journal'.
Aluminium Plates (AL_P)	Aluminium plates requirements of a particular order. It is a categorical variable that takes values 'typical' 4-color printing, '4+1' color printing or 'grayscale' printing
Ink Level Required (IR)	The amount of ink required for the completion of the order (gr.)
Offset Paper (OP)	The amount of paper used for the successful completion of a particular order
Machine (Mac)	The ID of the machine that the particular order was forwarded for printing, ranging from 1 to 5
Humidity (H)	Water vapor relative to air temperature
Temperature (Temp)	Air temperature at the factory ranging from 292 to 298 Kelvin
Paper Waste (PW)	The amount of paper wasted for each printing job

Table 2: Parameters and attributes for the input and target variables

Parameter Type	Parameter	Mean	Standard Deviation	Minimum	Maximum
Input variables	Quality	1.600	0.762	1	3
	Quantity	2331.810	1319.671	206	9956
	Type	2.0994	0.8309	1	3
	Aluminium Plates	3.681	0.947	1	5
	Ink Level Required	123.141	69.678	10.896	525.696
	Offset Paper	2565.441	1451.634	227	10952
	Machine	2.676	1.338	1	5
	Humidity	55.007	3.391	45.070	69.940
	Temperature	294.328	1.041	292.020	300.010
Target variable	Paper Waste	463.835	270.093	37	2486

3 MATERIALS AND METHODS

3.1 Dataset Description

The original dataset consisted of features (i.e., order-specific) and labels based on historical measurements from a 4-month period (03/07/2022 - 31/10/2022) coming from Pressious Arvanitidis, an Offset Printing manufacturer based in Greece. Each of the collected parameters and features, follows the process of a particular printing order (i.e., from the sales department to the quality assessment department). The order and factory related characteristics used in this paper are presented in Table 1.

Table 2 summarized the descriptive statistics of the independent and dependent variables of the complete dataset.

3.2 Data Preprocessing

Data preprocessing was performed in order to facilitate the training and testing processes of the ML models with high-quality data. To convert categorical values into numerical values, the one-hot encoding technique was utilized and applied to all categorical features in our original dataset.

Furthermore, to make the data measurements more symmetric to a normal distribution, we used the Log Transformation. This methodology, enables the data transformation to a range (0,1] with the following equation:

$$y' = log_{10}(y) / log_{10}(y_{\max})$$

where y_{max} indicates the maximum value of the label.

3.3 Candidate Models

To create the proposed stacking ensemble learning framework, five base ML models were trained, using the scikit learn package, including: i) Elastic Net (ENet), which is a penalized linear regression model, ii) Support Vector Machine (SVM), which uses an ϵ-insensitive tube to allow for more flexibility in errors, iii) Kernel Ridge Regression (KRR), which is identical to SVM, with the exception that it doesn't ignore errors smaller than ϵ, iv) Extreme Gradient Boosting (XGBoost), which provides parallel tree boosting and v) LightGBM, which is a gradient boosting framework that uses tree-based learning algorithms.

To evaluate the performance of the base models and find the suitable hyperparameters, the dataset was randomly divided into two subsets, a training dataset (90%) and a testing dataset (10%). Table 3 shows the ML models and their hyperparameter that provided the best cross-validation scores.

Table 3: Base ML models and their hyperparameter settings

ML Models	Hyperparameters	Values
ENet	alpha	0.00001
	l1_ratio	1
	max_iter	50000
	random_state	1
KRR	alpha	0.6
	kernel	polynomial
	degree	2
	coef0	2.5
SVM	kernel	poly
	coef0	0.5
XGBoost	colsample_bytree	0.65
	gamma	0.01
	learning_rate	0.05
	max_depth	3
	n_estimators	200
	reg_alpha	0.464
	reg_lambda	0.857
	subsample	0.55
	nthread	-1
LightGBM	objective	regression
	num_leaves	2
	learning_rate	0.05
	n_estimators	1000
	bagging_fraction	0.76
	bagging_freq	5
	feature_fraction	0.2319
	feature_fraction_seed	9
	bagging_seed	9
	min_data_in_leaf	9
	min_sum_hessian_in_leaf	11

3.4 Proposed Stacking Ensemble Learning Framework

To implement the proposed stacking, the following four steps were used:

Step 1: During the first step, a five-fold cross-validation technique was used to evaluate the performance of the ML models. As a result, SVM, KRR and XGB were selected as base learners.

Step 2: After the selection of the base models, the out-of-fold predictions from the training set were used as a new feature together with the average value of predictions coming from the testing set. The new features were then merged with the existing datasets.

Step 3: During the third step, a meta-learner was selected (ENet), which was trained and evaluated using the new generated datasets.

Step 4: Finally, the stacked meta-learner regressor is combined with the XGB model to accurately predict the paper waste generation for each printing run.

The proposed architecture of the stacking ensemble learning framework is showcased in Figure 1.

Table 4: Overall comparison of the proposed candidate models in terms of RMSLE

Models	RMSLE
ENet	0.1244
KRR	0.1234
SVM	0.1237
XGB	0.1249
LightGBM	0.1317
Proposed Model	0.1180

3.5 Evaluation

We used 5-fold cross-validation to compare the performance of the candidate regression models by using the Root Mean Squared Logarithmic Error (RMSLE), calculated as [15]:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1) - \log(a_i+1))^2}$$

where the natural logarithms are considered. In this equation, p_i and a_i are the predicted and actual values for data instance i, respectively.

4 RESULTS

Figure 2 shows the comparison of the frequency histograms and the quantile–quantile (Q–Q) plots of the PW and the log-transformed PW data, indicating that the log transformation tended to make the distribution more symmetric and normal.

While table 4 shows the overall cross-validation results (RMSLE) calculated by the candidate models and the proposed framework. The results demonstrate that the values of RMSE are almost constant for the ENet, KRR, SVM and XGB models, while the LightGBM model showed to be affected greatly by the larger penalty for underestimation of the RMSLE evaluation. Furthermore, the proposed stacking ensemble learning model, outperformed the base models, indicating that the framework is superior to single ML models.

5 CONCLUSSIONS

Printing environments are characterized by high uncertainty, and therefore, defects are an unavoidable and common phenomenon. These defects can have a severe impact on many levels, ranging from a direct economic loss to the environmental impact of wasted resources.

In this study, we proposed a stacking ensemble learning framework to improve predictions of paper waste in Offset Printing environments, to minimize the unnecessary paper surplus methodology currently utilized by the industry and make the manufacturing processes more sustainable. The two-layer stacking ensemble framework, consisting of the SVM, KRR, and XGB models as the first layer and an Elastic Net model as the second layer, was generated based on a dataset from an Offset Printing company. Overall, the novel stacking ensemble learning model outperformed the base models in terms of the RMSLE on the validation set, confirming it can provide accurate estimations against other state-of-the-art models.

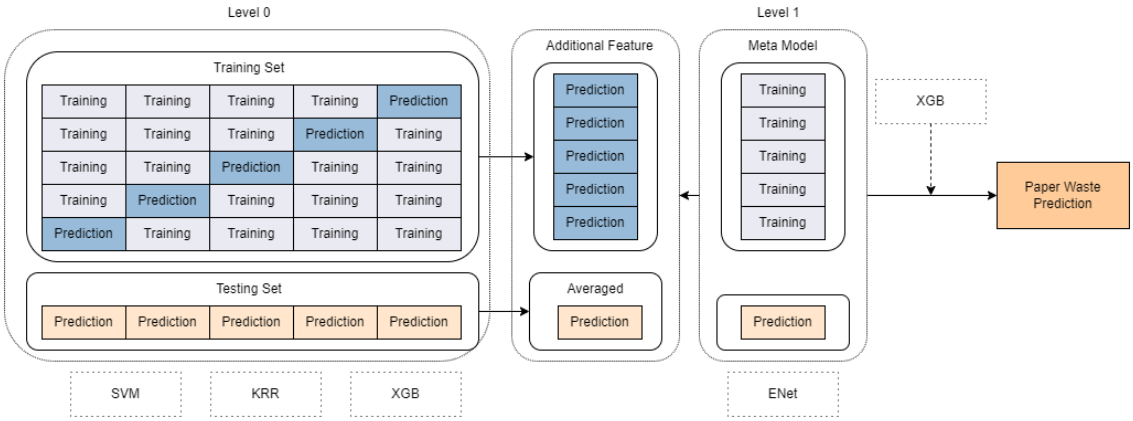

Figure 1: Proposed stacking model architecture.

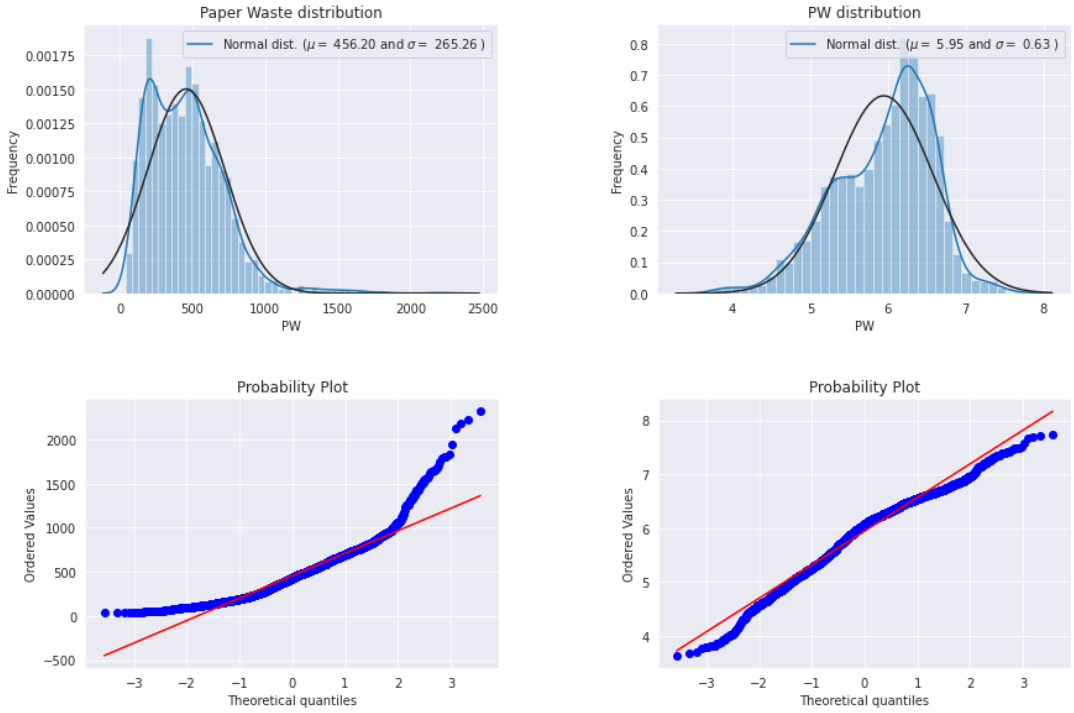

Figure 2: Frequency histograms and Q–Q plots of the original PW data and the log-transformed PW data.

Regarding the practical implications of the findings of this study, the adoption of the proposed framework is envisioned to facilitate the evolution of the industrial premises of Offset Printing manufacturers, assisting them to adopt Industry 4.0 principles, by moving their production chains from the state of ex post management to the state of ex ante prediction of resource management. Furthermore, the accurate estimations of paper waste, will also have a positive environmental impact on printing manufacturers, enabling efficient production and significant reduction of operating expenses.

Finally, unlike other previous studies that used only homogeneous ensembles or simple weighted average models, to the best of our knowledge, ours is the first study to explore the estimation of waste generation using ML in Offset Printing. In future work,

the proposed methodology can be further extended to predict additional raw materials, such as ink, water, and aluminum that are also consumed during each print run.

ACKNOWLEDGMENTS

This work has been partially supported by the PDS project, under the open call of the AI REGIO (Regions and Digital Innovation Hubs alliance for AI-driven digital transformation of European Manufacturing SMEs) project, funded by the European Commission under Grant Agreement number 952003 through the Horizon 2020 program (https://www.airegio-project.eu/) and by the ICOS (Towards a functional continuum operating system) project, funded by the European Commission under Grant Agreement number 101070177 through the Horizon 2020 program (https://www.icos-project.eu/).

REFERENCES

[1] European Commission,Printing Industry: Why the printing industry is important. Retrieved November 17, 2022 from https://single-market-economy.ec.europa.eu/sectors/raw-materials/related-industries/forest-based-industries/printing-industry_en.

[2] Richard Smith. 2011. The Environmental Sustainability of Paper. Graduate Studies Journal of Organizational Dynamics, 1(1).

[3] Philipp Baumann, Manuel Kammermann, and Silvan Elsaesser. Minimizing paper waste and setup costs in offset printing. 2021. In Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 858-862. 10.1109/IEEM50564.2021.9673049.

[4] Sotirios T. Spantideas, Anastasios E. Giannopoulos, Nikolaos C. Kapsalis, Angelos Angelopoulos, Stamatis Voliotis and Panagiotis Trakadas. 2022. Towards Zero-Defect Manufacturing: Machine Selection through Unsupervised Learning in the Printing Industry. In Proceedings of the Workshop of I-ESA. Valencia, SP.

[5] Angelos Angelopoulos, Anastasios E. Giannopoulos, Nikolaos C. Kapsalis, Sotirios T. Spantideas, Lambros Sarakis, Stamatis Voliotis, and Panagiotis Trakadas. 2021. Impact of Classifiers to Drift Detection Method: A Comparison." In International Conference on Engineering Applications of Neural Networks, 399-410. https://doi.org/10.1007/978-3-030-80568-5_33.

[6] Jacek Hamerliński and Yuriy Pyr'yev. 2014. A Method of Minimising Paper Requirements for Offset Printing. BioResources 9(3), 5147-5154.

[7] Alexandros Kalafatelis, Konstantinos Panagos, Anastasios E. Giannopoulos, Sotirios T. Spantideas, Nikolaos C. Kapsalis, Marios Touloupou, Evgenia Kapassa, Leonidas Katelaris, Panagiotis Christodoulou, Klitos Christodoulou and Panagiotis Trakadas. 2021. ISLAND: An Interlinked Semantically-Enriched Blockchain Data Framework. In Proceedigs of the International Conference on the Economics of Grids, Clouds, Systems, and Services, 207-214. https://doi.org/10.1007/978-3-030-92916-9_19.

[8] Angelos Angelopoulos, Anastasios Giannopoulos, Sotirios Spantideas, Nikolaos Kapsalis, Chris Trochoutsos, Stamatis Voliotis, and Panagiotis Trakadas. 2022. Allocating orders to printing machines for defect minimization: A comparative machine learning approach. In IFIP International Conference on Artificial Intelligence Applications and Innovations, 79-88. https://doi.org/10.1007/978-3-031-08337-2_7.

[9] Angelos Angelopoulos, Emmanouel T. Michailidis, Nikolaos Nomikos, Panagiotis Trakadas, Antonis Hatziefremidis, Stamatis Voliotis, and Theodore Zahariadis. 2019. Tackling faults in the industry 4.0 era—a survey of machine-learning solutions and key aspects. Sensors 20(1), 109. https://doi.org/10.3390/s20010109.

[10] Sara Nasiri and Mohammad Reza Khosravani. 2021. Machine learning in predicting mechanical behavior of additively manufactured parts. Journal of materials research and technology 14 (2021), 1137-1153. https://doi.org/10.1016/j.jmrt.2021.07.004.

[11] Nicholas E. Johnson, Olga Ianiuk, Daniel Cazap, Linglan Liu, Daniel Starobin, Gregory Dobler, and Masoud Ghandehari. 2017. Patterns of waste generation: A gradient boosting model for short-term waste prediction in New York City. Waste management 62, 3-11. https://doi.org/10.1016/j.wasman.2017.01.037.

[12] Atul Kumar, S. R. Samadder, Nitin Kumar, and Chandrakant Singh. 2018. Estimation of the generation rate of different types of plastic wastes and possible revenue recovery from informal recycling. Waste Management 79, 781-790. https://doi.org/10.1016/j.wasman.2018.08.045.

[13] G. Uganya, D. Rajalakshmi, Yuvaraja Teekaraman, Ramya Kuppusamy, and Arun Radhakrishnan. 2022. A Novel Strategy for Waste Prediction Using Machine Learning Algorithm with IoT Based Intelligent Waste Management System. Wireless Communications and Mobile Computing 2022. https://doi.org/10.1155/2022/2063372.

[14] Uyeol Park, Yunho Kang, Haneul Lee, and Seokheon Yun.2022. A Stacking Heterogeneous Ensemble Learning Method for the Prediction of Building Construction Project Costs. Applied Sciences 12(19), 9729. https://doi.org/10.3390/app12199729.

[15] Jiachen Zhang, Xingquan Zuo, Mingying Xu, Jing Han and Baisheng Zhang. 2021. Base Station Network Traffic Prediction Approach Based on LMA -DeepAR. In 2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS). 10.1109/ICCCS52626.2021.9449212.