Grenze International Journal of Engineering and Technology, June Issue

Evaluation of Machine Learning Techniques for High Dimension Dataset for Predicting Municipal Solid Waste Generation

G Jenilasree[1], Dr.Sujatha[2] and Dr.M.Bhuvaneswari[3]
[1]Research scholar, University college of Engineering, Anna University (BIT campus), Trichy
Email: jenilasree@gmail.com
[2]Professor, University college of Engineering, Anna University (BIT campus), Trichy
Email: sujathaaut@gmail.com
[3]Assistant Professor, University college of Engineering, Anna University (BIT campus), Trichy
Email: bhuvanakirran@gmail.com

***Abstract*—The forecast of waste generation is an essential step for an adequate planning of waste management, since it has various factors that can be used effectively. To make a decision process, the application of predictive and prognosis models are used as tools. In this paper ,the raw dataset is obtained from the Tiruchirappalli municipal corporation as case study.The real time data set has indicators such as : population , number of residents , bio degradable waste ,non- biodegradable waste, hazardous waste and other total municipal solid waste are applied as input variables into the models to predict the amount of solid waste fractions obtained from various cities. The real time raw data set is preprocessed and applied into Support vector machine(SVM).The SVM model shows with an accuracy of 85% that is analyzed for the forecasting of solid waste generation and composition.**

***Index Terms*— Forecasting, Municipal Solid waste, Data Preprocessing , Correlation and supervised machine learnings.**

I. INTRODUCTION

The drastic change in the lifestyle, speedy growth of consumerism triggered by aggressive competition among producers of consumer products and revolution changes in the packaging industry had all resulted in a shocking increase in the composition of the waste .An emerging problem for local and national governments is waste management due to the growing amount of solid waste [1,2].Considering the reality , the almost 50% of Indian population will be living in urban areas by 2030, and transforming consumer behaviors have resulted in the production of 62 million tons of garbage every year, of which less than around 15% is processed and 60% is collected. Recently, the Central Pollution Control Board, Delhi, India confirmed that it requires more than 650 acres of landfills for this process which is four times of the existing space in India, at least 9,000 metric tons of garbage remains uncollected on streets sides and bins. These affecting the environment and the hygiene of the people. The ministry of urban Development (MOUD) came out with the smart programme in selected municipalities and corporations. The focus is on sustainable and inclusive development at compact areas to create a replicable model which will act as a lighthouse for development of all the cities [3]. The core

Grenze ID: 01.GIJET.9.2.368

Infrastructure elements of urban areas has major issue of solid waste management, a series of guidelines for developing a SMART city plan to over 500 cities in the country. This was evaluated and based on several parameters laid down, Trichy city was selected to be a SMART city in Tamil Nadu .so, and the real time garbage dataset is collected from Trichy Corporation for further analysis and prediction process in order to know about the exact production obtained from various garbage's. In this paper, the real time garbage data set preprocessed to avoid the classification error rate in the given garbage data set. The generated output is then correlated. The correlation is a method used to check for data preprocessed data to check whether the classification is finished correctly or not[4]. After that, the feature extraction process is carried out and it is used to extract knowledge from a large amount of data to analyze and predict the given data set. It also uses data mining techniques and other techniques which are used to predict future outcomes to build models for the algorithm behind the data set. As algorithms can extract data automatically, machine learning algorithm and various regression plays an important role in the field of data science, data mining and Artificial intelligence. Feature Extraction is majorly divided into two categories namely unsupervised learning and supervised learning algorithms. Supervised learning is a technique in which it trains the machine using the data which is already ragged with the correct answer. Then, a new set of data is given to the algorithm for analysis to obtain accurate outcome [5]. In this paper a real time garbage data set for the year 2020,2021 and 2022.Further the feature is extracted using supervised machine learning algorithms .The remaining of this paper is organized as follows session 1 describes about Introduction, session 2 describes about Related Work, session 3 Municipal solid waste management in Tiruchirappalli ,session 4 describes about dataset description ,session 5 describes about Data preprocessing, session 6 describes about Data classification task, Session 7 describes about Performance Analysis using various Supervised Machine Learning Algorithm and Regressions, session 8 describes about Result and conclusion and session 8 describes about Future work.

II. Related Work

B. Li and P. D. Franzon, (2019) [6] to predict the value in physical design it uses machine learning. In this approach by using Global route with Machine learning the accuracy and prediction rate is approximately high.

M Khan (2018) proposed work is to find the patterns of learning and interaction of effective complex system using machine learning algorithms and this could be obtained by collaborative learning environment (CLE). The new frame work identifies the changes in the behavior at micro level itself.

Narges Shahidi, Mahmut T. Kandemir (2018)[7] In this paper they proposed a new Cached GC garbage collection which postpones valid pages during writing the data in cloud and erases the possible time by performing Solid State Drive. The request response time gets reduced during this process. Due to this 19% of recite tail latency is reduced and the same improves 27% of IOPS

B, D.Z. Pan, T., and X., (2015) [8]the proposed method is used to reduce gigantic data and complexity by using pattern matching and machine learning techniques. The adaption and applying process of ML and PM in VLSI physical design leads to provide prediction and matching which shows quality results.

Y. Li, P. P. Lee, J. C., and Y. (2015)[9] the proposed probabilistic method is to detect the exact location of garbage collector. The result shows data locality of garbage collected and improves the performance of data locality. By using trace driven method, we can find out the model of the garbage bin exactly in real time world.

The prediction of waste generation during different seasons were applied into artificial neural network, over short, medium and long time stages

III. Municipal Solid Waste Management in Tiruchirappalli ,Tamil Nadu

Tiruchirappalli is a city from Tamil Nadu, a region of India. The expectancy of urban life in the city has been increased day by day. The increasing amount of solid waste happens due to the higher consumption of population .Tiruchirappalli is the fourth largest city in Tamil Nadu and occupies an area of 4,407 square kilometer and the population is 2,388,831. One of the major challenges in Tamil Nadu is the collection and disposal of the mixed garbage waste via an Eco friendly manner and recover from the waste as much wealth as possible. Further the quantity of waste generated has gone up to about 500 tons per day and it has a risky task to collect this amount of garbage .A very pungent foul smell emanates from the decayed garbage during rainy season much to the discomfit of the residents .During summer, fire breaks out in the dumping yard as well as residential area and fanned by the sunlight and heavy winds spread fast among the scores of mound some of which rise up to 50 feet. During past four years, this has become a regular feature, forcing scores of

families to lock up their houses. On several occasions the pungent smoke emanating from the burning waste results in poor visibility on a stretch of National Highway and traffic have been stopped for periods ranging from 2 to 6 hours. In several parts of the country bio-mining method has been employed to convert such dumping yards to remove the bulk of the accumulated waste , use whatever wealth available in the dry garbage and also remove non bio degradable waste as scrap which could be sent to nearby cement factories as already mandated by solid waste management rules 2000, amended in 2016 and 2017.All the segregating the different types of waste are now in the final stage but the amount of various solid waste collected is not analyzed or predicted clearly .So most of the wastages from the residential areas are not collected properly due to unawareness of garbage produced in the particular area.
Figure 1 represents steps involved in waste management.

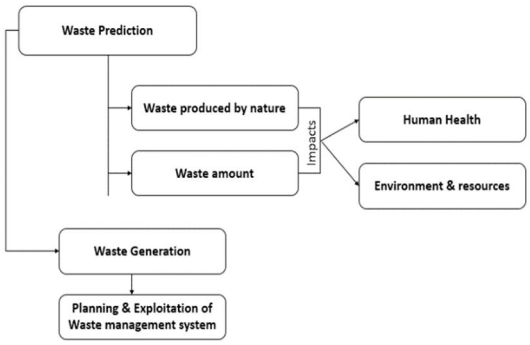

Figure 1. Steps involved in waste management

IV. DATA SET DESCRIPTION

The experimental study uses Garbage collection data set to assess the model performance. A large data set of garbage collection for each and every day was taken for the year 2020, 2021 and 2022 for analysis and prediction. The data set is obtained from Trichy Municipal Corporation consists of wet waste , bio degradable , non biodegradable , hazardous and other wastages. At the beginning, the data set is partially divided into train and test set and it is pro portioned as 60:40. The target is set and if the range is below the target it indicates low, average indicates medium and if the range is above the target then it is high.

V. DATA PREPROCESSING

Before going to Machine Learning, Data preprocessing is an initial step for real-world dataset which is incomplete, inaccurate and often lacks specific values. It is mainly helps to clean and organize the raw dataset to make it more effective for future processing steps like Machine Learning models. Data cleaning includes imputation , noisy data , smoothing and clustering are removed .In imputation it is calculated either by mean or median .Then the data transformation is done by using normalization method that converts the value into 0's or 1's (ie) standard form . The min max method is used for normalizing data in early phase. All the data used in the training have the same scale and the training time is improved by normalization shown in equation 1.

$$\text{Xnorm} = \text{X} - \text{Xmin Xmax} - \text{Xmin} \qquad (1)$$

Xnorm is said to be as normalization and before being normalized it is said to be as X. Both Xmax and Xmin describes about the maximum and minimum values of each feature. It is then carried out by data integration ,where we check correlation for the given dataset. Correlation is a statistical measure which describes the association between random variables and helps in predicting one quantity from another. It is mainly used when both variables taken for the study are normally distributed. This affects the extreme value of co-efficient, which may diminish the strength of relationship. The variables of the data set are independent. The below equation of correlation is represented as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - y)}{\sqrt{\sum_{i=1}^{n}(x^{(i)} - x^{-1})^2]}\sqrt{[\sum_{i=1}^{n}(y^{(i)} - y^{-1})^2]}} \qquad (2)$$

Where r is said to be sample statistics. The parameter x and y indicates the number of independent variables. In the garbage data set figure 4 shows the cross validation for the independent variable. The output of LFS is

applied to the correlation and the result show s that the relationship for the given garbage data set is classified correctly shown in figure 2.

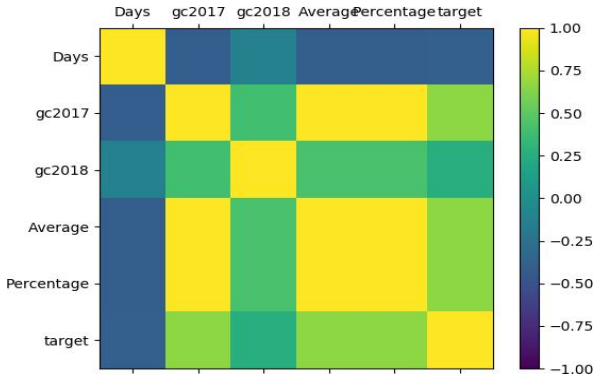

Figure 2. Correlation Relationship

VI. DATA CLASSIFICATION TASK

The class membership decides to classify the data $y^{'}$ with unknown data based on data set $D=(x_{1,}y_{1),...}$ $(x_{n,}\ y_{n})$ of data items x_i with known memberships yi. In this case we mainly focused on classification problems, in which the class labels are either 0 or 1. The m- dimensional vectors are x_i and the components that includes are input variables or independent variables and co- variants. The data classification method has two approaches namely : Supervised Learning , Regressions.. In Data mining it is concerned with finding understandable knowledge, while machine learning is concerned with improving the performance of an intelligent system or agent for problem-solving tasks. Machine learning algorithm consists of supervised and unsupervised learning algorithm helps to analyze and predict the dataset [5]. Figure 3 represents the flow graph of machine learning algorithms.

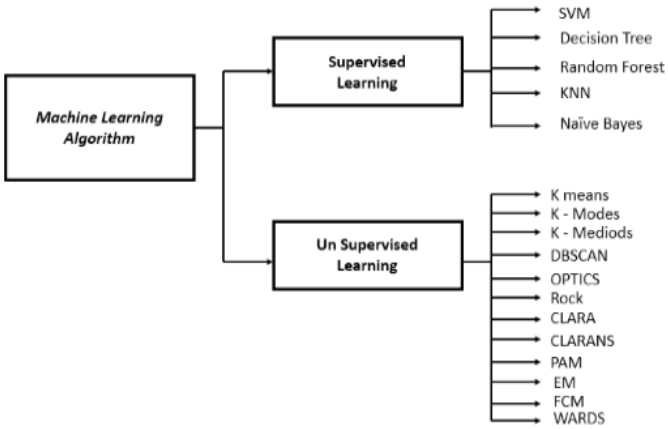

Figure 3 Machine Learning Algorithms

A. Supervised Algorithm

Supervised learning means learning from examples. A given training set acts as examples for the classes. The system finds the description of each class. Once this has been formulated, it is used to predict the class of the previously unseen objects. This is similar to discriminant analysis in statistics. The classification process is done easily.

B. Un supervised Algorithm

Unsupervised learning means observation and discovery. In this mode of learning, there is no training set or proper knowledge of the classes. The system analyses the given set of class descriptions and produces different clusters of data. This helps in finding the unlabeled data's hidden structure. This is similar to cluster analysis in statistics.

VII. Performance Analysis using Supervised Algorithm

reason for choosing Supervised Algorithm to pursue the study is that it produces accurate analysis and predicts processes effectively for the given data set because the algorithms involves in classification process. It will allow us to determine the unseen instances with class labels that help in generalizing the data in a reasonable way.

A. Support Vector Machine

Support Sector Machine (SVM) is an advanced supervised learning technique for regression and classification which automatically fits nonlinear and linear models. This algorithm has a special advantage of minimizing the expected error compared to minimize the classification error. It also takes a labelled data as input and creates learning functions that can be used for classification of unknown test data. SVM can perform as a linear and non-linear classification. In contrast to the decision theoretic minimum distance classifier that uses only one decision boundary , SVM uses a reference hyperplane and two decision boundaries to classify the points. In a training dataset all training samples that fall on the boundary lines are called support vectors. A SVM implements a binary classifier. This means that there are two classes say +1 and -1 .

$$D = \{(x_1, y_1), (x_2, y_2), \ldots \ldots (x_n, y_n)\}, x \in \Re^n, \text{and y is the target label.}$$

$$y = \{-1, 1\} \qquad (3)$$

With the help this to find a hyper plane that maximally separates the classes. The hyper plane equation.

$$h(x) = b + w_1x_1 + w_2x_2 + \cdots + w_nx_n = 0 \qquad (4)$$

where b is the intercept and $a_1, a_2, \ldots \ldots, a_n$ are coefficients and n is the dimension of the data point. Then the hyper plane equation separates the data points x_i into two classes +1 and -1.Here,w weight vector and b is the bias or offset of the origin. In real time dataset there is a need to extract features from these data prior for classification and we use Linear kernel ,Polynomial kernel. Linear kernals equation

$$k(x, y) = \emptyset(x)\emptyset(y) = x^Ty \qquad (4)$$

Where x and y are two vectors.

To get a dual problem that admits efficient computational methods, it goes for the duality theory of mathematical program. Support Vector Machine in which SVM Linear shows 100% precision of the given data set compared to SVM non Linear Precision that contains different kernels that shows precision97.34%(polynomial),97.20% and 95.47%Linear . The result shows 83% of accuracy and 79% precision.Table 1 represents show various of kernals of RBF , Linear and Polynomial.

TABLE I. SHOWS VARIOUS KERNALS

Kernals		
RBF	Linear	Polynomial
97.20%	95.47%	97.34%

VIII. Results & Conclusion

The data set of collected garbage is preprocessed and correlated. The data set is analyzed and predicted by using supervised learning algorithms accuracy of 85 percentage and precision of 93%. Figure 4 represents analysis of various garbage collected per day.With the help of this process we can calculate the exact collection of garbage for the upcoming year. As predicted, for the year 2021, per capital, it would produce 0.569 tons of garbage and the total garbage will be around 71.15 ton s. For the year 2023, per capital, it will be around 0.649 tons of garbage and over all garbage collected throughout India will be around 107.01 tons.

Figure 5 represents various comparison of garbage collected per month and figure 6 represents garbage's collected for 2020 and 2021.

IX. Future Work

In India ,the garbage segregated at source in the country will be around 43 per cent. From the mid of 2018, the collection of garbage is done door- to- door and this has increased the collection of garbage rapidly from 53 per cent to 85 percent. However mixed garbage ends up in the dumpsite and collection of segregated garbage has started. The future work is to get image data set containing the different types of garbage namely degradable,

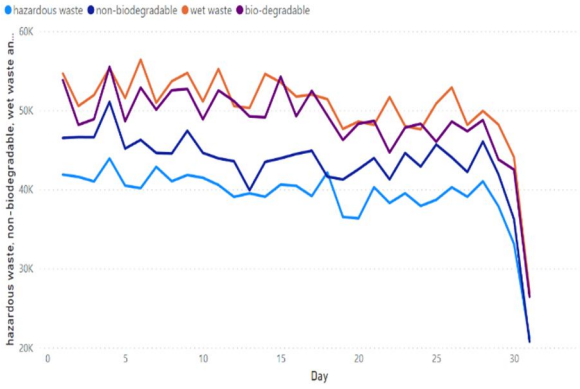

Figure 4 Analysis of various garbage's collected in a day

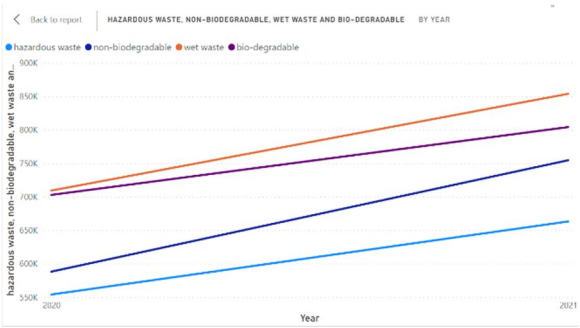

Figure 5 Comparsion of various garbages collected for month

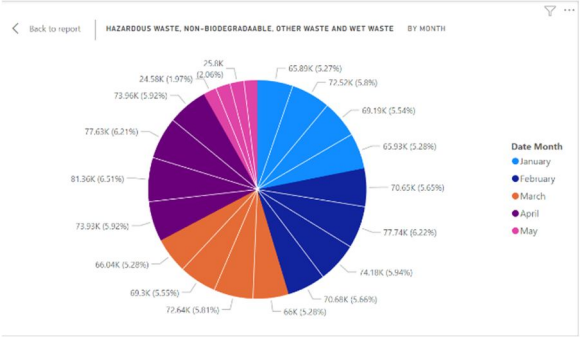

Figure 6 Comparison of various garbage's collected for 2020 and 2021

non-degradable and hazardous waste ,paper ,trash,plastic etc. and by using ensemble learning and deep learning algorithm to determine which type of feature is collected rapidly and to design a smart bin.. Due to this the environment gets cleaned, production of cobalt gas will be increased and level of landfill will be reduce.

ACKNOWLEDGEMENTS

This work is supported by Municipal corporation of Tiruchirappalli ,Tamil Nadu for giving garbage dataset for my research work and I sincerely thanks to Dr. N.Shenbagavadivu for guiding me.

REFERENCE

[1] https://www.tn.gov.in/ta/district_view Berecet,M.,Fischer,K .,2015,Ecological Evaluation of an optimized waste collection system – case study in a german community,Environment.Eng.Manag.J.12,1361-1369.

[2] Costiuc, L., Tierean, M., Baltes, L., Patachia, S., 2015. Experimental investigation on the heat of combustion for solid plastic waste mixtures. Environ. Eng. Manag. J.14, 1295-1302

[3] Lindawati C, Wang W, Cui N, Hari. Feasibility Analysis on Collaborative Platform for Delivery Fulfillment in Smart City. In: IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), 19-21 Dec 9 2015, Chengdu, China. IEEE ; 2015,. p. 147–152. Available from: https://ieeexplore.ieee.org/document/7463716.
[4] Anagnostopoulos T, Zaslavsy A, Medvedev A, Khoruzhnicov S. Top - k Query Based Dynamic Scheduling for IoT enabled Smart City Waste Collection. In: 16th IEEE International Conference on Mobile Data Management, 15-18 , June 2015, Pittsburgh, PA, USA. IEEE ; 2015,. p. 50–55. Available from: https://ieeexplore.ieee.org/document/7264372.
[5] Hahn SS, Lee S, Kim J. To collect or not to collect: Just-intime garbage collection for high-performance ssds with 5 long lifetimes. In: 52nd ACM/EDAC/IEEE Design Automation Conference (DAC), 8-12 June 2015, San Francisco, CA, USA. vol. 191. IEEE ; 2015,. p. 1–191. Available from: https://ieeexplore.ieee.org/document/7167376.
[6] Mcewan MZ, Komsul. Pre-emptive garbage collection for ssd raid. In: and others, editor. 2016 Euromicro Conference on Digital System Design (DSD), 31 Aug.-2 Sept. 2016, Limassol, Cyprus. IEEE ; 2016,. p. 356–363. Available from: https://ieeexplore.ieee.org/document/7723574.
[7] Maglogiannis I, Loukis E, Zafiropoulos E, Stasis. Support Vectors Machine-based identification of heart valve diseases using heart sounds. Journal Computer Methods and Programs in Bio medicine. 2009; 95: 47– Available from: 10.1016;https://dl.acm.org/citation.cfm?id=1534969.
[8] Li B, Franzon PD. Machine learning in physical design. In: IEEE 25th Conference on Electrical Performance Of Electronic Packaging And Systems (EPEPS), 23-26 Oct. 2016, San Diego, CA, USA. IEEE ; 2016,. p. 147–150 , Available from: https://ieeexplore.ieee.org/document/7835438.
[9] Nargis ,Shahidi N, Kandemir TM. CachedGC: Cache-Assisted Garbage Collection in Modern Solid State Drives. In: and others, editor. 2018 IEEE 26th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), 25-28 Sept. 2018. Milwaukee, WI, USAMilwaukee, WI, USA: IEEE ; 2018,. Available
[10] B Yu, JR Gao, D Ding, X Zeng, DZ Pan, Accurate lithography hotspot detection based on principal component analysis-support vector machine classifier with hierarchical data clustering,Journal of Micro/Nanolithography, MEMS, and MOEMS 14 (1), 011003
[11] Li Y, Lee PP, Lui JC, Xu Y. Impact of data locality on garbage collection in ssds: A general analytical study. In: Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering, ser. ICPE 15. ACM ;2015,. p. 305–315. Available from: https://dl.acm.org/citation.cfm?id=2688036.
[12] Chopade P, Saad M, Khan D, Davier AV. Machine Learning for Efficient Assessment and Prediction of Human Performance in Collaborative Learning Environments. In: 2018 IEEE International Symposium on Technologies for Homeland Security (HST), 23-24 Oct. 2018, Woburn, MA, USA. USA: IEEE ; 2018,. Available from: https://ieeexplore.ieee.org/document/8574203.
[13] Obayya M, Abou-Chadi F. Data fusion for heart diseases classification using multi-layer feed forward neural network. In: and others, editor. International Conference on Computer Engineering & Systems, 25-27 Nov. 2008, Cairo, Egypt. IEEE ; 2008,. p. 67–70. Available from: https://ieeexplore.ieee.org/document/4772968.
[14] Ahamed F, Farid F. Applying Internet of Things and Machine-Learning for Personalized Healthcare. In: and 24 others, editor. 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), 3-7 Dec. 25 2018, Sydney, Australia. IEEE ; 2018,. iCMLDE. Available from: https://ieeexplore.ieee.org/document/8613997. from: https://ieeexplore.ieee.org/document/8526873.
[15] Yu B, Pan DZ, Matsunawa T, Zeng X. Machine learning and pattern matching in physical design. In: The 20th Asia and South Pacific Design Automation Conference, 19-22 Jan. 2015, China, Japan. IEEE ; 2015,. p. 19–22,Available from: https://ieeexplore.ieee.org/document/7059020.
[16] Puiu D, Barnaghi PM, Toenjes R, Kuemper D, Ali MI et al. CityPulse: Large ScaleData Analytics Framework for Smart Cities. IEEE Access. 2016; 4: 1086–1108. Available from: 10.1109/ACCESS.2016.2541999;https://ieeexplore.ieee.org/document/7447851.
[17] Wang B, David-Chalon R. A Smart City Case Study: Dynamic Management of Road Lanes. In: Distributed ,Ambient, and Pervasive Interactions. Proceedings of the Second International Conference. Springer ; 2014,. p. 629– 640. Available from: 10.1007/978-3-319-07788-8_58;https://link.springer.com/chapter/10.1007/978-3-319-07788-8_58.
[18] Shahrestani S. Assistive 43 IoT: Deployment Scenarios and Challenges. In: Internet of Things and Smart Environments Springer, Cham ; 2017,. p. 75–95. Available from: https://doi.org/10.1007/978-3-319-60164-9_5.