Paper No: SE-04 **Systems Engineering**

Initiating customer relationship measurement at small and medium enterprises with low computational cost

Tharika Chalani Weerakoon*
Department of Physical Sciences and Technology
Sabaragamuwa University of Sri Lanka, Sri Lanka
thariweera@gmail.com

Kapila Tharanga Rathnayaka
Department of Physical Sciences and Technology
Sabaragamuwa University of Sri Lanka, Sri Lanka
kapila.tr@gmail.com

Abstract: **Modern business entities re-engineer all core business and secondary business activities for customer satisfaction, thereby boosting profit margins. In re-engineering efforts, businesses require immense data processing and decision-making on customers and buying patterns. Still, SMEs face challenges when moving on to customer-centric marketing due to the less accumulated data and resistance towards investing in erudite decision-making tools. Hence, this current research study aims to provide a feasible data mining approach for SMEs in customer recognition with low computation complexity. Data accumulation has happened during the introduction of SME's mobile application to its customers. Hence, the dataset consists of demographic features age, gender, residency region and occupation of each customer. The proposed approach has two phases as follows; the first phase, the customer demographic data with the target variable of purchase value had subjected to data preprocessing. Null values and noise have treated with a binning method. In the second phase, feature engineering had carried out so categorical variables are in numerically manner. Therefore, the binary encoding was used for categorical. Finally, the dimensionality reduction of the processed data had done using Principal Component Analysis (PCA) to extract the most prominent and customer explanatory attributes within the SME. The PCA yielded 10% of a reduction in total explained variance percentage, meaning that the data had compressed. Using K-Means clustering 4 distinctive clusters were extracted. The usage of PCA had leveraged quick clustering and obtained 4 clusters represented the most impactful customers between the age of 0-17 and occupational level 10. With the implementation of PCA, the dataset narrowed down only with the most prominent features that an SME should care of and with this methodology SME can initiate the practising of more efficient customer data analysis using data mining and machine learning.**

Keywords: *Customer Relationship Management, Data mining, Principal Component Analysis, Small and Medium Enterprises*

I. INTRODUCTION

Small and Medium Enterprise (SME) sector is an important contributor to many economies and has received special attention from government policy formulation bodies worldwide. More than 75% of enterprises are SMEs and according to World Bank reports, a quota of up to 60% of total employment and 40% of GDP is backed by SMEs in emerging economies like Brazil, India, China and South Africa. Furthermore, the World Bank emphasizes that the growth of SMEs substantially improves local economies of a particular country.

Technological advancements have influenced SMEs to engage in more competitive and dynamic business environments. Mainly, SMEs have focused on customer-oriented merchandising rather than traditional merchandising of goods and services. Hence, SMEs adopt mobile-based applications with online buying and selling approaches, accumulating data in databases. This accumulated data carry hidden information which can be used for effective and insightful decision making. To extract particularly hidden information, data mining concepts are being used.

Data mining yields productive outcomes for business acumen and SMEs and can be deployed to evaluate companies' own customer base. Usually, SMEs show resistance towards practising data mining at their enterprises due to lack of experience in data mining techniques and complexity in formulating the best data mining strategy at the enterprise level. Further, SMEs have the major concern of having small data sets compared to large scale organizations. Therefore, they are reluctant to invest since they perceive that data mining will not be effective in generating returns for their investment.

The study primarily aims to introduce an approach to initiating data mining at SMEs with respect to Customer Relationship Management (CRM).

II. LITERATURE REVIEW

A. Data Mining

Data mining is the excavating of useful and relevant information from considerably large data sets or databases. Data mining fundamentally objectify to provide more insight and understanding about the data structure and the important features of the dataset. The two goals of data mining are information description and prediction. SMEs can gain immense advantage within the industry by integrating data mining and knowledge discovery in accumulated data [1].

The process of data mining is a combination of an intensive set of steps from data gathering to knowledge extraction. Data will be gathered from different sources, primarily from data warehouses and secondary data from third-party data repositories. Then feature selection will be used to identify best data attributes which facilitate desired data-driven objectives set in the business organization. The target data will be formulated and pre-processing of data will commence since many inconsistencies may exist within the data. The pre-processed data will be transformed into tensors or datagrams, which are the forms that will leverage effective data mining. At the data mining stage, the data will undergo pattern recognition, regression analysis and different statistical analysis along with machine learning. At the final stage of extracted information, the knowledge will be accumulated and will be used for the organization's decision-making process.

Data mining leverages enterprises to escalate business and financing competencies. The top management including the board of directors in business organizations requires information for better decision making in the marketing of products, modes and privileges for satisfying customers, competitor analysis and human resources management. The back-end of the business analytics function usually supports data mining. The business analytics front-end functions compose of executive reporting metrics and organized information. An effective and efficient data mining process execution becomes a core competency for the organization, thereby providing a competitive advantage over its rivals in a form of valuable business intelligence. The practice of data mining within the organization, assures better strategic and contingency plan formulation and execution at organizations [2].

Data mining induces core competencies within the business's internal operations and resources. This is the most suitable and probable cause for an SME to adopt data mining. SMEs also aspire to reach appropriate markets and to evolve the products' brand values within the industry they do business with. It is critical to understand how the core business functions of an SME adds value to its end products or services as per the perspective of customers. Hence, if SMEs can see the benefits that they can yield by data mining different sources of data, then they will be motivated to get started by articulating effective business intellectual inferences [3].

B. CRM

CRM is identified as a business model that integrates all business functions; primary and secondary as logistics, procurement, sales and marketing, and also emphasizes customer expressed satisfaction and unrevealed needs to assists business organizations to build loyal, profitable relationships with customers. It assists the monitoring process, along with, evolving the relationships for a better larger market share. The CRM concept first emerged among the vendor community in the mid-1990s and has since generated much interest among the academic community, afterward [4].

CRM has 4 different dimensions as Customer Identification, Customer Attraction, Customer Retention and Customer Development. These dimensions can be identified as the customer management life cycle. Data mining assists SMEs to analyze their datasets and model information and provides recommendations for each CRM dimensions. Research has reviewed and demonstrated the major data mining techniques that can be used in CRM. Accordingly, most deployed data mining techniques for a CRM are, Association rule, Decision tree, Genetic algorithm, Neural networks, and Clustering and Regression analysis. For an SME, which is initiating CRM, needs to identify its customers. Hence, the customer identification phase is vital to engage in. At the stage of initial customer identification, businesses tend to engage in TCA and customer segmentation. TCA involves seeking the profitable segments of customers through analysis of customers' underlying characteristics, whereas customer segmentation involves the subdivision of an entire customer base into smaller customer groups or segments, consisting of customers who are relatively similar within each specific segment [5]. To commence TCA, businesses engage in Exploratory Data Analysis (EDA) with the aid of internally generated data, as well as third party data. Customer segmentation is identifying different clusters of customers according to enterprise defined objectives. For customer segmentation, there are several data mining techniques such as Principal Component Analysis (PCA), Association rule, Classification and Clustering.

C. Initial customer identification at SMEs

When initiating data mining at SMEs, there should be predetermined customer-oriented objectives, that are being clearly defined by the board of directors. Unless otherwise, a lack of the desired goal in data mining will retard the effectiveness of decision making.

Internally generated data at SMEs tend to have complex dimensionalities, which will create unwanted analytical burdens with regard to data attributes. More data attributes within a dataset will induce distractions. Hence, it is wistful to use dimensionality reduction techniques before the initial customer analysis.

SMEs face several challenges at the commencement of data mining, for initial customer identification. The most challenging factor that influences the retardation in initial customer identification is having small datasets. According to experiments on data mining, satisfactory results were only obtained with a large set of data. For small training sets, the performance and effectiveness in data mining approaches may not be good enough, or the learning task for a computer can even not be accomplished. As a result of this deficiency, the applications of neural networks get severely limited. When considering the root cause for why small datasets cannot provide enough information, the existence of gaps between samples can be noted [6].

The analysis will also be complicated due to a large number of attributes associated with the dataset of SMEs. When SMEs gather data, due to lack of experience, they gather unnecessary, misleading demographical attributes pertaining to customers. Analyzing the most relevant attributes is advantageous to SMEs in terms of computational costs. Hence, the PCA can be deployed to reduce dimensionality and thereby promote the analysis of the most important and relevant attributes of customers for SMEs.

D. PCA

PCA is a data mining approach which is used for dimensionality reduction in a dataset. Usually, customer data gathered at SMEs has lots of components and attributes as discussed. Hence, the dataset is bulky and tedious to be used for an effective data mining process. PCA is a statistical technique of data reduction and aims to produce a more significant, smaller quantity of derived variables representing a larger number of original variables. This analysis helps to simplify subsequent data analysis [7]. PCA is implemented to extract the most important information from the statistical data. It represents those important information as a set of newly formulated orthogonal variables which are known as principal components. PCA also focuses to display the similarity patterns that can be derived between the observations and variables as points in spot maps.

Mathematically, PCA has a greater dependence upon the positive semi-definite matrix Eigen-decomposition and also upon the rectangular matrix Singular Value Decomposition. More importantly, it is determined by eigenvectors and eigenvalues [8]. The main goals of PCA are extraction of the most important data from the dataset, compression of data, data description simplification, increase of the ability to engage in observation and variable structure analysis and data compression through dimension reduction. This would be achieved without a significant information loss and is used in image compression.

III. Methodology

The methodology was formulated considering the research objective of introducing feasible data mining for SMEs in customer identification, considering the need for keeping costs and complexity low. The methodology composed of two phases as TCA is done with the aid of the EDA data mining techniques. The second phase was carried out to identify clusters of customers within the SME using PCA.

The dataset used was accumulated as a result of the campaign hosted for the introduction of mobile applications. Mobile applications at its user registration gather demographical data for each customer. These demographic data were under the attributes of age, gender, the region of residence and occupation along with purchases made by each customer. All the demographic data are in categorical form, whereas the attribute, purchase, the target variable is in numerical form.

Initially, the extracted dataset was preprocessed to reduce the noise and to handle null values. Then the noise reduced dataset was used to engineer the features. Feature extraction conducts an attribute reduction process. It should be clarified that feature selection only ranks the existing attributes considering the predictive significance of each attribute, whereas feature extraction essentially transforms the attributes. The transformation of attributes or features generates linear combinations of the original attributes presented in the dataset [9]. The customer-related dataset mostly consisted of categorical data which need to be encoded for further data mining. Binary encoding was used to transform categorical data attributes to arithmetic-friendly values for further analysis. This feature was deployed with the aid of the Python data mining library, Scikit Learning.

Then the processed customer data underwent the first phase of the analysis, which is EDA. There, the overall customer data distribution for customer profitability was measured. This approach was used to identify the behaviour of customers for the organization's profit margins. Based on different demographic attributes, mainly focusing on age, gender, occupation and residency level, EDA was carried out. This enabled the linking of the profitable target market with customer attributes. Specifically, the univariate distribution of the target variable is considered, concerning purchases versus purchase count. Hence, at a glance, the indirect representation of the profitability distribution at the SME was illustrated. Then further analysis was carried out considering other customer demographic features, to understand the target customers in terms of the demography. The EDA was implemented with the use of Python data mining libraries; Numpy, Pandas, Matplotlib.

The second phase of the methodology was designed to identify the underlying customer segmentation within the SME. PCA was used since the dimensionality of the dataset gets reduced. This improves the analysis with the most prominent features that affect the SME's customer base, highlighted. The total dimensionality was reduced to 12 facets. As data being compressed, without a data loss, two meaningful and prominent features were claimed to be principal components for customer segmentation. Then K-Means clustering was deployed using R to identify the most relevant clusters with less computational efforts.

IV. Results

A. *Phase 1: EDA for TCA*

The EDA conduct checks to ensure the quality of data, summary statistics calculation, and appropriate graph plotting. For SMEs, EDA is crucial, and considering the management viewpoint, objectives were defined. Accordingly, the following EDA results were obtained.

Fig. 1 illustrated below, shows Purchase versus the Purchase Counts, for the target variable, purchase distribution. The Pearson r is given as 0.27.

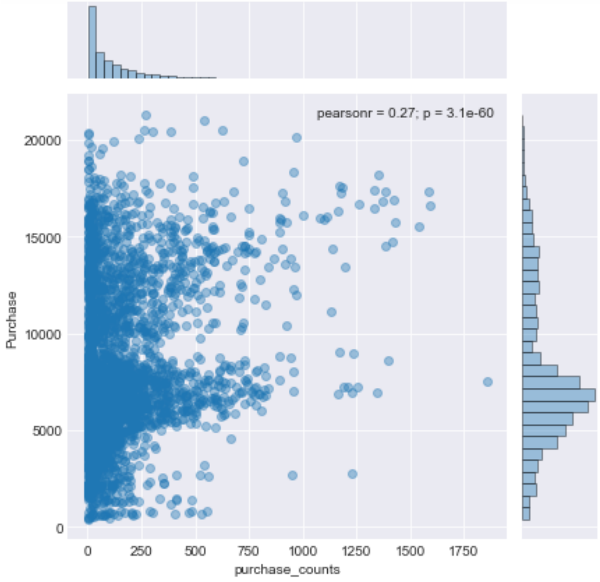

Fig. 1. Purchase vs Purchase count plot

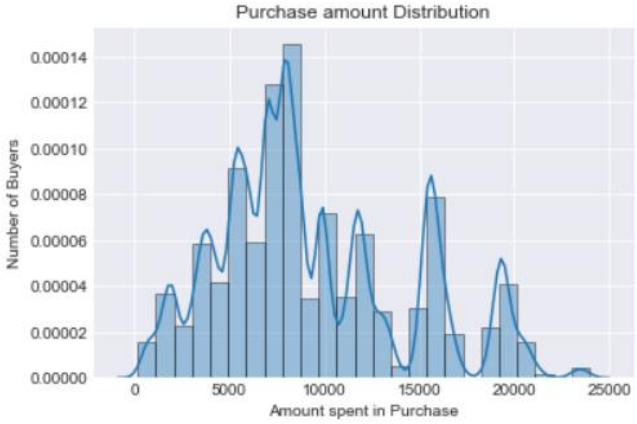

Fig. 2. Purchase amount distribution

Fig. 2, represents how to purchased values are distributed to buyers. It is revealed that purchase amounts between 5000 and 10,000 have attracted many numbers of buyers. This means that the products within the range of 5000-10,000 indicate a trend of fast-moving products.

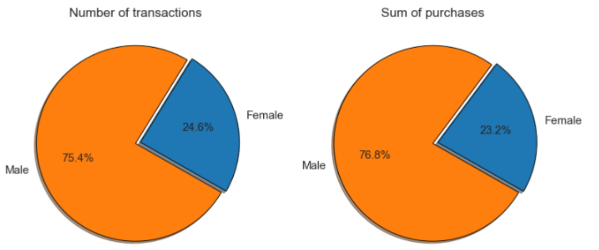

Fig. 3. Sales based on Gender

The above illustration in Fig. 3 shows how the target variable varies based on Gender. The male and female proportion in the number of transactions and sum of purchases show similar behaviours.

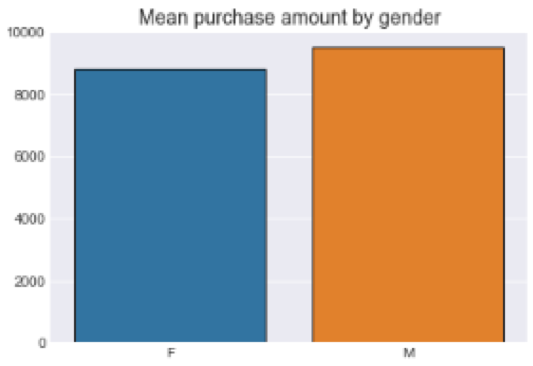

Fig. 4. Mean sales consideration based on Gender

The representation in Fig. 4 implies that the mean purchases of the two genders are almost equal.

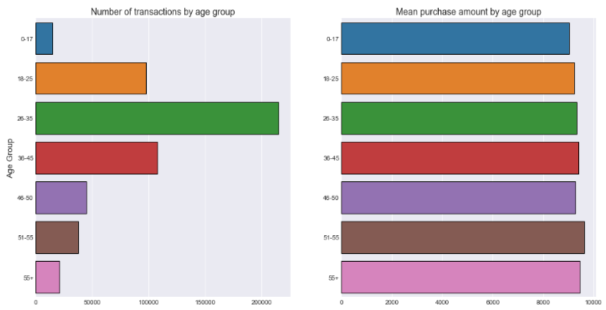

Fig. 5. Age-based sales distribution

As in Fig. 5, the number of transactions indicates wide fluctuations. But the mean purchase amount depicts that the age groups have contributed to purchases in similar amounts.

The illustration in Fig. 6 depicts how the existing customer base behaves within their demographic features for profit margins.

EDA has shown conflicting results which confuse and retard the entire data mining approach within SMEs. Only with an EDA, SMEs get misleading business intelligence and may give up the opportunity to implement the most promising applications of machine learning and deep learning.

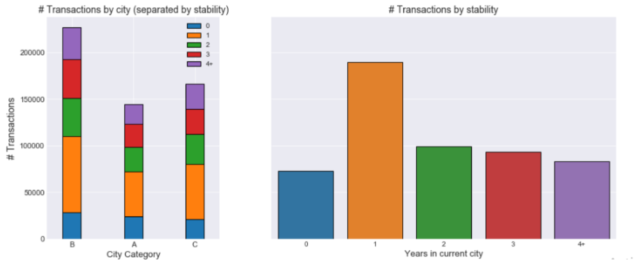

Fig. 6. Sales consideration on customer residency

Hence, PCA can be used to infer the most effective demographical features that an SME should consider for CRM and target marketing. The following results were obtained via PCA conducted using R studio. According to the used dataset, the extended version of PCA, the MCA has used, so that the best mix of sub-demographical features can be extracted for further analysis.

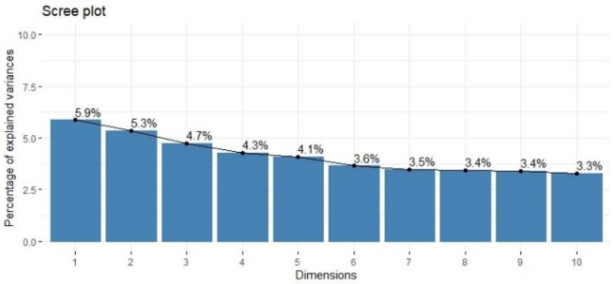

Fig. 7. Dimensionality scree plot

According to the scree plot in Fig. 7, the first data feature dimension holds 5.9% Multiple Component Analysis (MCA) and the second dimension shows 5.3% MCA. The aggregated MCA is around 10%. This means that the dataset does not hold distinguishably different data features. Hence, customers are not distinguishable or different from one another when it comes to purchases.

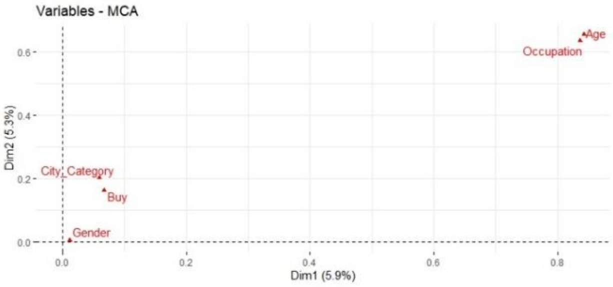

Fig.8. Dimensionality comparison with principal component

According to the representation in Fig. 8, Occupation and Age factors need to have greater focus as these factors will help SMEs to diversify its customer base. Here, the comparative association was considered between the most prominent, first two dimensions.

The following heat map (Fig. 9) shows how sub-attributes impact on the initial customer identification (the first phase of TCA) at SMEs.

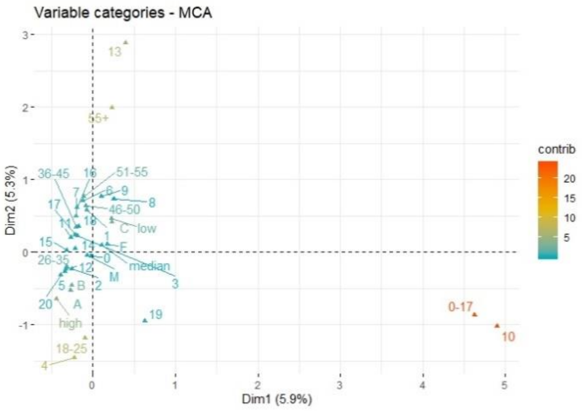

Fig. 9. Feature contribution

Features shown in red color are the directly, impactful features when comes to customer diversification along with the segmentation. Hence, Occupation 10 and Age range 0-17 need to be carefully considered when formulating strategic decisions.

Finally, customer segmentation and identification can be formulated and it is shown in Fig. 10 with 4 clusters.

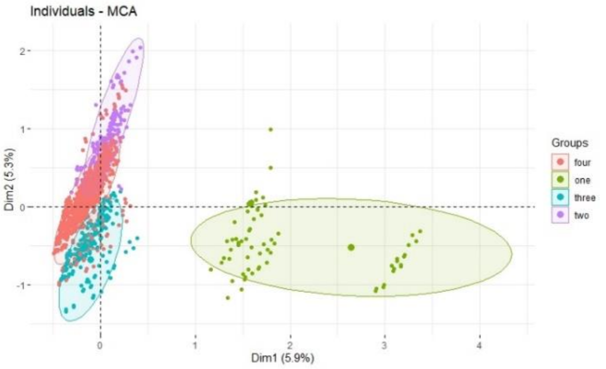

Fig. 10. Final customer segmentation

V. Conclusion

The study initially aimed to introduce a simple, yet robust data mining approach for SMEs to initiate fully pledged CRM. The most crucial step of CRM, which was identifying the existing customer base was achieved via two customer analytical approaches; TCA and customer segmentation.

For TCA, the data mining technique of EDA was used and effective decisions were drawn with simple computational efforts. The extracted, processed dataset yielded meaningful information that facilitates decision making at the SMEs.

The customer segmentation, identifying the most effective and promising customer base in terms of profits was done using PCA. The use of PCA technique compressed the dataset dimensionality and had facilitated to determine the principal components and sub-features which are impacting the customer behavioral and property-based segmentation.

Data mining techniques; EDA and PCA have improved the feasibility of data mining at SMEs in recognition of customers with low computation cost and low complexity.

The use of PCA has induced several theoretical implications. PCA analysis was based on the assumption of linearity. PCA aims to identify the orthogonal projections within the dataset, which has the highest variance, thereby determining the hidden linear correlations between variables. But when considering the data generated at SMEs, there can be non-linearly correlated variables within datasets. Hence, PCA is not adequate for the determination of the most prominent variable. PCA was also based on the assumption that the highest variance and lowest variance are the most important. This assumption holds implications when engaging in blind source separation. But in this study, PCA was used to determine the most prominent data attribute, which is a means of noise reduction. Therefore, this assumption is advantageous to study prediction precision. Theoretically, PCA is not a scale variant analysis, i.e. PCA makes rotation transformation within the dataset, which nullifies the effect of data scale or in other words, PCA does not normalize the dataset. This mentioned effect has little impact on the SMEs datasets, as the scale of data does not change predominantly over a while. But in a case of scale differentiation in variables, PCA will give quite a different set of components as principal components. This can only happen if the SMEs try to deviate from their core business activities or adopt a serious change within their organizations.

The main practical implication that can be encountered is the reluctance of management at SMEs to adopt data mining and machine learning approaches within the business intelligence formulation process. This is mainly due to the belief within SMEs, that predicting the future is not for the small or medium-sized businesses, but for large scale, multinational companies. SMEs need to adopt these approaches to sustain within the industry and also to gain a competitive edge than its rivalries.

References

[1] M. Baer, T. Ariyachandra, and M. Frolick, "Initiating and Implementing Data Mining Practices within a Small to Medium-Sized Business Organization," *J. Econ. Bus. Manag.*, vol. 1, no. 4, 2013.

[2] P. M. Lee, "Use Of Data Mining In Business Analytics To Support Business Competitiveness," *Rev. Bus. Inf. Syst. - Second Quart. 2013*, vol. 17, no. 2, pp. 53–58, 2013.

[3] S. Y. Coleman, "Data-Mining Opportunities for Small and Medium Enterprises with Official Statistics in the UK," *J. Off. Stat.*, vol. 32, no. 4, pp. 849–865, 2016.

[4] N. Gordini, I. Sanpaolo, and V. Veglio, "Customer relationship management and data mining : A classification decision tree to predict customer purchasing behavior in global market," in Handbook of *Research on Novel Soft Computing Intelligent Algorithm*s: Theory and Practical Applications, ch 1, Eds. Vasant Pandia, 2014.

[5] E. W. T. Ngai, L. Xiu, and D. C. K. Chau, "Expert Systems with Applications Application of data mining techniques in customer relationship management : A literature review and classification," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 2592–2602, 2009.

[6] R. Andonie, "Extreme Data Mining: Inference from Small Datasets," *Int. J. Comput. Commun. Control*, vol. V, no. 3, pp. 280–291, 2010.

[7] S. S. Bhadauria and M. Pradesh, "Introduction to Principal Component Analysis in Applied Research," *New Man Int. J. Multidiscip. Stud., vol 1,* 2014.

[8] U. Sarkar, F. Sciences, S. Datta, and F. Sciences, "Principal Component Analysis," *Int. J. Livest. Res.*, vol. 7, pp. 60–78, 2017.

[9] C. Nithya and V. Saravanan, "A Survey of Feature Extraction and Feature engineering In Data Mining," *IOSR J. Eng.*, no. Iccids 2018, pp. 83–87, 2018.