Employee Attrition Prediction Using Random Forest Classifier

DR.A.K.Mariappan Professor
Department of Information technology
Easwari Engineering College
Chennai -89,India.

Rohith Amrose R
Department of Information Technology
Easwari Engineering College
Chennai – 89,India.

Sumitha S
Department of Information Technology
Easwari Engineering College
Chennai – 89,India.

Tharagai S
Department of Information Technology
Easwari Engineering College
Chennai – 89,India.

Abstract -

The application of people analytics in organizations can assist HR managers in reducing employee turnover by changing talent acquisition and retention methods. High employee attrition rates can severely impact a company's productivity and planning continuity, making it a significant threat to its success. By identifying key employee characteristics that affect attrition, this data-driven approach uses a hybrid strategy to create a useful employee attrition model. With decision-makers and business leaders showing an increased interest in machine learning, researchers must explore its application within business organizations. To help HR recruiters make better hiring and placement decisions, we offer a thorough analytics framework that can be used in real-world situations as a decision support tool. A subfield of analytics called HR analytics seeks to improve efficiency and overall results by developing HR units' systems within enterprises. The use of analytics by human resources has been prevalent for many years.

Keywords- Attrition Prediction, support vector machine, Machine Learning, Data Science, Random Forest.

I INTRODUCTION

The procedure of gathering, analysing, and reporting HR data is referred to as HR analytics, sometimes known as people analysis, workforce analysis, or talent analysis. It helps firms to evaluate how different HR indicators affect overall company performance and make data-driven choices. In essence, HR analytics is a well approach to managing human resources. HR analytics is an innovative tool that helps HR managers gain a better understanding of the types of employees who are likely to leave and the characteristics that lead to their departure. Organizations frequently apply analytics to predict an employee's plan to resign, allowing them to place the appropriate people in the appropriate positions at the appropriate times. In order to examine what has happened, descriptive analytics is frequently employed to condense or turn data into valuable information. While descriptive analytics have some value in explaining past events, they are not very useful in predicting future events or outcomes.

This paper's goal is to suggest a thorough data-driven strategy for anticipating and early spotting employee

intends to depart. This technique, in contrast to other efforts, concentrates on little yet information-rich HR data inside large data. The emphasis is on finding the best prediction models with excellent performance to anticipate employee attrition using benchmark and simulated open data. However, the research adds that in order to provide accurate and timely forecasts that are devoid of bias, HR data must be carefully prepared and filtered in addition to model performance. The rationale behind the suggested deep-data-driven strategy is that little data may provide the most business value at the lowest cost compared to vast amounts of big data. The functional dimension and the data dimension are the two dimensions that the study focuses on. The goal of the functional dimension is to evaluate, contrast, and choose the most precise prediction model that can anticipate staff attrition in advance. To assist HR managers in creating retention strategies, the data dimension aims to comprehend the reasons for positive attrition. The study switches from big data to deep data, which is the core feature of the suggested strategy, to solve data difficulties that enterprises may have while using HR analytics. Overall, the research suggests an unique method for predicting and detecting employee intention to quit while resolving data challenges in HR analytics. It makes use of the advantages of tiny but information-rich data.

II LITERATURE REVIEW

Middle level officers are more likely to leave, possibly as a result of a difference in opinion with their senior officer. They noted the key variables that affected employee churn at the company. He derives the two rules in a moderate manner. With the help of both parties, a set of questions was posed, and based on their responses, the researcher came to some conclusions about the workload, goals, employment opportunities, and firm management. According to some research, organizational level effects of termination and dismissal rates. The three-basic entity for the negative was described by S.S.Alduayj and K.Rajpoot in 2018.

Turgut, Y., Kose, Y., Ustundag, A., and Cevikcan's claim is correct that individuals make choices in every aspect of their personal and professional life, including opening a business, purchasing goods, and making future investments. Business managers also have to make a lot of decisions on a daily basis, and it is true that they cannot bear the weight of their decisions alone. That's why business analytics has become an essential tool for businesses in recent years. Business analytics involves complex analysis, optimization, and some algorithms to help businesses improve their operational efficiency and decision-making. By providing faster and more accurate analysis of data, business analytics can help businesses make better decisions, reduce costs, and increase profits.

Hlel, Colomo-Palacios, and Yahia [3] have proposed a method that aims to enhance the accuracy of employee turnover predictions. Unlike traditional big data approaches that focus on collecting large amounts of data, this method emphasizes collecting deep data that is relevant to predicting employee turnover. By analysing data that includes employee satisfaction surveys, performance reviews, and demographic information, HR professionals can gain valuable insights into why employees leave and what factors contribute to employee retention. This approach can help HR professionals develop effective strategies for attracting and retaining talent, which in turn can reduce employee turnover and its negative impact on productivity and workforce continuity. The combination of people analytics, data science, and big data analytics can provide valuable insights that can inform and improve HR decision-making processes.

The article by Jain and Nayyar[4] aims to improve upon the limitations of simple HR-based database systems by developing a more accurate and reliable model for predicting employee attrition. They recognize that these existing systems are not effective in helping enterprises make informed decisions due to their lack of accuracy and predictive power. By using a more sophisticated approach, they hope to provide businesses with more accurate information that can help them make better decisions about employee retention. It's interesting to hear that the method they recommend for predicting employee attrition is able to produce highly accurate results, as this could have significant implications for businesses looking to improve their employee retention rates.

N. Bhartiya, S. Jannu, P. Shukla, and R. Chapaneri[5] used machine learning and data analysis approaches in their study to forecast attrition in businesses. The study consisted of four phases: data acquisition, data conditioning, visualization, and classification. To carry out the study, the researchers employed various, algorithms, within the Python environment. The work of the classification predictions was evaluated using three measures: accuracy score, confusion matrix, and ROC curve. Overall, the study aimed to provide insights into how algorithms and data can be used to predict attrition in firms and help them take proactive measures to reduce employee turnover.

Employee turnover can be a costly and time-consuming process for organizations, particularly when it comes to recruiting and training new personnel. That's why many companies are investing in HR analytics to better understand and predict employee turnover, and to develop strategies to retain top talent. By leveraging data and machine learning algorithms, HR analytics can help companies identify the factors that contribute to turnover and take proactive measures to prevent it. R. Chakraborty, K. Mridha, R. N. Shaw and A. Ghosh[6].

G. Raja Rajeswari, R. Murugesan, R. Aruna, B. Jayakrishnan, and K. Nilavathy [7] all concur. The study aims to find the factors that helps to predict attrition using HR analytics. The researchers argue that retaining permanent personnel is a significant challenge for companies, and when employees leave, it can harm long-term working relationships between employees and the organization. By identifying the factors that contribute to attrition, companies can take steps to reduce turnover and retain their valuable employees.

Srivastava, D.K.Nair, and P[8] argue that employee attrition can have a impact on an organization, both financially and operationally. Predictive analytics, which involves the use of algorithms, can be used to accurately forecast future events, such as employee attrition, based on past and present data. By identifying attrition risks, management can take proactive measures to retain valuable employees, thereby reducing corporate risk. Workforce analytics can help companies in predicting and controlling attrition, resulting in higher quality outcomes.

Akarsu [9] identifies the process of identifying current talent within a company as one of the most significant challenges and top concerns for talent management. In order to make strategic decisions, organizations must take into consideration the human resources at their disposal. The result of the recent study is to determine why some of most experienced and talented employees are leaving their jobs prematurely. Additionally, the research aims to predict when the next group of valuable workers will depart from the company.

Attrition, according to Chowdhury, A.H., and Malakar, S [10]. That sounds like an interesting study. Machine learning techniques can be useful in predicting which employees are at risk of leaving, allowing companies to take proactive measures to address their concerns and keep them on board.

III PROPOSED SYSTEM

The aim of this study is to predict employee attrition rates using various classification algorithms such as Support Vector Machines (SVM), Decision Tree Classifier (DTC), and Random Forest Classifier (RFC). The study conducted a comparative analysis of these algorithms and found that RFC had the highest accuracy at 86.00% followed by logistic regression at 81.40%. The product developed in this study can be used by businesses and government agencies in various industries, such as finance, education, and IT, and can be tailored to meet their specific needs, such as addressing issues related to working from home. To better understand the causes of high attrition, an exploratory study was conducted by reviewing literature, papers, and public datasets offered by researchers and HR specialists, and the identified features were compared.

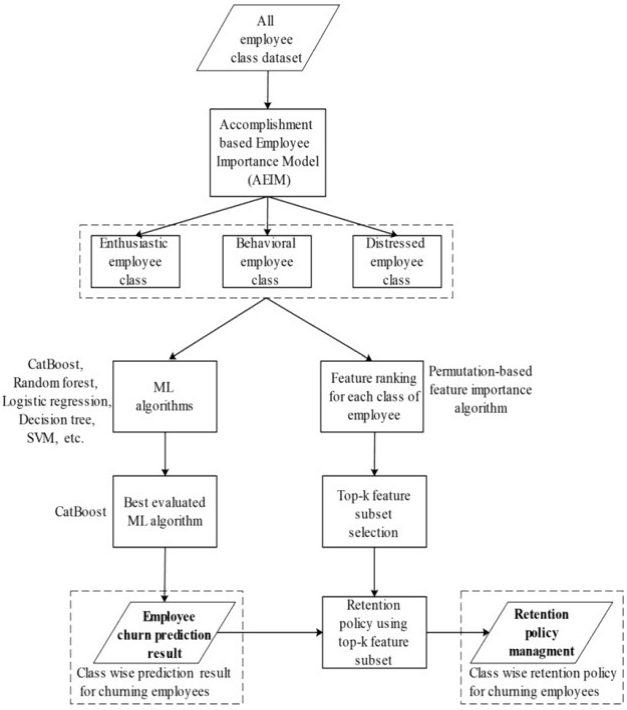

Fig 1: Architecture Diagram

IV.THE PROPOSED SYSTEM

A. Data gathering and Pre-processing

Finding and gathering employee characteristics that are appropriate for our analysis was the first step in this study. This step involves conducting an exploratory analysis of the causes of employee turnover using secondary research and a review of the literature. Predictive model performance may be improved by doing data pre-processing. By transforming and encoding the data provided by respondents, the predictive model can better learn and understand the ideas in the data. One common approach used in pre-processing categorical features is One-Hot Encoding, which involves converting each unique value in the categorical fields to a numerical value. This method can help prevent outliers from influencing the predictions by normalizing data in ranges format. Python's Scikit-learn library offers several functions for data pre-processing, such as Standard Scaler, which can be used to scale numerical features, and Label Encoder, which can be used to encode categorical features. It's important for ML algorithms to know the whole context of the data they're working with, especially when working with data from unknown origins. The model's efficiency and the machine learning training system as a whole might be jeopardized by these statistics.

A method such as visualization must be used to identify these outliers or abnormalities After collection, they are pre-processed for the machine learning algorithm. Memory and processing resources needed for repetitions during training are directly proportional to the size of a dataset. Data pre-processing is the procedure of converting and cleaning raw data in order to create precise predictions using ML algorithms. Real Time projects benefit from ML's decreased process complexity, despite the fact that the most difficult stage of an ML project is data preparation. The term "label encoding" refers to the process of turning labels into a form that can be read by machines, which involves transforming the labels into a numerical form. In addition to this, we are going to remove all of the Nan values from the dataset. During the data preparation phase of machine learning, the major pre-processing works include the removal of Null values.

	Age	Attrition	BusinessTravel	DailyRate	Department
0	41	Yes	Travel_Rarely	1102	Sales
1	49	No	Travel_Frequently	279	Research & Development
2	37	Yes	Travel_Rarely	1373	Research & Development
3	33	No	Travel_Frequently	1392	Research & Development
4	27	No	Travel_Rarely	591	Research & Development
5	32	No	Travel_Frequently	1005	Research & Development
6	59	No	Travel_Rarely	1324	Research & Development
7	30	No	Travel_Rarely	1358	Research & Development
8	38	No	Travel_Frequently	216	Research & Development
9	36	No	Travel_Rarely	1299	Research & Development

Fig 2: Snippet of the dataset

This approach is commonly used in the field of machine learning for binary classification problems, such as predicting whether a customer will purchase a product or not, or whether an email is spam or not. The performance of each model on the same dataset may be compared when various classifiers are used, and tweaking hyperparameters improves each model's performance. During training, the models' performance is assessed using the validation set, and their ultimate performance on untried data is assessed using the test set. This approach is standard practice in machine learning to avoid overfitting to the training data and to ensure that the models generalize well to new data.

B. Feature Analysis

In 2001, Bremen made the initial RF algorithm proposal. It employs a random forest classification for discrete results and a random forest regression for continuous results. The idea that the random forest method has strong noise and anomalous value tolerance as well as a high degree of prediction accuracy has been supported by a number of empirical tests.

Step 1:
The bootstrap random sample technique involves randomly sampling the original dataset with replacement to create new datasets of the same size as the original dataset.

Step 2:
Classification regression tree is created for each of the training sets , and then generate K decision trees.As internal nodes for branches, this approach, which looks at how each tree develops, chooses Modal Features at random rather than the best characteristics.

Step 3:
Makes voting simple.
Due to the independence of each decision tree's training process, random forest training may be completed in parallel, greatly accelerating the process.

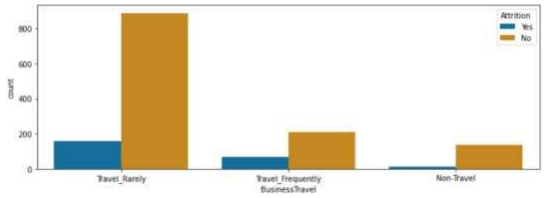

Fig 3: Attrition Business Travel

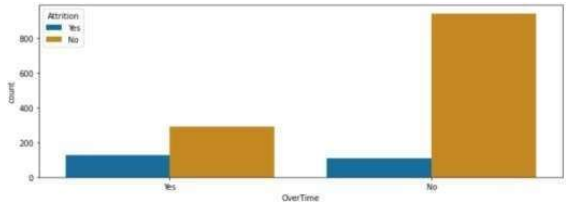

Fig 4: Attrition based on overtime

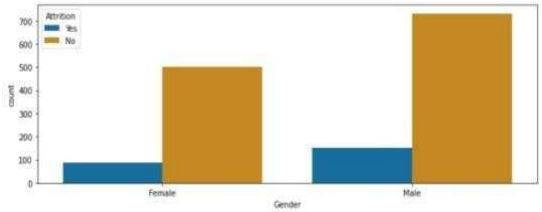

Fig 5: Attrition based on Gender

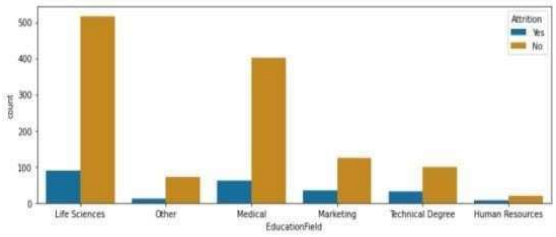

Fig 6: Attrition based on Education

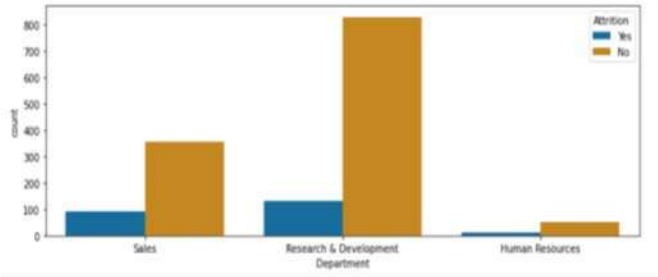

Fig 7: Attrition based on Department

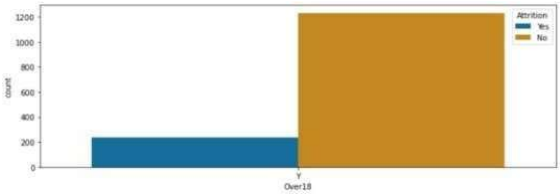

Fig 8: Attrition based on age over18

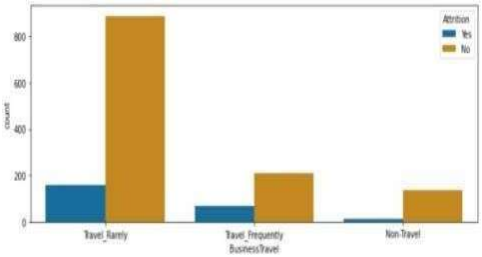

Fig 9: Attrition Business Travel

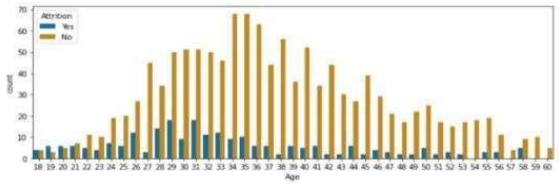

Fig 10:Attrition based on Age

Classifier Evaluation Index: Area under the curve (AUC), recall (RR), precision (PPV), and accuracy (ACC), among others, are often used evaluation indices for the effectiveness of prediction models. The columns of the matrix serve as placeholders for the prediction categories, and the total of those values corresponds to the data observations for that category. Also, the individual values of the matrix indicate the categories, while the total of the values in the matrix represents observations. The binary categorization of employee turnover is the main focus of this study. Turnover falls into the good category, while no turnover falls into the bad category.

In the study, the researchers identified and gathered various employee characteristics that may impact employee attrition, such as basic information, job position information, education and training history, performance reviews, and attendance information. They compared these features using various machine learning algorithms to determine which ones are most predictive of employee attrition. It's to note that the specific features that are most impactful may vary depending on the industry, company size, and other factors. Therefore, it's essential to tailor the feature selection process to the specific context in which the model will be applied.

$$TPR = TP\ (TP + FN)$$

$$\text{FOR} = \text{FP}\ (\text{FP} + \text{TN})\ (2).$$

$$\text{TP} + \text{TN} = \text{TP} + \text{FP} + \text{FN} + \text{TN}.$$

$$\text{TP}\ (\text{TP} + \text{FP}).$$

If all characteristics (dimensions) were considered, the application of the algorithm would be very time and space consuming, which would be detrimental to the method's performance.

As a result, the RF technique is used to first minimise the dimensions.

C. Training the model and result evaluation

One common approach for feature selection in RF is to add noise to each feature and measure the decrease in prediction accuracy. Features that result in the greatest decrease in accuracy when their values are randomly permuted are considered to be the most important. This process can be repeated multiple times to obtain a more accurate ranking of feature importance. By removing less important features, the dimensionality of the dataset can be reduced, which can lead to faster and more accurate machine learning models.

Random Forest Algorithm:

Using the unused data, determine the error rate for each decision-tree in the given algorithm. Furthermore, the efficacy of the decision tree can be assessed using the data.

To validate the model, perform the following steps:

Step 1: Aggregate the results from all sub-classifiers using the F-measure-based weights to determine the final weighted classifier performance for each sample in the unstructured sample set. This is the result of the calculation:

$$H(x) = \text{argmax}\ (\Sigma j\ Wj\ hj(x))$$

where H(x) represents the final classification result for the input sample x, and Wj and hj(x) represent the weight and classification result, respectively, for sub-classifier j.

Step 2: With the test set and the expected results, compute metrics like accuracy, precision, recall, and F1-score to assess the model's performance. The model can also be compared to other models using metrics such as ROC curves and AUC scores. Overall, the weighted random forest algorithm using F-measure-based weights can effectively classify employee attrition with high accuracy and provide insights into the most significant features for predicting employee turnover.

V RESULTS

Fig.11 shows a correlation matrix, which is a useful tool to identify patterns in the data. According to the dataset, working overtime does not have a significant impact on attrition. In terms of gender, men tend to leave the company more often than women. Education also plays a big role in attrition, with employees who studied life science tending to leave the company more frequently. This suggests that HR should consider a particular education background for long-term retention. Research and Development is the department with the highest attrition rate. HR should pay extra attention to this department to improve retention. This may be due to the fact that this is a crucial time for career development, so HR can improve retention by offering educational sessions and career guidance.

Fig.12 shows the prediction of the proposed system, which uses the random forest algorithm to predict the output. Overall, the findings imply that random forest is a useful algorithm for anticipating attrition and that retention strategies have to be customised to the many facets of HR management and worker requirements.

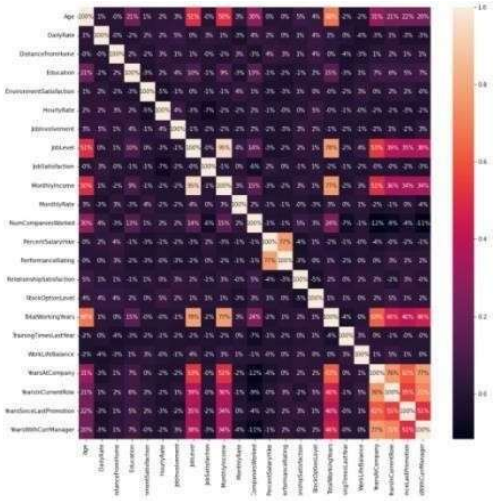

Fig 11. Correlation matrix

A correlation matrix is indeed a useful tool to analyse the relationships between different variables in a dataset. Based on the analysis of the dataset, it seems that working overtime does not have a significant impact on employee attrition. However, it appears that men tend to leave the company more often than women, and employees who have studied life science tend to leave more frequently. This suggests that HR may want to consider educational background when selecting employees to promote long-term retention Additionally, the research and development department has a high attrition rate, indicating that HR should focus on improving retention in this department.

The age range of 27 to 31 appears to be a critical period for employee turnover, as employees in this age range may be looking to improve their career paths. To address these issues, HR may want to provide additional support, such as educational sessions and career guidance, to help employees in this age range develop their skills and achieve their career goals. By addressing these factors, HR may be able to improve employee retention and reduce attrition rates in the organization. Where we tend to travel but in case of business travel it's a sickness for employees so people may find to leave the company. Age 18 is a start of the career and also the confusion part. Where students may find other ideas and tend to leave the company. It works by creating multiple decision trees and combining their results to make a final prediction.

Fig 12: Prediction of Proposed System

VI CONCLUSION

The proposed employee attrition model, along with the predictive analytics methods, can help HR managers identify employees who don't want to work more in t6he organisation and take preventative measures to retain them. Only 11 characteristics and a mixed

research technique are needed for the model to reliably forecast positive attrition and detect departure intent. The recommendation and experimentation of various predictive models can help HR managers choose the best model for their specific needs and dataset size. Finally, the interpretation and explanation of the model's findings can help HR managers understand why employees are leaving and take appropriate actions to improve retention. Overall, this study's contributions can assist HR managers in reducing employee attrition and improving their workforce.

REFERENCES:

1. S. S. Alduayj and K. Rajpoot's paper "Predicting Employee Attrition Using Machine Learning" was presented at the 2018 World Congress on Information Technology Innovations (IIT) in Uae, United Arab Emirates.

2. Ustundag, A., Turgut, Y., Kose, Y., and Cevikcan, E. (2022). Analytics for Managers in Business. Business Analytics for Professionals, edited by A. Ustundag, E. Celikcan, and O. F. Beyca. Advanced Manufacturing Springer Series. Cham Springer.

3 In IEEE Access, N. B. Yahia, J. Hlel, and R. ColomoPalacios published a paper titled "Beyond Big Data to Deeper Data to Support Priorities Of people for Employee ’s turnover Prediction.".

4. Forecasting Employee Turnover intention With Gradient boosted Machine Learning Approach, at SMART 2018, an international conference held in Moradabad, India. A. Nayyar and R. Jain.

5. Employee ’s turnover Prediction Using Classifier Models, 2019 IEEE 5th International Conference for Innovation in Technologies (I2CT), Bombay, India. N. Bhartiya, S. Jannu, P. Shukla, and R. Chapaneri.

6. Study and Forecast Analysis of the Employee Turnover International Conference on Computers, Power and Telecommunications, R. Chakraborty, K. Mridha, R. N. Shaw, and A. GhoshKuala Lumpur, Malaysia: Communication Technologies (GUCON), 2021.

7. Forecasting Employee Attrition Using Machine Learning, 3rd Annual Meeting on Sustainable Communication and Electronics (ICOSEC), Trichy, India, 2022, G. Raja Rajeswari, R. Murugesan, R. Aruna, B. Jayakrishnan, and K. Nilavathy.

8. D.K. Srivastava and P. Nair (2018). Analysis of Employee Attrition Using Predictive Methods. In: Joshi, A., and Satapathy, S. (eds).

9. O. Akarsu, C. Kadayifci, and S.C. (2022). Analytics for human resources. Business Analytics for Professionals, edited by A. Ustundag, E. Celikcan, and O. F. Beyca. Advanced Manufacturing Springer Series. Cham Springer.

10. Seal, D.B., Goswami, S., Malakar, S., and Chowdhury, A.H. (2022). Making Sense of High Employee turnover using Machine Learning. Data Management, Analytics, and Innovation, edited by N. Sharma, A. Chakrabarti, V.E. Balas, and A.M. Bruckstein. Volume 71 of the Lectures on Data Science and analytics and Communications Technologies. Singapore's Springer.