

# A machine learning model for predicting innovation effort of firms

Ruchi Rani<sup>1</sup>, Sumit Kumar<sup>2</sup>, Rutuja Rajendra Patil<sup>2</sup>, Sanjeev Kumar Pippal<sup>3</sup>

<sup>1</sup>Department of Computer Science Engineering, Indian Institute of Information Technology, Kottayam, India

<sup>2</sup>Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India

<sup>3</sup>Department of Technology, Nath School of Business and Technology, Mahatma Gandhi Mission University, Aurangabad, India

## Article Info

### Article history:

Received Sep 1, 2022

Revised Dec 24, 2022

Accepted Feb 3, 2023

### Keywords:

Classification and regression tree

Data mining

Innovation

Innovation predictors

Machine learning

## ABSTRACT

Classification and regression tree (CART) data mining models have been used in several scientific fields for building efficient and accurate predictive models. Some of the application areas are prediction of disease, targeted marketing, and fraud detection. In this paper we use CART which widely used machine learning technique for predicting research and development (R&D) intensity or innovation effort of firms using several relevant variables like technical opportunity, knowledge spillover and absorptive capacity. We found that accuracy of CART models is superior to the often-used linear parametric models. The results of this study are considered necessary for both financial analysts and practitioners. In the case of financial analysts, it establishes the power of data-driven prototypes to understand the innovation thinking of employees, whereas in the case of policymakers or business entrepreneurs, who can take advantage of evidence-based tools in the decision-making process.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Sumit Kumar

Department of Electronics and Telecommunication, Symbiosis Institute of Technology, Symbiosis International (Deemed University)

Pune, Maharashtra, India

Email: er.sumitkumar21@gmail.com

## 1. INTRODUCTION

Research and development (R&D) are one of the most critical issues presently upsetting the growth of entrepreneurs. Over the past decade, a modern production factor has developed a fundamental cause of competitive advantage-knowledge, learning, and ingenuity. All these aspects certainly lead to increased innovation activity and the creation of innovations. This is feasible to attain by engaging monetary bodies in cooperative chains. So, the so-called knowledge spillover [1] effects become a side effect of any defined cooperation with a knowledge base. Smart cities present a highly conducive environment for innovative and user-friendly innovation and can allocate resources to develop new urban and regional innovation systems.

Classification and regression trees (CART) are a family of machine learning techniques popular in several scientific fields that do not assume data normality and user-specified model statements like ordinary least square (OLS) regression. These methods are easy to use and interpret. Regression trees are non-parametric and computationally intensive methods. They can be applied to a large dataset with high dimensions and are resistant to outliers [2]. These methods are helpful even if someone plans to use conventional methods for identifying the essential variables if there are many variables.

Some important algorithms for tree-based regressions are automatic interaction detection (AID), Chi-squared automatic interaction detection (CHAID), CART, and C 5.0. as numerous implementations of these algorithms exist due to a very active machine learning community. In this paper we will discuss only

the CART algorithm followed by an application on a business performance and enterprise survey (BEEPs) dataset from German industry, 2005. Nonlinear innovation models like CART have recently been introduced to accommodate interactive and recursive concepts. However, little attention has been paid to predicting innovation activity using nonlinear models. The authors would like to fill this gap and show that a much more accurate prediction can be achieved using an ensemble of decision trees.

BEEPs data provides detailed performance and innovation indicators, market environment, and company characteristics. Regarding innovation research, center for internet security (CIS) is the most widely used research in recent innovation research. However, CIS does not know whether a company has discontinued a product/service. No data (other than BEEPs) asks the company if they have stopped the product/service. Our experiment is done on a sample of German industry data for the period 2002-2004 chosen from BEEPS 2005 to predict the innovation efforts of firms. We implement CART using the package recursive partitioning and regression trees (*rpart*) in R statistical computing environment.

The error versus the number of split plots was obtained in the results from the first plot. We know that the first split gives the maximum information and adds the most to the model fit. The second plot suggests pruning the tree to include only three splits. The mean difference in innovation effort of firms in the training dataset predicted values of innovative effort (IE) from the linear and the tree model are plotted against the actual values of IE for firms in the testing datasets. Also, the superior fitness of the tree model for predictive purposes is apparent from a visual inspection of the data points. Our findings also indicate a non-linear relationship between the outcome and independent variables. A hierarchical list of the independent variables according to their importance in predicting innovation effort was obtained from the tree model. Knowledge spillovers (KS), a structural variable, top the list for determining firms' R&D intensity or innovation effort.

The work in this paper is divided as follows. The subsequent section will explain various machine learning algorithms and their importance. The section 3 describes the classification and regression trees (CART). Section 4 will describe different application of CART and results and the last the research paper is concluded in the conclusion section.

## 2. MACHINE LEARNING ALGORITHMS

There are numerous machine learning algorithms [3] available in the literature to acquire from a given data so-called train data. A read model analyzes and predicts the desired class when new or invisible data emerges. Algorithms for learning the same predictive machine are explained below.

### 2.1. Naïve Bayes method

This method is based on bayes theorem. The naive bayes method [4] is a standard classification method based on possible possibilities. It usually performs well on complex data sets that are difficult to learn using traditional algorithms.

### 2.2. Support vector machine

Support vector machine (SVM) [5]–[9] is a type of separation strategy in which data points are divided byline in the instance of line SVM and hyperplane in the event of indirect SVM. Parting is chosen so that; the two sides of the hyperplane divide the data into two classes. When anonymous data arrives, it forecasts which side/category it fits to. The boundary among the hyperplane and the supporting vectors is as huge as likely to minimize error in separation.

### 2.3. K-nearest neighbor

The K-nearest neighbor (K-NN) classifier [10] is an indolent learner among the machine learning algorithms. It not ever reads data and does not construct models. Instead, it discovers examples from the training database close to the unknown sample. Based on the criteria of the neighbors will predict a new model. The number “k” governs the number of nearby data points or instances to be nominated from the training model.

### 2.4. Random decision forest

It is an ensemble learning technique employed for classification and regression and can be constructed using a large number of decision trees. One of the study methods for ensembles [11] contains fewer cut trees than a single deciding decision tree. Although separating all the trees from the forest randomly gives the class an anonymous instance, and the course with the most votes will be offered an unknown sample.

## 2.5. Decision trees classifier

Decision tree is among the most widespread learning algorithm in machine learning. C 4.5 [12] is a yardstick learning classifier on the decision tree that is frequently associated to developed novel algorithms. Here is the learning algorithm C 5.0 [13], an improved form of the decision tree classifiers. The nodes and edges are a series of shapes and leaves for class labels.

## 3. LITERATURE REVIEW

As discussed above, a knowledge-based ecosystem can enhance knowledge spillover outcomes in knowledge networks moreover can assist in promoting R&D activities. Spillovers arise inside knowledge-based networks involving knowledge bodies, such as academic research centers and innovation centers. The variety of innovation actions, in-house research and development, and outside knowledge procurement is also a significant factor. Additionally, several surveys and reports confirm the amount of in-house research and development [14], [15]. This kind of R&D improves the likelihood of innovation interest. Several articles address the significance of community monetary assistance for innovation interests [16]. They indicate that it is beneficial when encouraging worldwide cooperating companies. Numerous experimental studies disagree and show that collaboration and KS encourage innovation interests [17]–[19]. The majority of the earlier investigations have been inadequate in using linear (logistic) regression models, and thus the fundamental non-linear attributes of the innovation procedure have not been believed.

## 4. CLASSIFICATION AND REGRESSION TREES

Numerous experimental studies disagree and show that collaboration and KS encourage innovation interests [17]–[19]. The majority of the earlier investigations have been inadequate in using linear (logistic) regression models, and thus the fundamental non-linear attributes of the innovation procedure have not been believed. CART are built using recursive partitioning, a heuristic technique [20]–[22]. There are many algorithms to implement this technique. The first challenge of any decision tree algorithm is to identify the attribute which can be used to obtain the best split of the sample data using some criteria of purity and subsequently apply these criteria to each subgroup recursively. While Gini or log-likelihood function is used for splitting by *rpart* for classification tree, for regression tree "anova" method is used. The criteria in "anova" is maximizing SST-(SSL+SSR) whereas  $SST = \sum (y_i - \hat{y})^2$  is the sum of the squares for the node, and secure sockets layer (SSL) and solid state relay (SSR) are the sum of the squares for the left and right child nodes respectively [23].

The second important thing is to avoid overfitting to data sample and keeping it right for the purpose of prediction. It can be summarized as: If  $T_1, T_2, \dots, T_K$  are the terminal nodes of the tree  $T$ ,  $|T|$  is number of terminal nodes in (1).

$$\text{Risk of } T = R(T) = \sum_{i=1}^k P(T_i)R(T_i) \quad (1)$$

Here,  $|T|$  is equivalent to degrees of freedom and  $R(T)$  to the remaining sum of squares when compared to regression. Let  $\alpha$  be the complexity parameter which represents the cost of addition of alternative variable to the model and it ranges from 0 to  $\infty$ . Then we can calculate the cost of a tree as (2).

$$R\alpha(T) = R(T) + \alpha|T| \quad (2)$$

The function *rpart* tries to minimize this cost by selecting subtrees  $T\alpha$  with minimal cost. Cross-authentication is used to find a finest value for  $\alpha$ . It is implemented through the advisory threshold complexity parameter of *cp* as (3).

$$Rcp = R(T) + cp * |T| * R(T_1) \quad (3)$$

where  $|T|$  is the number of splits in the tree,  $T_1$  is the tree with no splits and  $R$  is the risk. This equation has no unit and hence more user responsive than the original CART formula in (2) given above. When the value of *cp* is put equal to 1, it results in a tree with no splits. For regression models, a split is decreed apriori not worth pursuing if it does not lead to value by at least *cp*. The default value of 0.01 has been very successful and the cross-validation often needs to remove one or two layers only.

A number of techniques can be used for measuring prediction accuracy of a regression tree. Summary statistics [24]–[26] of the model predictions can be used for comparing the range of the projected values with the real values for the outcome variable using a test dataset. Correlation between the projected

values and the real values also provide a simple measure about the model's performance. Another statistical measure for a model's performance, mean absolute error (MAE), measures on an average how far the predictions are from the true values. When  $n$  is the number of predictions and  $e$  is the error of prediction  $i$ , MAE is given in (4). Prediction accuracy of regression trees can be further improved by ensemble learning techniques. However, ensemble learning is out of scope for this work.

$$MAE = \frac{1}{n} \sum |e_i| \quad (4)$$

## 5. APPLICATIONS AND RESULTS

We apply the CART algorithm using *rpart* package on a sample of German industry data for the period 2002-2004 chosen from BEEPS 2005. Our model consisted of the following variable some of which were suitable pre-processed for goodness of use. It consisted of a total of 1,196 firms from which 1,122 data were usable.

### 5.1. The variables

The different variables used in this research are IE, technological opportunity (TO), KS and absorptive capacity measured using firm size. The effort to innovate is the dependent variable in the model. It imitates the volume of resources a company offers to carry out innovative actions over a certain period. IE is numeric and is equal to the yearly R&D expenditure of a firm/annual sales. At the same time, TO is a dummy with levels 1 and 0 for a high and low technological opportunity, respectively. KS is also numeric. Absorptive capacity measured using firm size (numeric), percentage of a firm's annual sales from direct export (numeric), Acquisition of new production technology in the last 36 months (dummy), percentage of engineers and scientists among firm employees (numeric).

### 5.2. Results

We first randomized the data sample and then divide two sets, training, and testing dataset. The training dataset had 900 observations, 80% of the total; and the testing dataset contained the rest 20% equaling 222. Subsequently a regression tree model was created for the training data set with *Inn3* as the outcome variable using *rpart* function from the R library *rpart*. This produces a tree with 14 splits. Six variables except the variable *firm\_age* is used to construct this tree. The importance of the variable for splitting the data in chronological order is KS, firm size, export sales, marketing expenditure, firm age, qualified personnel, and technological opportunities.

Using the function *rsq.rpart* jackknifed error versus the number of splits plots were obtained and plotted as shown in Figures 1 and 2. From the first plot we know that the first split gives the maximum information and adds the most to the model fit. The second plot suggests pruning the tree to include only three splits. We found that complexity parameter (*cp*) value of the tree with only three splits is 0.049569 and gave the least value for cross-validated error. The pruned tree uses only two variables, namely KS and firm size and plotted in Figure 3.

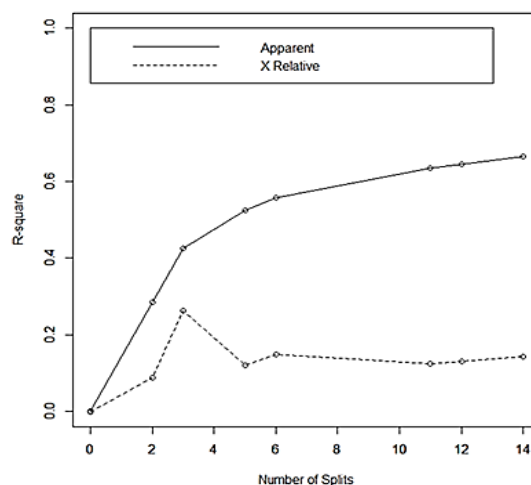


Figure 1. Effects of selecting different switching under dynamic condition

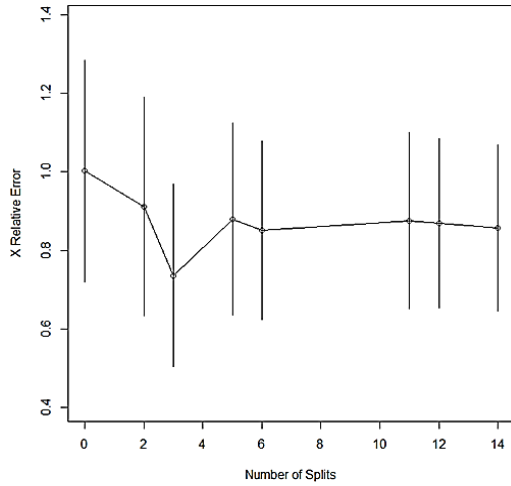


Figure 2. X relative error vs number of splits

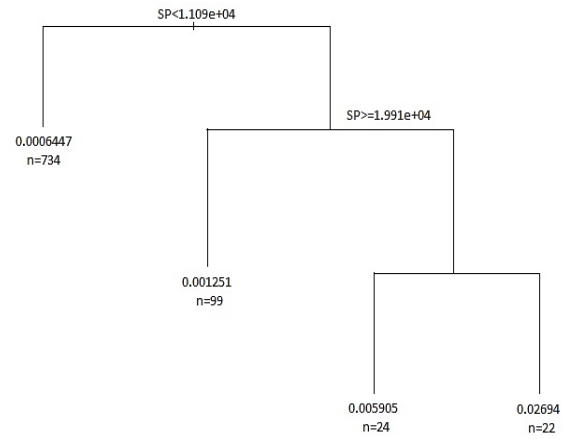


Figure 3. Pruned tree

The tree model was then validated for their capability to accurately predict the innovation effort of firms by using the validation set method. The validation set was previously unexposed to the model. The correlations between the predicted and actual values of the innovation effort for the validation set was found to be 0.7624708. and the testing dataset is 0.002779919. While the prediction of the linear model is 7.84% closer than this mean difference, the tree model does a much better job with 61.41% of the mean value. The same value for a linear model OLS was 0.3087624. Similarly, the mean absolute average error (MAE) of prediction for the linear regression model was calculated at 0.002562028 and at 0.001072886 for the tree model. The mean difference in innovation effort of firms in the training dataset predicted values of IE from the linear and the tree model are plotted against the actual values of IE for firms in the testing datasets in Figures 4 and 5 respectively. The superior fitness of the tree model for predictive purpose is apparent from a visual inspection of the data points. Our findings also indicate toward a non-linear relationship among the outcome variable and the independent variables. A hierarchical list of the independent variables according to their importance in predicting innovation effort was obtained from the tree model. KS, a structural variable, tops the list for determining the R&D intensity or innovation effort of firms.

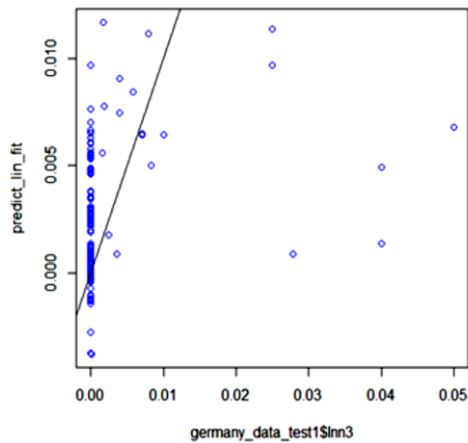


Figure 4. Linear model fit

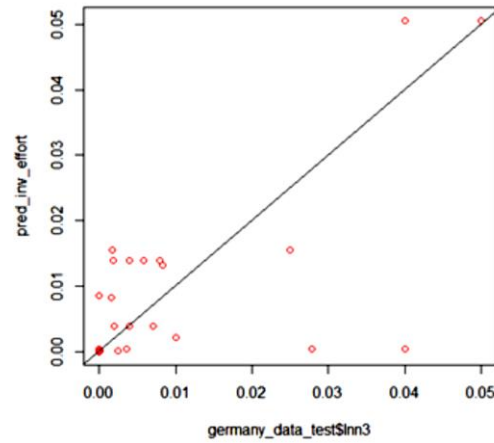


Figure 5. X tree model fit

## 6. LIMITATION AND FUTURE WORK

The main limitation of this study is the execution of the model, which achieves superior accuracy, only shifting the operating point: we do not look for a constant value of innovation bias, but we attempt to recognize those firms that are expected to turn out innovation expenditures. Another limitation of the study is the exclusion of micro firms. We can add these two points as a future work extension for better results.

## 7. CONCLUSION

Our findings here indicate toward a non-linear relationship among the outcome variable and the independent variables. The superior fitness of the tree model for predictive purpose when data normality and linearity of the relationship is doubtful is apparent from a visual inspection of the model fit figures. Additionally, a hierarchical list of the independent variables according to their importance in predicting innovation effort can also be obtained from the tree model. This list can subsequently be used for traditional modeling technique. So, in this paper, CART which widely used machine learning technique for predicting R&D intensity or innovation effort is successfully employed.

This paper analyses the prediction of R&D intensity or innovation effort of firms using several relevant variables like technical opportunity, knowledge spillover and absorptive capacity. We found that accuracy of CART models is superior to the often-used linear parametric models. The results of this study are considered necessary for both financial analysts and practitioners. In the case of financial analysts, it establishes the power of data-driven prototypes to understand the innovation thinking of employees, whereas in the case of policymakers or business entrepreneurs, who can take advantage of evidence-based tools in the decision-making process.

We achieved significantly higher accuracy by bagging the decision tree. Still, it added complexity because the regression tree model was built on the training dataset using 1,196 companies and generated a 14-split tree. Therefore, in the future, we not only proposed to use alternative machine learning and soft computing techniques such as rule-based evolutionary systems and fuzzy rule-based systems but also used the trade-off between bias and decentralization for predicting innovation activities by encouraging further research. Finally, we recommend that you include additional input variables relevant to the business cluster initiative and the regional innovation system.




## REFERENCES

- [1] M. Gu, C. Shu, and D. De Clercq, "Knowledge filters and employee venturing behaviors: a cross-institutional study of the U.S. and Indian firms," *IEEE Transactions on Engineering Management*, pp. 1–14, 2022, doi: 10.1109/TEM.2021.3065955.
- [2] D. Steinberg and P. Colla, "CART: tree-structured non-parametric data analysis," San Diego, CA: Salford Systems, 1995.
- [3] R. R. Patil, S. Kumar, and R. Rani, "Comparison of artificial intelligence algorithms in plant disease prediction," *Revue d'Intelligence Artificielle*, vol. 36, no. 2, pp. 185–193, Apr. 2022, doi: 10.18280/ria.360202.
- [4] K. P. Murphy, "Naive bayes classifiers," University of British Columbia, vol. 18, no. 60, pp. 1–8, 2006.
- [5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
- [6] R. R. Patil and S. Kumar, "Rice-fusion: a multimodality data fusion framework for rice disease diagnosis," *IEEE Access*, vol. 10, pp. 5207–5222, 2022, doi: 10.1109/ACCESS.2022.3140815.
- [7] R. R. Patil and S. Kumar, "Priority selection of agro-meteorological parameters for integrated plant diseases management through analytical hierarchy process," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 1, pp. 649–659, Feb. 2022, doi: 10.11591/ijece.v12i1.pp649-659.
- [8] R. R. Patil and S. Kumar, "Predicting rice diseases across diverse agro-meteorological conditions using an artificial intelligence approach," *PeerJ Computer Science*, vol. 7, Sep. 2021, doi: 10.7717/peerj-cs.687.
- [9] R. R. Patil and S. Kumar, "A bibliometric survey on the diagnosis of plant leaf diseases using artificial intelligence," *Library Philosophy and Practice*, pp. 1–26, 2020.
- [10] L. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, 2009, doi: 10.4249/scholarpedia.1883.
- [11] D. Gupta, D. S. Kohli, and R. Jindal, "Taxonomy of tree based classification algorithm," in *2011 2<sup>nd</sup> International Conference on Computer and Communication Technology (ICCT-2011)*, Sep. 2011, pp. 33–40, doi: 10.1109/ICCT.2011.6075191.
- [12] L. Breiman, "Random forests-random features," Technical Report 567, Department of Statistics, UC Berkeley, 1999.
- [13] A. Galathiya, A. Ganatra, and C. Bhensdadia, "Classification with an improved decision tree algorithm," *International Journal of Computer Applications*, vol. 46, no. 23, pp. 1–6, 2012.
- [14] L. Abramovsky, E. Kremp, A. López, T. Schmidt, and H. Simpson, "Understanding co-operative innovative activity: evidence from four European countries," *Economics of Innovation and New Technology*, vol. 18, no. 3, pp. 243–265, Apr. 2009, doi: 10.1080/10438590801940934.
- [15] W. Becker and J. Dietz, "R&D cooperation and innovation activities of firms-evidence for the German manufacturing industry," *Research Policy*, vol. 33, no. 2, pp. 209–223, Mar. 2004, doi: 10.1016/j.respol.2003.07.003.
- [16] H. Hottenrott and C. Lopes-Bento, "(International) R&D collaboration and SMEs: The effectiveness of targeted public R&D support schemes," *Research Policy*, vol. 43, no. 6, pp. 1055–1066, Jul. 2014, doi: 10.1016/j.respol.2014.01.004.
- [17] M. J. Nieto and L. Santamaría, "The importance of diverse collaborative networks for the novelty of product innovation," *Technovation*, vol. 27, no. 6–7, pp. 367–377, Jun. 2007, doi: 10.1016/j.technovation.2006.10.001.
- [18] M. Frenz and G. Ietto-Gillies, "The impact on innovation performance of different sources of knowledge: Evidence from the UK community innovation survey," *Research Policy*, vol. 38, no. 7, pp. 1125–1135, Sep. 2009, doi: 10.1016/j.respol.2009.05.002.
- [19] J. Gallego, L. Rubalcaba, and C. Suárez, "Knowledge for innovation in Europe: The role of external knowledge on firms' cooperation strategies," *Journal of Business Research*, vol. 66, no. 10, pp. 2034–2041, Oct. 2013, doi: 10.1016/j.jbusres.2013.02.029.
- [20] B. Li, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees (CART)," *Biometrics*, vol. 40, no. 3, pp. 358–361, 1984.
- [21] C. D. Sutton, "Classification and regression trees, bagging, and boosting," in *Handbook of Statistics*, Elsevier, 2005, pp. 303–329.
- [22] T. M. Therneau and E. J. Atkinson, "An introduction to recursive partitioning using the RPART routines," Technical Report Department of Health Science Research, Mayo Clinic, Rochester, Minnesota, 1997.




- [23] P. Yang, Y. Hwa Yang, B. B. Zhou, and A. Y. Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, no. 4, pp. 296–308, Dec. 2010, doi: 10.2174/157489310794072508.
- [24] B. H. Hall, "The financing of research and development," *Oxford Review of Economic Policy*, vol. 18, no. 1, pp. 35–51, Mar. 2002, doi: 10.1093/oxrep/18.1.35.
- [25] European Commission, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs, *Regional Innovation Scoreboard 2017*, Publications Office, 2017, doi: 10.2873/593800.
- [26] F. Bogliacino and M. Pianta, "The Pavitt Taxonomy, revisited: patterns of innovation in manufacturing and services," *Economia Politica*, vol. 33, no. 2, pp. 153–180, Aug. 2016, doi: 10.1007/s40888-016-0035-1.

## BIOGRAPHIES OF AUTHORS






**Ruchi Rani**    received the bachelor's degree in Computer Science engineering from Kurukshetra University, Kurukshetra, India in 2008, the master's degree from Maharshi Dayanand University, Haryana, India in 2012. She is currently pursuing the Ph.D. degree with the department of Computer Science Engineering, Indian Institute of Information Technology Kottayam, Kerala, India. Her research interests include machine learning, and deep learning. She can be contacted at email: ruchiasija20@gmail.com.






**Sumit Kumar**    received the bachelor's degree in Electronics and Telecommunication from Kurukshetra University, Kurukshetra, India in 2005, the master's degree from Guru Jambheshwar University of Science and Technology, Haryana, India in 2008, and Ph.D. degree from Jamia Millia Islamia, Delhi, India in 2017. Currently, He is working as an Associate Professor at Electronics and Telecommunication Department of Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra, India. His research areas are artificial intelligence, antenna design, IoT, wireless networks, wireless communication, and computational intelligence. He can be contacted at email: er.sumitkumar21@gmail.com.



**Rutuja Rajendra Patil**    received the bachelor's degree as well as master's degree in Information Technology from Pune University, Pune, India, in 2006 and 2015, respectively. She is currently pursuing the Ph.D. degree with the Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune. She is currently a Research Scholar at Symbiosis International (Deemed University). Her research interests include machine learning and deep learning. She can be contacted at email: rutujapat@gmail.com.



**Sanjeev Kumar Pippal**    he is B. Tech from MJP Rohilkhand University, M. Tech and Ph.D from MNNIT Allahabad, his area of interest is cloud computing, distributed computing and block chain. He is certified in Machine and Deep Learning. He has published more than 30 research papers in SCI/Scopus International Journals and Conferences. He has filed four patents and published 02 patents. He was instrumental in setting up a 21 cr AI lab at AKTU, Lucknow. He has participated and coordinated as chair four BOS meetings at NIT Kurukshetra, Galgotias University and Chandigarh University. He has vast administrative experience as hod, dean, director at above mentioned organizations. He can be contacted at email: sanpippalin@gmail.com.