IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol. 8, Issue.10, October 2020 E-ISSN: 2347-2693

Machine Learning Techniques for Cancer Prediction: A Survey

Archana Pathak[1*], Nidhi Ruthia[2], Chetan Agrawal[3]

[1,2,3]Computer Science & Engineering, Radharaman Institute of Technology & Science, Bhopal, India

[*]*Corresponding Author: priya.pathak148@gmail.com, Tel.: 9304951335*

DOI: https://doi.org/10.26438/ijcse/v8i10.174179 | Available online at: www.ijcseonline.org

Received: 20/Oct/2020, Accepted: 25/Oct/2020, Published: 31/Oct/2020

***Abstract*—** In the present age of innovation medicinal field researcher are very much interested in disease classification for the analysis of disease. It is a point of concern on the grounds that real treatment of this disease isn't found till date. Patients having this ailment must be spared if and just in the event that it is found in beginning period (arrange I and stage II). In the event that it is identified in last stage (arrange III and stage IV) at that point possibility of endurance will be exceptionally less. Machine learning and information mining system will assist medicinal with handling to handle with this issue. Cancer growth has different manifestations, for example, tumor, abnormal bleeding, more weight reduction and so forth. It isn't vital that a wide range of tumors are harmful. Tumors are fundamentally of two kinds one is benign and the other one is malignant. To give suitable treatment to the patients, side effects must be contemplated appropriately and a programmed expectation framework is required which will characterize the tumor into benevolent or harmful. In the present web world, majority of information is created via web-based networking media or medicinal services sites. From this immense measure of information, side effects can be gotten by utilizing information mining method, which will be further helpful for disease location or classification. This paper makes study of such most recent research study that utilizes on the web and disconnected information for malignant growth arrangement.

***Keywords*—** Tumors, Machine Learning, binary classification, malicious, gene expression

I. INTRODUCTION

Early detection of cancer is very much important for a quick reaction and better chances for cure. But early recognition of malignancy is very difficult on the grounds that the side effects of the malady toward the start are missing. In this manner, cancer stays one of the subjects of health research, where numerous specialists have contributed with the point of making proof that can improve treatment, avoidances and diagnostics. Research around there is a mission of information through studies, studies and trials directed with applications so as to find and decipher new information to forestall and limit the risk adverse consequences. To comprehend these issues more accurately, devices are as yet expected to assist oncologists with choosing the treatment required for mending or counteraction of repeat by decreasing the hurtful impacts of specific medicines and their expenses. To develop devices for cancer management, the machine learning techniques and clinical variables, for example, patient age and histopathological factors structure the reason for day by day decision making. A few investigations have been created in this theme by utilizing the gene expressions [1] or using image processing [2]. In machine learning there are two types: the supervised and unsupervised learning. The first concedes that the classes used to arrange the information are known ahead of time and the second, the classes are not known. Among the techniques, there are: Support Vector Machines [3], Decision Tree [4], Neural Network [5], Bayesian systems [6], K-Nearest Neighbors [7], and so on.

Given the noteworthiness of customized medication and the developing pattern on the use of ML procedures, we here present an audit of concentrates that utilize these techniques with respect to the cancer prediction and prognosis. In these examinations prognostic and prescient features are viewed as which might be free of a specific treatment or are incorporated so as to direct treatment for malignant growth patients, separately. Likewise, we examine the sorts of ML strategies being utilized, the kinds of information they coordinate, and the general execution of each proposed plan while we additionally talk about their upsides and downsides.

A conspicuous pattern in the proposed works incorporates the joining of blended information, for example, clinical and genomic. In any case, a typical issue that we saw in a few works is the absence of outside approval or testing with respect to the prescient presentation of their models. Plainly the utilization of ML strategies could improve the exactness of disease defenselessness, repeat and endurance expectation. Based on accuracy of cancer classification growth expectation result has altogether improved by 15%–20% the most recent years, with the utilization of ML procedures.

A few examinations have been accounted in the literature and depend on various procedures that could empower the

early malignant growth analysis and visualization [8]. In particular, these investigations portray approaches identified with the profiling of flowing miRNAs that have been demonstrated a promising class for cancer identification and detection. Although these techniques experience the ill effects of low affectability with respect to their utilization in screening at beginning periods and their trouble to separate benign from harmful tumors. Different angles with respect to the expectation of malignancy result dependent on quality articulation marks are examined in [9]. These examinations list the potential just as the impediments of microarrays for the forecast of malignant growth result. Despite the fact that quality marks could essentially improve our capacity for forecast in malignant growth patients, poor advancement has been made for their application in the centers. In any case, before quality articulation profiling can be utilized in clinical practice, considers with bigger information tests and progressively sufficient approval are required.

In this paper we have discuss about the various machine learning algorithms and how these algorithm can perform during cancer classification. In the sections 2 there is a detailed review done by different authors in the area of machine learning classification on the dataset based on different type of diseases. After this in section 3 we discuss about the different machine learning classification algorithms followed with the problem identification during analysis of cancer classification in section 4. The last section 5 covers the conclusion and the future work in the area of cancer research.

II. LITERATURE SURVEY

Chen Y-C et al. [10] has performed cross-laboratory validations for the cancer patient data from 4 hospitals. The investigation is about the feasibility of survival risk prediction using gene expression data. The analysis is done on lung cancer data and has used ANN architecture on the training data. Five survival correlated genes are identified from the four microarray gene expression data and data is taken from multiple sources based on lung cancer diagnosis. After constructing multiple ANN architecture on the training data they have achieved the accuracy of 83% and this is based on the trusted data.

Park K et al. [11] has done their research on three prominent machine learning models for breast cancer survival prediction. The models used were support vector machines, artificial neural networks, and semi-supervised learning models. They have used the dataset which is based on the cancer incidence in the United States. In this work semi supervised learning model has shown the improved performance than the other two. They have also improved the model accuracy by reducing the noise in the available data.

Xu X et al. [12] had developed an efficient feature selection method: the support vector machine-based recursive feature elimination (SVM-RFE) approach for gene selection and prognosis prediction. Using the leave-one-out evaluation procedure on a gene expression dataset including 295 breast cancer patients. They discovered a 50-gene signature that by combing with SVM, achieved a superior prediction performance with 34%, 48% and 3% improvement in Accuracy, Sensitivity and Specificity, compared with the widely used 70-gene signature.

Gevaert et al. [13] has evaluated three methods for integrating clinical and microarray data. The decision integration, partial integration and full integration are used to classify publicly available data on breast cancer patients into a poor and a good prognosis group. The partial integration method is most promising and has an independent test set area under the ROC curve of 0.845. After choosing an operating point the classification performance is better than frequently used indices.

Rosado et al. [14] is to develop an intelligent and efficient model, based on Support Vector Machines (SVM), able to predict prognosis in patients with oral squamous cell carcinoma (OSCC). A total of 34 clinical and molecular variables were studied in 69 patients suffering from an OSCC. Variables were selected by means of two methods applied in parallel (Non-concave penalty and Newton's methods). The implementation of a predictive model was performed using the SVM as a classifier algorithm. Finally, its classification ability was evaluated by discriminant analysis. Recurrence, number of recurrences, and TNM stage have been identified as the most relevant prognosis factors with both used methods. Classification rates reached 97.56% and 100% for alive and dead patients, respectively (overall classification rate of 98.55%). SVM techniques build tools able to predict with high accuracy the survival of a patient with OSCC.

Delen et al. [15] has used two popular data mining algorithms that are artificial neural networks and decision trees along with a most commonly used statistical method logistic regression to develop the prediction models using a large dataset almost more than 200,000 cases. They used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes. The results has shown that the decision tree (C5) is the best predictor with 93.6% accuracy on the holdout sample artificial neural networks came out to be the second with 91.2% accuracy and the logistic regression models came out to be the worst of the three with 89.2% accuracy.

Kim and Shin [16] has utilized unlabeled patient data, which is relatively easier to collect. Therefore, it is regarded as an algorithm that could circumvent the known difficulties. However, the fact is yet valid even on SSL that more labeled data lead to better prediction. To compensate for the lack of labeled patient data, they consider the concept of tagging virtual labels to unlabeled patient data,

that is, 'pseudo-labels,' and treating them as if they were labeled. The implemented algorithm 'SSL Co-training', based on SSL. SSL Co-training was tested using the surveillance, epidemiology, and end results database for breast cancer and it delivered a mean accuracy of 76% and a mean area under the curve of 0.81.

III. MACHINE LEARNING TECHNIQUES

ML, a part of Artificial Intelligence, relates the issue of learning from data samples to the general idea of interference [17]. Each learning procedure comprises of two stages: (I) estimation of obscure conditions in a framework from a given dataset and (ii) utilization of assessed conditions to anticipate new yields of the framework. ML has likewise been demonstrated an intriguing zone with regards to biomedical research with numerous applications, where a satisfactory speculation is gotten via looking through a n-dimensional space for a given arrangement of organic examples, utilizing various systems and calculations [18]. In supervised machine learning a marked arrangement of preparing information is utilized to gauge or guide the info information to the ideal yield. Interestingly, under the unsupervised learning strategies no marked models are given and there is no thought of the yield during the learning procedure. Accordingly, it is up to the learning plan/model to discover designs or find the gatherings of the info information. In directed learning this strategy can be thought as an order issue. The errand of arrangement alludes to a learning procedure that orders the information into a lot of limited classes. Two other basic ML undertakings are regression and clustering. On account of regression issues, a learning capacity maps the information into a genuine worth variable. Along these lines, for each new example the estimation of a prescient variable can be evaluated, in light of this procedure. Clustering is a typical solo assignment where one attempts to discover the classes or groups so as to portray the information things. In light of this procedure each new example can be doled out to one of the distinguished clusters concerning the comparable attributes that they share.

Assume for instance that we have gathered therapeutic records significant to breast cancer and we attempt to anticipate if a tumor is harmful or favorable dependent on its size. The ML question would be alluded to the estimation of the likelihood that the tumor is harmful or no (1 = Yes, 0 = No). Fig. 1 portrays the classification procedure of a tumor being dangerous or not. The surrounded records portray any misclassification of the sort of a tumor created by the strategy.

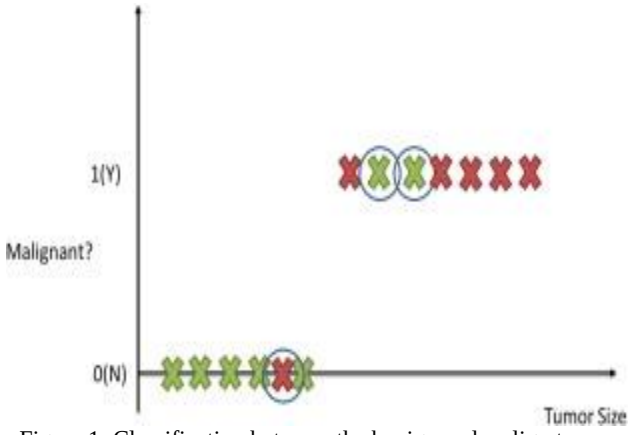

Figure 1: Classification between the benign and malign tumor

While applying a ML strategy, information tests establish the essential segments. Each example is depicted with a few features and each component comprises of various kinds of qualities. Besides, knowing ahead of time the particular sort of information being utilized permits the correct determination of apparatuses and strategies that can be utilized for their investigation. A few information related issues allude to the nature of the information and the preprocessing steps to make them progressively reasonable for ML. Information quality issues incorporate the nearness of clamor, anomalies, absent or copy information and information that is one-sided unrepresentative. While improving the information quality, normally the nature of the subsequent examination is additionally improved. Furthermore, so as to make the crude information increasingly appropriate for additional investigation, preprocessing steps ought to be applying that subject to the change of the information. Various procedures and systems exist, pertinent to information preprocessing that emphasis on altering the information for better fitting in a particular ML strategy. Among these strategies the absolute most significant methodologies incorporate (I) dimensionality decrease (ii) include determination and (iii) feature extraction. There are numerous advantages with respect to the dimensionality decrease when the datasets have countless features. ML calculations work better when the dimensionality is lower [19]. Also, the decrease of dimensionality can kill immaterial features, diminish clamor and can deliver increasingly hearty learning models because of the association of less features. When all is said in done, the dimensionality decrease by choosing new features which are a subset of the old ones is known as feature determination. Three principle approaches exist for feature selection to be specific inserted, channel and wrapper approaches [20]. On account of feature extraction, another arrangement of features can be made from the underlying set that catches all the critical data in a dataset. The production of new arrangements of features takes into consideration assembling the portrayed advantages of dimensionality decrease.

In any case, the use of feature determination strategies may bring about explicit variances concerning the formation of

prescient element records. A few examinations in the writing talk about the marvel of absence of understanding between the prescient quality records found by various gatherings, the need of thousands of tests so as to accomplish the ideal results, the absence of natural translation of prescient marks and the perils of data release recorded in distributed investigations [21].

The fundamental goal of ML strategies is to deliver a model which can be utilized to perform arrangement, forecast, estimation or some other comparative assignment. The most widely recognized undertaking in learning process is classification. As referenced already, this learning capacity groups the information thing into one of a few predefined classes. At the point when an arrangement model is created, by methods for ML strategies, preparing and speculation blunders can be delivered. The previous alludes to misclassification blunders on the preparation information while the last on the normal mistakes on testing information. A decent characterization model should fit the preparation set well and precisely group every one of the examples. On the off chance that the test mistake paces of a model start to increment despite the fact that the preparation error rates decline then the wonder of model over-fitting happens. This circumstance is identified with model intricacy implying that the preparation mistakes of a model can be decreased if the model unpredictability increments. Clearly, the perfect unpredictability of a model not vulnerable to over-fitting is the one that creates the least speculation mistake. A conventional technique for investigating the normal speculation mistake of a learning calculation is the predisposition fluctuation decay. The inclination segment of a specific learning calculation gauges the blunder pace of that calculation. Also, a second wellspring of mistake over all conceivable preparing sets of given size and all conceivable test sets is called difference of the learning strategy. The general expected mistake of a characterization model is established of the whole of predisposition and change, in particular the bias–variance decomposition.

When a classification model is gotten utilizing at least one ML systems, it is essential to assess the classifier's exhibition. The presentation investigation of each proposed model is estimated as far as affectability, particularity, precision and area under the curve (AUC). Affectability is characterized as the extent of genuine positives that are effectively seen by the classifier, though explicitness is given by the extent of genuine negatives that are accurately distinguished. The quantitative measurements of exactness and AUC are utilized for surveying the general execution of a classifier. In particular, exactness is a measure identified with the complete number of right expectations. In actuality, AUC is a proportion of the model's exhibition which depends on the ROC bend that plots the tradeoffs among affectability and 1-explicitness. Let’s see some of the well known classification algorithms.

Naïve Bayes – This is one of the well-known classification algorithms. It is mainly used when probability prediction belongs to a certain class. It always provides increased accuracy and is one of the fastest algorithms used for train data. Commonly used for large data sets. This is a sequential algorithm that follows the stages of execution, followed by classification, evaluation and forecast. There are various types of data mining algorithms to find the relationship between a normal person and a sick person, but many of the algorithms have their own limitations, such as numerous iterations, long computation time and grouping of continuous arguments, etc. Naive Bayes has overcome several limitations and is one of the best for use in large data sets. Consider odds as factors for class prediction if a set of tests is provided.

Support Vector Machine - It is based on linear classification and it act as a binary classifier. Firstly It was introduced by Vladimir Vapnik and it has shown its effectiveness mainly in the area of pattern recognition problem. Many times it has shown better classification than other classifiers mainly in case of small dataset. Let us discuss how it works, it segregate a pair of training vectors for two dissimilar groups (x_1, y_1), (x_2, y_2),…, (x_m, y_m), where xi €Rd Represent vectors in d-dimensional attribute space and yi € {-1, +1} is a group label.

The following figure shows a linear core SVM procedure that allocates a non-linear input space in a new linearly separable space. It is shown that all vectors that are on one side of the hyperplane are labeled -1, and vectors that are aligned on the other side are labeled + 1. Learning points that are near the hyperplane in the transformed space are considered as reference vectors. Compared to the training set, the size of the support vectors is smaller, these support vectors define the boundaries of the hyperplane and the decision surface.

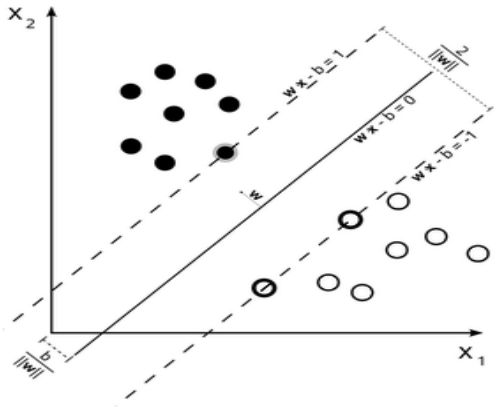

Figure 2: Support Vector Machine

KNN(k nearest neighbors)

K-NN is one of the most popular classification algorithms in machine learning algorithm based on distances. You do not need any training stage, based primarily on cases. The input sample, which we consider as a training set, was combined with the distance function, and the choice or

forecast of a new object depends on how close it is to the given classes. In this case, we take the value of neighbors, it can be taken as n, which is an integer, and, depending on the value of n, new objects are classified into different classes. Before classifying a new object, its distance measurement is taken from another object that belongs to other classes, where the distance between objects is less, a specific object will belong to the class of this previous resident object. Most of the time we use the Euclidean method to measure distances [22]

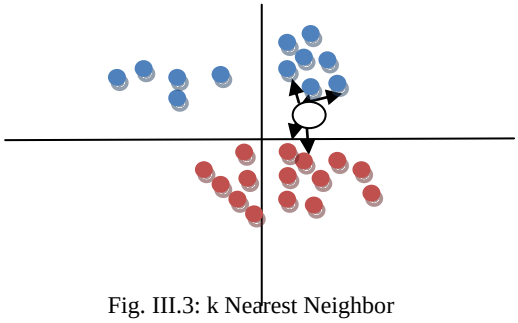

Fig. III.3: k Nearest Neighbor

Logistic regression

Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model (glm). In 1972, Nelder and Wedderburn proposed this model with an effort to provide a means of using linear regression to the problems which were not directly suited for application of linear regression [23]. In fact, they proposed a class of different models (linear regression, ANOVA [24], Poisson Regression [25] etc) which included logistic regression as a special case.

The fundamental equation of generalized linear model is:

$$G(E(y)) = \alpha+\beta x_1 + \gamma x_2 \quad (1)$$

Here, g() is the link function, E(y) is the expectation of target variable and $\alpha + \beta x1 + \gamma x2$ is the linear predictor (α ,β, γ to be predicted). The role of link function is to 'link' the expectation of y to linear predictor.

It is one of the best model for prediction [26] , classification [27] and regression [28]. It is also known as logit method. It is also probabilistic linear classifier [29]. There are many other algorithms like decision tree, artificial neural network which has not discussed in detail in this paper.

IV. PROBLEM STATEMENT

A heterogeneous disorder of many distinct subtypes was characterized by cancer. Throughout cancer research, early diagnosis and forecasts have become a requirement as it can facilitate the ongoing treatment of patients. A number of research teams from the biomedical and bio-informatics industries researched machine learning (ML) technique implementations, due to the importance of classifying cancer patients into high-or low-risk categories. Such methods were therefore used as a blueprint for cancer development and treatment. However, ML methods can be used to classify key characteristics from complex data sets. Although it is observed that some of the cancer datasets are considered to be high dimensional and it is very complex task to process high dimensional dataset. So dimensionality reduction method helps us to reduce the number of features during prediction and classification.

V. CONCLUSION AND FUTURE SCOPE

In this paper we have discuss about the various machine learning algorithms which can be used for cancer analysis prediction with their application. The work also describes the previous work done in the cancer classification using machine learning algorithm. The main focus is on the supervise machine learning algorithms which are able to predict the cancer disease on the basis of certain parameters. It has also been observed that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and classification can provide promising tools for inference in the cancer domain. In future we will discuss the application of deep learning on cancer classification using data based on genes expression.

ACKNOWLEDGMENT

We are very thankful to the management and colleagues of Radharaman Institute of Technology & Science, Bhopal

REFERENCES

[1] R. A. Radcliffe, "Gene expression," in Neurobehavioral Genetics: Methods and Applications, Second Edition, **2006**.
[2] D. Oliva, M. Abd Elaziz, and S. Hinojosa, "Image Processing," in Studies in Computational Intelligence, 2019.
[3] N. Guenther and M. Schonlau, "Support vector machines," Stata J., **2016**.
[4] C. Bulac and A. Bulac, "Decision Trees," in Advanced Solutions in Power Systems: HVDC, FACTS, and AI Techniques, **2016**.
[5] T. G. Clarkson, "Introduction to neural networks," Neural Netw. World, **1996**.
[6] K. R. Koch, Introduction to bayesian statistics. **2007**.
[7] P. J. García-Laencina, J. L. Sancho-Gómez, A. R. Figueiras-Vidal, and M. Verleysen, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation," Neurocomputing, **2009**.
[8] O. Fortunato et al., "Assessment of circulating micrornas in plasma of lung cancer patients," Molecules, **2014**.
[9] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: A multiple random validation strategy," Lancet, **2005**.
[10] Y. C. Chen, W. C. Ke, and H. W. Chiu, "Risk classification of cancer survival using ANN with gene expression data from multiple laboratories," Comput. Biol. Med., **2014**.
[11] K. Park, A. Ali, D. Kim, Y. An, M. Kim, and H. Shin, "Robust predictive model for evaluating breast cancer survivability," Eng. Appl. Artif. Intell., **2013**.
[12] X. Xu, Y. Zhang, L. Zou, M. Wang, and A. Li, "A gene signature for breast cancer prognosis using support vector machine," in 2012 5th International Conference on Biomedical Engineering and Informatics, BMEI 2012, **2012**.

[13] O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau, and B. De Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks," in Bioinformatics, **2006**.
[14] P. Rosado, P. Lequerica-Fernandez, L. Villallain, I. Pena, F. Sanchez-Lasheras, and J. C. De Vicente, "Survival model in oral squamous cell carcinoma based on clinicopathological parameters, molecular markers and support vector machines," Expert Syst. **Appl**., **2013**.
[15] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," Artif. Intell. Med., **2005**.
[16] J. Kim and H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," J. Am. Med. Informatics Assoc., 2013.
[17] "Pattern Recognition and Machine Learning," J. Electron. Imaging, **2007**.
[18] A. Agah, A. Niknejad, and D. Petrovic, "Introduction to Computational Intelligence Techniques and Areas of Their Applications in Medicine," in Medical Applications of Artificial Intelligence, **2013**.
[19] "Introduction to data mining," Intell. Syst. Ref. Libr., **2011**.
[20] R. Nair and A. Bhagat, "A Life Cycle on Processing Large Dataset - LCPL," Int. J. Comput. Appl., **2018**.
[21] Y. Drier and E. Domany, "Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes?," PLoS One, **2011**.
[22] M. Greenacre and R. Primicerio, "Measures of distance between samples: Euclidean," in Multivariate Analysis of Ecological Data, **2013**.
[23] J. A. Nelder and R. W. M. Wedderburn, "Generalized Linear Models," J. R. Stat. Soc. Ser. A J. R. Stat. Soc. Ser. A (General J. R. Stat. Soc. A, **1972**.
[24] Martin, "Two-way ANOVA and ANCOVA," None, 1000.
[25] R. Berk and J. M. MacDonald, "Overdispersion and poisson regression," J. Quant. Criminol., **2008**.
[26] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," CSBJ, **vol. 13, pp. 8–17, 2015**.
[27] E. Alpaydın, Introduction to machine learning, vol. 1107. 2014.
[28] D. P. R. G and R. T. Sriramaneni, "Literature Survey on Various Software Cost," **vol. 4, no. Iv, pp. 868–874, 2016.**
[29] D. Mladenić, J. Brank, M. Grobelnik, and I. Natasa Milic-Frayling, "Feature Selection using Linear Classifier Weights: Interaction with Classification Models," in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, **2004**.

AUTHORS PROFILE

Archana Pathak is M.Tech Scholar from Radharaman Institute of Technology & Science, Bhopal. She has completed her BE from Radharam Engineering College.Her area of research is Machine Learning.

Nidhi Ruthia studied Master of Engineering in CSE at Sagar Institute of Research Technology and Science, Bhopal. She completed her Bachelor of Engineering from Sagar Institute of Research and Technology, Bhopal. Currently, working as Assistant Professor at Radharaman Institute of Technology and Science, Bhopal. Her area of interest in the field of Research is Data mining, Big data, Data Analytics and Artificial Intelligence.

Chetan Agrawal Studied Master of Engineering in CSE at TRUBA Institute of Engineering & Information Technology Bhopal. He has studied his Bachelor of Engineering in CSE at BANSAL Institute of Science & Technology Bhopal. Currently He is working as Assistant professor in CSE department at RADHARAMAN Institute of Technology & Science Bhopal M.P. India. His research area of interest is Social Network Analysis, Data Analytics, Machine Learning, Cyber Security, Network Security, Wireless Networks, and Data Mining.

