


The Prediction of Diabetes: A Machine Learning Approach

Lalit Kumar, Galgotias University, India*

Prashant Johri, Galgotias University, India

 <https://orcid.org/0000-0001-8771-5700>

ABSTRACT

Diabetes is a widely spread disease globally. This issue is a matter of great concern, and the disease is spreading at an alarming rate across the country. We can analyse, visualize the data appropriately, and forecast the chances of having diabetes for a person with the highest level of accuracy and exactness. This indefatigable investigation and paper aims to analyze, compare different neural networks, machine learning algorithms, and classifiers that can predict the probability of disease in patients. The results obtained from the proposed methods are assessed using recollection techniques and making assessments based on exactness of the outputs, which are tested for a number of cases consisting of correct forecasts and wrong forecasts. A thorough study is done on diabetes dataset, and experiments have been carried out using neural networks and several different classifiers.

KEYWORDS

Classifier, Linear Regression, Machine Learning Neural Network, Predictive Analytics, Python

INTRODUCTION

Diabetes is regarded as a very baleful and constantly recurring disease among a set of health-related problems. That is also referred to as 'Diabetes Mellitus, and it is considered one of the foremost reasons for deaths in India. The disease is constantly recurring and occurs when the pancreas does not create sufficient insulin levels. (World Health Organization, 2003) It is also happening in the situation when the affected person is incapable of utilizing the produced insulin. The regulatory function of insulin is to maintain a suitable blood sugar level. The increased sugar level in a person's blood is generally visible and could affect the nervous system and blood cells. Diabetes can also cause other diseases such as blindness, kidney failure, jolting, and cardiovascular disease. Recent research has revealed that approximately 98 million persons might be affected by diabetes in India by 2030 (Weir. & Bonner-Weir, 2004). Hence, there is a need to diagnose and prevent the disease at an early stage.

Diabetes is categorized as Type-1, Type-2, and Gestational diabetes

Type-1 of the disease (diabetes): Type-1 or 'Juvenile category of diabetes' is found when the person's body is incapable of generating a proper insulin level. Since the person suffering from this disease category has a dependency on insulin, it is advised that the person have an insulin intake, which is artificially available (Lee et al., 2011).

DOI: 10.4018/ijrqeh.298630

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

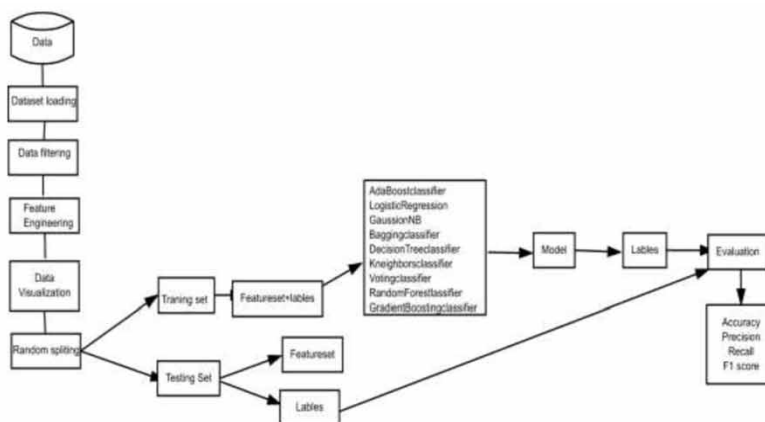
Type-2 of the disease (diabetes): Type 2 and impacts insulin use by the body cells. Although the generation of insulin is satisfactory, the cells cannot respond and adequately use insulin. Most diabetes cases are of the type-2 category (Jayanthi & Babu, 2017).

Gestational category of the disease (diabetes): This disease category is found in females during pregnancy as the body reflects a lower degree of sensitivity to insulin. This disease category is not found in every female and is generally resolved after the delivery.

In this paper, the authors propose a machine learning-based scheme to analyze sample data sets of PIMA to classify the data and forecast the presence or absence of diabetes (Soofi, 2017; Patel, 2017)). The emphasis is to assess the outputs of classification-based frameworks and forecast the chances of occurrence of diabetes in a person with the highest level of accuracy (WHO, 2011; WHO, 2016). Here authors have implemented nine different classification models: AdaBoost Classifier, Logistic Regression, GaussianNB, Bagging Classifier, Decision Tree Classifier, k-NN Classifier, Voting Classifier, Random Forest Classifier, Gradient Boosting Classifier to perfectly examine the dataset (Greenwood et al., 2015).

The architecture of the proposed approach is shown in Figure-1 entitled ‘Diabetic data pre-processing’.

Figure 1. Diabetic data pre-processing



BACKGROUND

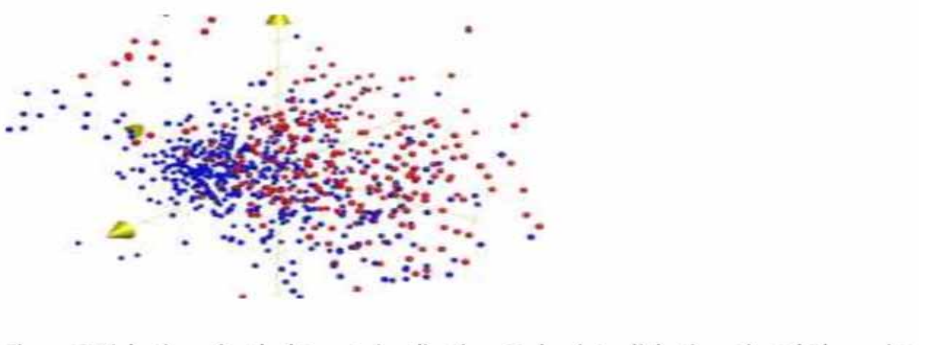
Although numerous machine-learning algorithms are used in this research, several representations and forms capable of forecasting a category of diabetes are evaluated for their exactness. The selected algorithms with the highest level of accuracy are discussed, and their results are compared (Perveen et al., 2019). Currently, designing monitoring models for actual data of patients is getting attraction, and many models are available for use in the health monitoring system (Chauhan et al., 2019).

Further, there is a great need for the online availability of these models. However, a robust design of such a model, a GSM and IoT based structure, is described in (Chauhan et al., 2019). Moreover, the role of controlling glucose levels that help to prevent complications in diabetes is handled in (Vehí et al., 2019); the risks are described, and the application of machine learning models is provided. The application includes an ANN (Artificial Neural Network), SVM & data mining techniques. Jakka & Rani (2019) describe the prediction role of SVM, Random forest and Decision tree classifier.

Besides, as the health monitoring system relies on newer technologies, a comparison of machine learning algorithms in the prediction of diabetes is dealt in (Gupta & Gill, 2020). This scheme uses algorithms like RF, SVM, LDA etc. The use of machine learning techniques and classification tasks are also described in (Subhash & Kumar, 2019), while the methods are compared for their performance. Finally, Yildirim et al. (2019) use a technique based on data mining and machine learning for diabetes prediction.

Using ECG data and cardiac rate information based on deep learning methods is described in (Ignatius et al., 2019), while Greenwood et al. (2015) address the benefits of health monitoring systems (Greenwood et al., 2015). Classification techniques based on machine learning algorithms for diabetes prediction are employed in Kumari et al. (2020). A method for classification of the disease is given in the paper, while a comparative analysis of the accuracy of these methods is provided.

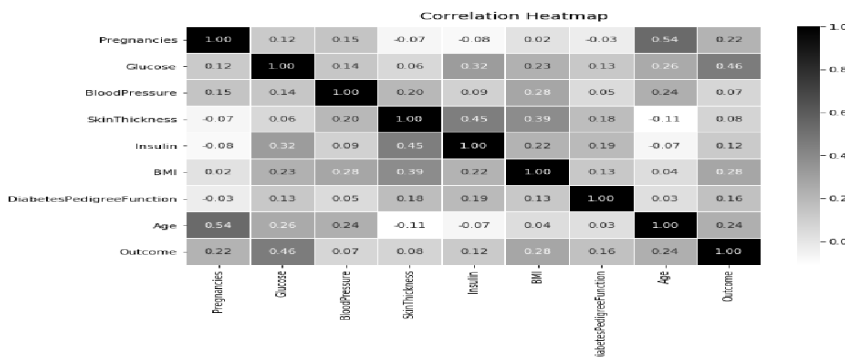
Figure 2. Diabetic patient's dataset visualization. Redpoint - diabetic patients' Blue points -non-diabetic patients



Visualization

For the above-given data to fit best in the underlined algorithms, it is helpful to visualize the data before applying any machine learning tool. Thereby principal component analysis (PCA) is used to

Figure 3. Diabetic data Co-relation



reduce the eight-dimension verse into three dimensions. Moreover, some clusters can be viewed in Figure-2, while exceptional points for diabetic and non-diabetic patients can also be found.

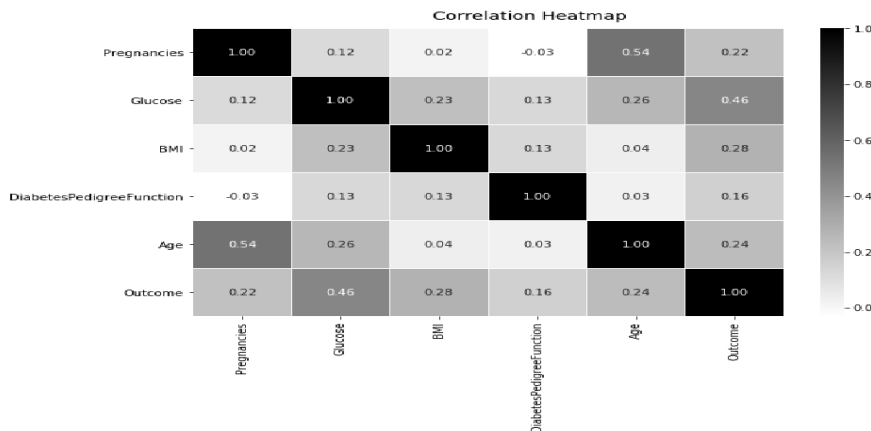
Overall, clustering algorithms may find a hard way to fit the data correctly and predict the best outcome, while an artificial neural network (ANN) can identify the patient's nature. Random-forest and Decision trees might be practical tools to classify the sequential data (for instance, only females can be diagnosed with Gestational diabetes). Such problems might go unrealized while solely working on ANN.

METHODOLOGY

The proposed approach comprises of different steps. Step 1 is data loading; in this step, we load the dataset into Jupyter Lab for manipulation (Bai et al., 2019). In the next step, we clean the dataset not to contain any missing or null values; then, we perform exploratory data analysis (Srivastava et al., 2018; Raspberry, 2018).

Next, we compute a correlation matrix between the data features. Our correlation matrix is shown in Figure-4. Taking insights from the correlation matrix, we drop the low correlation factor with the class label (Wang et al., 2018). We then randomly split this pre-processed data into two sets; training sets and testing sets using 80% records in the training set (Pima Indians Diabetes Database); (Aljumah, 2013). Finally, we apply different machine learning algorithms to understand the data patterns and trends in the diabetes dataset and train the model to make a prediction. Later, we make predictions based on the testing dataset and calculate our machine learning model (Aljumah, 2013).

Figure 4. Diabetic data Co-relation Heatmap



THE PROPOSED MODEL

The proposed model comprises an Artificial Neural Network consisting of five layers; before inputting the data, we have added some polynomial features to increase the accuracy and fit the given data (a problem of overfitting has been considered and taken care of. Normalization and scaling of data were done at this stage (Bai et al., 2019; Capobianco, 2017). A different dropout approach of the neural

network was added with the decrease in probability as we move further into the subsequent layers. The model was allowed to learn from 150 epochs using Adam optimizer.

Performance

The model's performance was evaluated using matrices such as recall, accuracy, precision and F1-score (Lomte et al., 2019). A confusion matrix was used to assess the machine learning algorithms (Zecchin et al., 2012; Lekha. & Suchetha, 2017). It used the following scores or metrics. Finally, the accuracy scored by the machine learning algorithms is summarized in Table-2.

A1 metric was used for representing total cases that are genuinely positive, B1 metric for actual negative points, C1 metric for false-positive cases, and D1 for false-negative patients. The scores or indices were computed using the following metrics:

$$Accuracy = \frac{A1 + B1}{A1 + B1 + C1 + D1} \quad (1)$$

$$Recall = \frac{A1}{A1 + D1} \quad (2)$$

$$Precision = \frac{A1}{A1 + C1} \quad (3)$$

$$F1 - Score = \frac{2 * RM * P1}{RM + P1} \quad (4)$$

The Dataset and Environment

The National Institute of Diabetes collected the PIMA diabetes dataset to study whether a patient has diabetes based on nine features. The dataset includes 2676 patients but has various limitations. All the features of the PIMA dataset, along with their representation, are listed in Table. The diabetes data correlation with eight features are shown in Figure-3

The dataset was split into an 80:20 ratio to form the training and testing—test models like Adaboost Classifier, Logistic Regression, GaussianNB, Bagging Classifier, Decision Tree Classifier, K Neighbor classifier, Voting classifier, Random Forest Classifier & Gradient Boosting Classifier (Lomte et al., 2019; Orabi et al., 2016) were used.

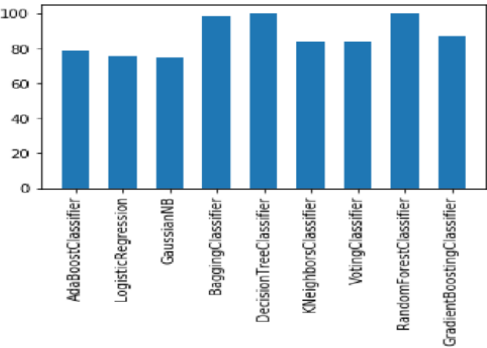
Table1. Diabetic Data Attributes

S. No	Attribute	Representation
1	No of time pregnant	Discrete type of data (int 64)
2	Plasma glucose	Discrete type of data (int 64)
3	BP(mm Hg)	Discrete type of data (int 64)
4	Skin Thickness	Discrete type of data (int 64)
5	Insulin mu U/ml	Discrete type of data (int 64)
6	Body mass index(Weight/Height)(kg/M ²)	Continuous type of data (int 64)
7	Diabetes Pedigree Function	Continuous type of data (int 64)
8	Age	Discrete type of data (int 64)
9	Class	Discrete type of data (int 64)

RESULTS AND ANALYSIS

The machine learning models of AdaBoost Classifier, Logistic Regression, GaussianNB, Bagging Classifier, Decision Tree Classifier, *k*-NN Classifier, Voting Classifier, Random forest Classifier, Gradient Boosting Classifier, trained on the diabetes database, evaluated the efficiency of our classification models on recall precision and accuracy. The authors predicted the model using the test set, and the predicted labels were compared with the actual labels. We found that the neural network performed the best and achieved the highest accuracy of 94.8 per cent. The Decision Tree Classifier also performed well and reached an accuracy of 91.8%. The Bagging Classifier followed with an accuracy of 89.6%. The GaussianNB classifier performed the worst, with an accuracy of just 74.5%. Figure-5 represents the accuracy score achieved by various machine learning algorithms.

Figure 5. Set of classifiers with accuracy



The loss was lowest in the Neural Net, Random Forest Classifier, and Decision Tree Classifier compared to other classification models (Perveen et al., 2018; Kavakiotis et al., 2017).

Table2. Comparison of algorithms with accuracy

Serial No	Machine Learning & Deep Learning Algorithms	Accuracy (%)
1	AdaBoost Classifier	79.6
2	Logistic Regression	77.26
3	GaussianNB	74.5
4	Bagging Classifier	89.6
5	Decision Tree Classifier	91.8
6	k-NN Classifier	84.3
7	Voting Classifier	85.56
8	Random Forest Classifier	94.8
9	Gradient Boosting Classifier	87.91
10	Neural Network	94.8

A comparative study was also conducted on all these classification algorithms using the diabetes dataset. The results showed that the Neural Network, Random Forest Classifier & Decision Tree Classifier are the best performing classification models. Both Random Forest Classifier and Neural Network had a precision score of 1.0 and a recall score of 1.0. Moreover, they were able to classify 371 actual cases of diabetes as positive, and 182 cases of non-diabetic as unfavourable, and one record were classified as false positive out of a total of 554 records in the test set. As a result, the Random Forest Classifier and Neural Net can make accurate predictions and classifications.

CONCLUSION & FUTURE WORK

Diabetes has become a dreadful disease that can be developed at any age and irrespective of gender. Its presence in the human body needs to be detected as soon as possible to prevent life threat. In this paper, a well-structured and organized study is done on various machine learning algorithms and classifiers, and a detailed comparison is made amongst them. The achieved results show that Neural Network's performance was best with the highest accuracy of 94% and with the least Misclassification rate (3.2%) in estimation to more algorithms.

However, the same method might be implemented on other diseases and larger samples, as only a small size of the dataset (2667 instances) was considered. Larger datasets will immensely help the scope of disease prediction and provide the much needed early detection and diagnosis. That way, they might timely help to keep health issues under control and find a way to eliminate them in the future.

REFERENCES

- Aljumah, A. A., Ahamad, M. G., & Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University-Computer and Information Sciences*, 25(2), 127–136.
- Bai, B. M., Nalini, B. M., & Majumdar, J. (2019). Analysis and detection of diabetes using data mining techniques—a big data application in health care. In *Emerging Research in Computing, Information, Communication and Applications* (pp. 443–455). Springer.
- Capobianco, E. (2017). Systems and precision medicine approaches to diabetes heterogeneity: A Big Data perspective. *Clinical and Translational Medicine*, 6(1), 1–10.
- Chauhan, A., Gupta, S. K., & Gupta, R. (2019). Patient Healthcare Monitoring system for Emergency Situations. *International Journal of Innovative Technology and Exploring Engineering*, 8(12S).
- Greenwood, D. A., Blozis, S. A., Young, H. M., Nesbitt, T. S., & Quinn, C. C. (2015). Overcoming clinical inertia: A randomized clinical trial of a telehealth remote monitoring intervention using paired glucose testing in adults with type 2 diabetes. *Journal of Medical Internet Research*, 17(7), e178.
- Gupta, C., & Gill, N. S. (2020). Machine Learning Techniques and Extreme Learning Machine for Early Breast Cancer Prediction. *International Journal of Innovative Technology and Exploring Engineering*, 9(4).
- Ignatius, H., Chandra, R., Bohdan, N., & Dharma, A. (2019). Comparison of Convolutional Neural Network Model in Classification of Diabetic Retinopathy. *Jurnal Penelitian Pos dan Informatika*, 9(2), 141-150.
- Jakka, A., & Rani, V. J. (2019). Performance Evaluation of Machine Learning Models for Diabetes Prediction. *International Journal of Innovative Technology and Exploring Engineering*, 8(11).
- Jayanthi, N., & Babu, B. (2017). Survey on clinical prediction models for diabetes prediction. *Journal of Big Data*, 4, 26.
- Johri, P., Singh, T., Das, S., & Anand, S. (2017, December). Vitality of big data analytics in healthcare department. In *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)* (pp. 669-673). IEEE.
- Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, 15, 104–116.
- Kumari, G. L. A., Padmaja, P., & Suma, J. G. (2020). ENN-Ensemble based Neural Network method for Diabetes Classification. *International Journal of Engineering and Advanced Technology*, 9(3).
- Lee, J., Keam, B., Jang, E., & Park, M. (2011). Development of a predictive model for type 2 diabetes mellitus using genetic and clinical data. *Osong Public Health and Research Perspectives*, 2(2), 75–82.
- Lee, Y. H., Bang, H., & Kim, D. J. (2016). How to establish clinical prediction models. *Endocrinology and Metabolism (Seoul, Korea)*, 31(1), 38.
- Leeflang, P. S., & Wittink, D. R. (2000). Building models for marketing decisions: Past, present and future. *International Journal of Research in Marketing*, 17(2-3), 105–126.
- Lekha, S., & Suchetha, M. (2017). Real-time non-invasive detection and classification of diabetes using modified convolution neural network. *IEEE Journal of Biomedical and Health Informatics*, 22(5), 1630–1636.
- Lomte, R., Dagale, S., Bhosale, S., & Ghodake, S. (2019). Survey of Different Feature Selection Algorithms for Diabetes Mellitus Prediction. In *The 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBE)* (pp. 1-5). Academic Press.
- Orabi, K. M., Kamal, Y. M., & Rabah, T. M. (2016). Early predictive system for diabetes mellitus disease. In *Industrial Conference on Data Mining* (pp. 420-427). Springer.
- Patel, P. B., Shah, P. P., & Patel, H. D. (2017). Analyze Data Mining Algorithms for Prediction of Diabetes. *International Journal of Engineering Development and Research*, 5(3).

- Perveen, S., Shahbaz, M., Keshavjee, K., & Guergachi, A. (2018). Metabolic syndrome and development of diabetes mellitus: Predictive modeling based on machine learning techniques. *IEEE Access: Practical Innovations, Open Solutions*, 7, 1365–1375.
- Pima Indians Diabetes Database Raspberry Pi. (2018). *Raspberry Pi 3 Model B*. Raspberry Pi. Retrieved July 10, 2020, from <https://www.kaggle.com/uciml/pima-indians-diabetes-database>
- Soofi, A. A. (2017). Classification Techniques in Machine Learning: Application and Issues. *Journal of Basic and Applied Sciences*, 13, 459–465.
- Srivastava, S., Singh, M., & Gupta, S. (2018, October). Wireless sensor network: a survey. In *2018 International Conference on Automation and Computational Engineering (ICACE)* (pp. 159-163). IEEE.
- Subhash, A. R., & Kumar, A. (2019). Accuracy of Classification Algorithms for Diabetes Prediction. *International Journal of Engineering and Advanced Technology*, 8(5S).
- Vehí, J., Contreras, I., Oviedo, S., Biagi, L., & Bertachi, A. (2019). Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning. *Health Informatics Journal*, 1–16.
- Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, 126, 3–13.
- Ward, M. O., Grinstein, G., & Keim, D. (2010). *Interactive data visualization: foundations, techniques, and applications*. CRC Press.
- Weir, G. C., & Bonner-Weir, S. (2004). Five stages of evolving beta-cell dysfunction during progression to diabetes. *Diabetes*, 53(suppl 3), S16–S21.
- WHO. (2016). *Definition and diagnosis of diabetes and intermediate hyperglycemia*. Report of a WHO/IDF consultation. WHO Libr. Cat. Data.
- WHO. (2013). *Use of glycated hemoglobin (HbA1c) in the diagnosis of diabetes mellitus. Abbreviated report of a WHO consultation 2011*. WHO.
- World Health Organization. (2003). Diet, nutrition, and the prevention of chronic diseases : report of a Joint WHO/FAO expert consultation. World Health Organization. Retrieved December 12, 2020, from.
- Yildirim, O., Talo, M., Ay, B., Baloglu, U. B., Aydin, G., & Acharya, U. R. (2019). Automated detection of diabetic subject using pre-trained 2D-CNN models with frequency spectrum images extracted from heart rate signals. *Computers in Biology and Medicine*, 113, 103387. doi:10.1016/j.combiomed.2019.103387 PMID:31421276
- Zecchin, C., Facchinetti, A., Sparacino, G., De Nicolao, G., & Cobelli, C. (2012). Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration. *IEEE Transactions on Biomedical Engineering*, 59(6), 1550–1560. doi:10.1109/TBME.2012.2188893 PMID:22374344

Prashant Johri is currently a professor at the School of Computing Science & Engineering, Galgotias University, Greater Noida, India. He completed his B.Sc.(H) from Aligarh Muslim University and M.C.A. from Aligarh Muslim University in 1995 and his Ph.D. in Computer Science in 2011 from Jiwaji University, Gwalior, India. He has also worked as a Professor and Director (M.C.A.) at the Galgotias Institute of Management and Technology, (G.I.M.T.) and worked as a Professor and Director (M.C.A.), Noida Institute of Engineering and Technology, (N.I.E.T.) Gr. Noida. He has served as Chair in many conferences and affiliated as member of program committee in many conferences of India and abroad. He has supervised 2 Ph.D. student and many M.Tech. students for their thesis. He published more than 100 research papers in National and International Journals and Conferences. He has published edited books in Elsevier and Springer. He has also contributed numerous book chapters in the several books published with publishers of high international repute. Apart from scholarly contribution towards scientific community, he organized several Conferences/Workshops/Seminars at the national and international levels. He voluntarily served as reviewer for various International Journals and conferences. His research interest includes Artificial Intelligence, Machine Learning, Data Science, Information Security, Cloud Computing, Block Chain, Healthcare, Agriculture, Image Processing, Software Reliability. He is actively publishing in these areas.