

Prediction of diabetes disease using machine learning algorithms

Monalisa Panda¹, Debani Prashad Mishra¹, Sopa Mousumi Patro¹, Surender Reddy Salkuti²

¹Department of Electrical Engineering, International Institute of Information Technology Bhubaneswar, Bhubaneswar 751029, India

²Department of Railroad and Electrical Engineering, Woosong University, Daejeon 34606, Republic of Korea

Article Info

Article history:

Received Jan 3, 2021

Revised Dec 24, 2021

Accepted Jan 5, 2022

Keywords:

Classification

Diabetes

Gradient boost

K-nearest neighbor

Support vector machine

ABSTRACT

Diabetes mellitus is a powerful chronic disease, which is recognized by lack of capability of our body for metabolization of glucose. Diabetes is one of the most dangerous diseases and a threat to human society, many are becoming its victims and, regardless of the fact that they are trying to keep it from rising more, are unable to come out of it. There are several conventional diabetes disease health monitoring strategies. This disease was examined by machine learning (ML) algorithms in this paper. The goal behind this research is to create an effective model with high precision to predict diabetes. In order to reduce the processing time, K-nearest neighbor algorithm is used. In addition, support vector machine is also introduced to allocate its respective class to each and every sample of data. In building any sort of ML model, feature selection plays a vital role, it is the process where we select the features automatically or manually and it contributes most to our desired performance. Overall, four algorithms are used in this paper to understand which can easily evaluate the total effectiveness and accuracy of predicting whether or not a person will suffer from diabetes.

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



Corresponding Author:

Surender Reddy Salkuti

Department of Railroad and Electrical Engineering, Woosong University

Jayang-dong, Dong-gu, Daejeon-34606, Republic of Korea

Email: surender@wsu.ac.kr

1. INTRODUCTION

Diabetes mellitus is a condition characterized by a metabolic process disorder and an excessive rise in blood sugar concentration due to lack of insulin, a peptide hormone secreted by pancreatic islet beta cells [1], [2]. If we step around the dangerous impact of diabetes, we will undoubtedly conclude that it can lead to significant complications, or even premature death. So, to decrease the mortality rate and improve a patient's health status. Therefore, machine learning algorithms are now used to identify and diagnose diseases in order to minimize the death risk and improve a patient's health status, as machine learning (ML) contributes to specific decisions. There are essentially two main clinical forms, type 1 diabetes and type 2 diabetes[3], which are indicated as (T1D) and (T2D) [4] respectively, according to the origin and progression of the condition, which is nothing but the disorder's etiopathology. Nearly 90% of all diabetic patients have T2D, which is predominantly characterized by insulin hormone resistance. Lifestyle, physical activity, food or dietary patterns and inheritance are the real causes of T2D, T1D is believed to be due to the autoimmune degradation of pancreatic- β cells in the langerhans islets.

Different researchers are designing a multiple diabetes prediction method based on a variety of algorithms. In [5] previously suggested a method for classifying diabetes disease via the use of the support vector machine (SVM). For diagnosis, the Pima Indian Diabetes (PID) dataset is used. Using the radial basis function (RBF) SVM kernel as the classifier, 78% of the accuracy was achieved. Orabiet *al.* [6] designed a

method for the prediction of diabetes. Pradhan and Bamnote [7] presents several genetic programming related algorithms and several tests are performed on this dataset. As a classifier for diabetes disease prediction, in [8] uses J48 decision tree (DT) (74.8% accuracy) and naïve bayes (79.5% accuracy). A prediction model with two sub-modules was developed in [9] to predict diabetes-chronic disease.

In [10] estimates the 250 million individuals are currently affected by diabetes and will cross 500 million by 2025. DT is used to find ways of extracting attributes and features from a fixed dataset [11]. Until testing, the dataset is trained to predict and store the results for each and every new instantiated object in a separate class. In [12] presents an algorithm that classifies the risk of diabetes mellitus [13]. This paper illustrates diabetes disease prediction based on the characteristics of the datasets. Gradient boost is adequate to equate logistic regression (LR), SVM, and k-nearest neighbor (KNN) to the rest of the classifier. Like dataset selection, extraction of attributes, implementation of algorithms after breaking the full dataset into a training and test dataset. Finally, the outcome shown in this paper demonstrates the proposed model's ability to predict diabetes with less time in the earlier process.

Section 1 is all about the introduction regarding Diabetes mellitus i.e., its cause, symptoms, types. Section 2 is about the research method, the model diagram and a brief description about the test and training dataset, the algorithms used for predicting the disease like SVM, KNN, LR, and gradient boost. Section 3 is describing the dataset in a clear way and the figure of outcome indicates the percentage patients with and without the disease. The section 4 depicts the results where we get gradient boost classifier give 81.25%. In the last section, the final conclusion of our processes is described based on our model.

2. RESEARCH METHOD

The required information and necessary steps in order to build the model to predict diabetes by using the classifiers are described as separate sections. The research approach mentioned in this paper explicitly defines the entire model's working criteria. Figure 1 depicts the procedure of proposed approach. The brief discussion of steps involved in the proposed approach are presented next:

- The first step is collection of data Kaggle [14]. This dataset contains all total of 768 instances and 9 attributes [14]. The dataset is briefly discussed in the dataset section.
- The second step is data-preprocessing, in this step the null values are checked and removed also the categorical values converted into numerical ones.
- In the third step exploratory data analysis [15] is performed where each columns correlation matrix was formed and also some visualizations like box plots [15] were done to check the outlier values. Some other visualization [16] was also done to check how [16] the features are related to the label column. Feature selection [17] which plays a major role is also done in this step where the important feature has been selected for the model and after the selection the data has been fit into the model for the prediction [17].
- In this step the dataset was divided into training and testing test [18]. For this work, the dataset has been divided into training and testing part with test size of 0.25. i.e., the training data [19] consists around 75% of the whole data whereas testing data contains 25% of the whole dataset. In the dataset there is a total of 768 instances and 9 attributes, so the training data will contain 576 instances (75% of 768) and testing data will contain 192 instances (25% of 768).
- Algorithms used: There are 4 algorithms used in this paper. Those are LR, KNN, SVM, and gradient boost. These 4 algorithms are discussed next.

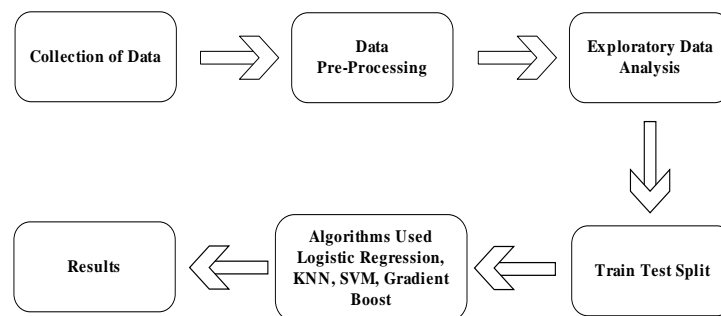


Figure 1. Proposed approach model diagram

2.1. Logistic regression (LR)

Various algorithms are used to decide which is the best match for this dataset and will provide better results. Logistic regression, KNN, SVM, and gradient boost are such algorithms. LR helps us in solving classification problems, it uses S-curve instead of a straight line for fitting the points [20]. Logistic is taken from the function logit that is used in this method of classification[20].

2.2. K-nearest neighbor (KNN)

It has two properties: i) lazy learning algorithm: since there is no separate step of preparation. It utilizes all the data during classification for training and ii) algorithm for non-parametric learning: about the underlying data, it does not assume anything. First of all, the data set is fed as input, including the dataset for testing and training. The loading of the dataset and data preprocessing takes place in the next step. After that decision, to get the desired results, the KNN algorithm is implemented. Steps involved in this algorithm are:

- Using euclidean, manhattan or hamming distances, the distance between the test data and each row of training data is measured.
- Now, arrange them in an ascending order based on the distance values.
- Next the top k rows [21] are picked from the sorted list.
- The class is allocated to the test point based on the frequent classes of these rows [21].

2.3. Support vector machine (SVM)

Support vectors are the most important data points of the training dataset. If we remove the data points then the position of dividing hyper plane will change rather than being 2 non-overlapping classes [22]. And in constant visualization nonlinear separation works well as compare to linear one. Steps involved in this algorithm are: i) import the dataset; ii) explore the data to figure out what they look like [23]; iii) then data is split into attributes and labels [23]; iv) the data is divided into training and testing datasets; and v) SVM algorithm is trained for our desired output or results.

The last step involved in this algorithm is the comparing all the precision [24] of the algorithms to get the results. This performance evaluation is carried out for all algorithms in this phase performance evaluation, the performance of the models has been evaluated using the confusion matrix and classification report where the accuracy, recall, f1-score for each algorithm has been calculated, the comparison between all algorithms is discussed in the results section. The output of a classification model is represented using a confusion matrix. The uncertainty/confusion matrix can be represented by,

$$\text{Confusion Matrix} = \begin{bmatrix} \text{TRUE}^+ & \text{FALSE}^+ \\ \text{FALSE}^- & \text{TRUE}^- \end{bmatrix} \quad (1)$$

True positive (TP): cases in which the classifier predicted TRUE (they have the disease) and TRUE was the correct class (patient has disease). Real negatives (TN): cases where FALSE (no illness) was predicted by the model and FALSE was the right class (patient do not have disease). False positives (FP) (type I error): the classifier predicted TRUE, but FALSE was the correct class (patient did not have disease). False negatives (FN) (type II error): instances where FALSE (patients have no disease) has been predicted by the machine learning model, but they actually have the disease.

Key Performance Indicator (KPI) calculation is as follows:

- Accuracy of classification=(TP+TN)/(TP+TN+FP+FN)
- Misclassification rate=(FP+FN)/(TP+TN+FP+FN)=(error rate)
- Precision=TP/Total TRUE Predictions=TP/(TP+FP) (how much was it accurate when the model predicted the TRUE class?)
- Recall=TP/Real TRUE=TP/(TP+FN) (how much did the classifier get it right when the class was actually TRUE?)

3. DATASET DESCRIPTION

This dataset has been taken from Kaggle [14]. The datasets consist of 768 rows and 9 columns, with 8 rows being instances, and the target variable being 1 row (output). The target variable is outcome, while Predictor variables include the patient's number of births, Triceps skin fold thickness measurement (mm), their body mass index (BMI) (weight in kg/(height in m)²), blood pressure diastolic (mm Hg), insulin level, era, and so on. Figure 2 depicts the instances of outcome column. The outcome column is the target variable and contains zeros and ones where zeros represent that patient don't have diabetes disease whereas one represents patient have diabetes disease. From Figure 2 we can say that there are 500 instances present in the

outcome column are 0 and 268 are 1 that is around 66.10% don't have diabetes and around 34.90% have disease in the dataset.

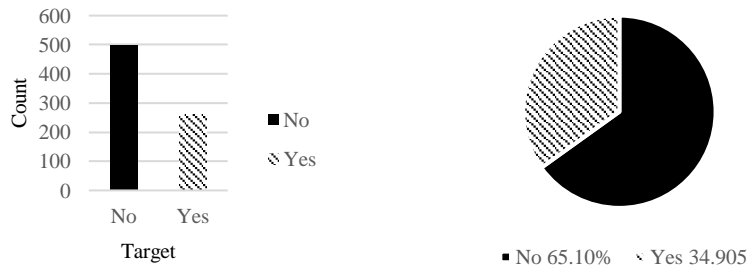


Figure 2. Instances of outcome

4. RESULTS AND DISCUSSION

The final result has been derived successfully using the mentioned four machine learning algorithms. Table 1 shows the accuracy of all the models used in this paper, their precision, recall and f1-score is also shown in Table 1. By comparing all the algorithms, it can be observed that the best algorithm based on accuracy is gradient boost with an accuracy of 81.25%.

Figure 3 depicts the accuracy comparison among the models used for this work. From this figure it can be concluded that Gradient Boost algorithm gives the best accuracy of around 81.25% and KNN gives the lowest accuracy of 78% among these four algorithms, in addition to this logistic regression gives 81% whereas Support vector classifier gives 80% accuracy.

Table 1. Comparison among the models

Name	Accuracy	Precision	Recall	F1-score
Gradient Boost	0.8125	0.7600	0.6290	0.6964
Logistic Regression	0.8073	0.7660	0.5806	0.6605
SVM	0.8021	0.7400	0.5968	0.6607
KNN	0.7813	0.7273	0.5161	0.6038

Model	Accuracy
GradientBoostingClassifier	0.81
LogisticRegression	0.81
SVC	0.8
kNeighborsClassifier	0.78

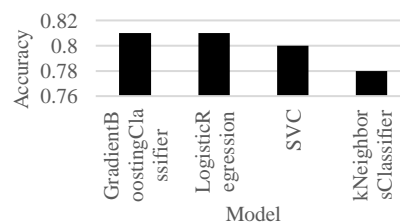


Figure 3. Model accuracy comparison

Figure 4 depicts the precision comparison among the models used for this work. From this figure, it can be concluded that gradient boost algorithm gives the best precision score of around 0.76 and k-neighbors classifiers gives the lowest precision score of 0.73 among these four algorithms, in addition to this LR [25] gives 0.77 whereas support vector classifier gives 0.74 precision score. Figure 5 depicts the precision comparison among the models used for this paper. From this figure, it can be concluded that gradient boost algorithm gives the best recall score of around 0.63 and K-Neighbors classifiers [26] gives the lowest recall score of 0.52 among these four algorithms, in addition to this logistic regression gives 0.58 whereas Support vector classifier gives 0.6 recall score.

Figure 6 depicts the F1-score comparison among the models used for this work. From this figure, it can be concluded that gradient boost algorithm gives the best F1-score [27] of around 0.7 and k-neighbors classifiers gives the lowest F1-score of 0.6 among these four algorithms, in addition to this logistic regression and support vector classifier gives same F1-score of 0.66. From the simulation results, it can be concluded that in this paper there are total 4 algorithms are used LR [28], KNN [29], gradient descent [30], and the best accuracy achieved is 81.25% which has given by gradient descent classifier.

Model	Precision
GradientBoostingClassifier	0.76
LogisticRegression	0.77
SVC	0.74
kNeighborsClassifier	0.73

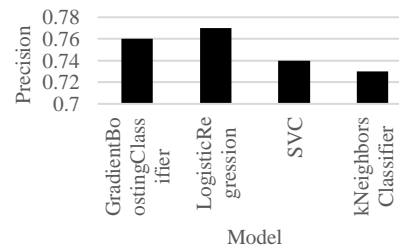


Figure 4. Model precision comparison

Model	Recall
GradientBoostingClassifier	0.61
LogisticRegression	0.58
SVC	0.6
kNeighborsClassifier	0.52

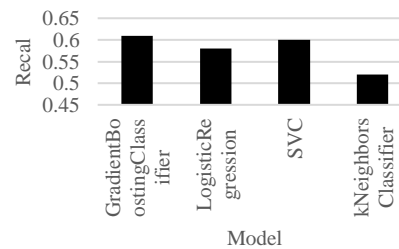


Figure 5. Model recall comparison

Model	F1-Score
GradientBoostingClassifier	0.68
LogisticRegression	0.66
SVC	0.66
kNeighborsClassifier	0.6

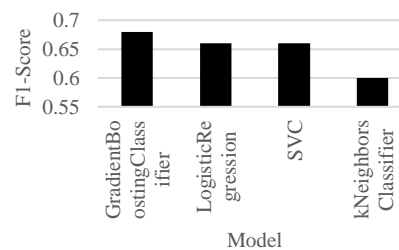


Figure 6. Model F1-score comparison

5. CONCLUSION

Using sophisticated statistical techniques and the availability of a large number of epidemiological and genetic diabetes risk datasets, ML has the considerable potential to restructure or shake up the risk of diabetes prediction. It is clearly seen from this paper that gradient boosting classifier works well for this type of dataset, which is also confirmed by model accuracy and recall. And KNN works well for the dataset includes a large number of datasets that it is easier to minimize processing time. And SVM deals with a wide number of functions for the dataset in a better way. This model can be used for future work, this application can be used by taking patients' past health records and showing whether or not the person has diabetes.

ACKNOWLEDGEMENTS

This research work was funded by "Woosong University's Academic Research Funding - 2022".





REFERENCES

- [1] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics in Medicine Unlocked*, vol. 10, pp. 100–107, 2018, doi: 10.1016/j.imu.2017.12.006.
- [2] J. S. Kaddis, B. J. Olack, J. Sowinski, J. Cravens, J. L. Contreras, and J. C. Niland, "Human pancreatic islets and diabetes research," *Journal of the American Medical Association (JAMA)*, vol. 301, no. 15, pp. 1580–1587, Apr. 2009, doi: 10.1001/jama.2009.482.
- [3] A. B. Olokoba, O. A. Obateru, and L. B. Olokoba, "Type 2 diabetes mellitus: a review of current trends," *Oman Medical Journal*, vol. 27, no. 4, pp. 269–273, Jul. 2012, doi: 10.5001/omj.2012.68.
- [4] T. Zheng *et al.*, "A machine learning-based framework to identify type 2 diabetes through electronic health records," *International Journal of Medical Informatics*, vol. 97, pp. 120–127, Jan. 2017, doi: 10.1016/j.ijmedinf.2016.09.014.




- [5] V. A. Kumari and R. Chitra, "Classification of diabetes disease using support vector machine," *International Journal of Engineering Research and Applications (IJERA)*, vol. 3, no. 2, pp. 1797–1801, 2018.
- [6] K. M. Orabi, Y. M. Kamal, and T. M. Rabah, "Early predictive system for diabetes mellitus disease," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9728, Springer International Publishing, 2016, pp. 420–427.
- [7] M. Pradhan and G. R. Bamnote, "Design of classifier for detection of diabetes mellitus using genetic programming," in *Advances in Intelligent Systems and Computing*, vol. 327, Springer International Publishing, 2015, pp. 763–770.
- [8] A. Iyer, J. S., and R. Sumbaly, "Diagnosis of diabetes using classification mining techniques," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 1, pp. 01–14, Jan. 2015, doi: 10.5121/ijdkp.2015.5101.
- [9] T. A. Rashid, S. M. Abdullah, and R. M. Abdullah, "An intelligent approach for diabetes classification, prediction and description," in *Advances in Intelligent Systems and Computing*, vol. 424, Springer International Publishing, 2016, pp. 323–335.
- [10] P. Samant and R. Agarwal, "Machine learning techniques for medical diagnosis of diabetes using iris images," *Computer Methods and Programs in Biomedicine*, vol. 157, pp. 121–128, Apr. 2018, doi: 10.1016/j.cmpb.2018.01.004.
- [11] N. Yilmaz, O. Inan, and M. S. Uzer, "A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases," *Journal of Medical Systems*, vol. 38, no. 5, May 2014, Art. no. 48, doi: 10.1007/s10916-014-0048-7.
- [12] N. Nai-arun and R. Moungrmai, "Comparison of classifiers for the risk of diabetes prediction," *Procedia Computer Science*, vol. 69, pp. 132–142, 2015, doi: 10.1016/j.procs.2015.10.014.
- [13] K. G. M. M. Alberti and P. Z. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO Consultation," *Diabetic Medicine*, vol. 15, no. 7, pp. 539–553, Jul. 1998, doi: 10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S.
- [14] Kaggle, "Pima Indians diabetes database." <https://www.kaggle.com/uciml/pima-indians-diabetes-database>.
- [15] C. H. Yu, "Exploratory data analysis in the context of data mining and resampling," *International Journal of Psychological Research*, vol. 3, no. 1, pp. 9–22, Jun. 2010, doi: 10.21500/20112084.819.
- [16] M. N. O. Sadiku, A. E. Shadare, S. M. Musa, and C. M. Akujuobi, "Data visualization," *International Journal of Engineering Research And Advanced Technology(IJERAT)*, vol. 2, no. 12, pp. 11–16, 2016, doi: 10.1007/978-1-4020-4409-0_56.
- [17] J. Miao and L. Niu, "A survey on feature selection," *Procedia Computer Science*, vol. 91, pp. 919–926, 2016, doi: 10.1016/j.procs.2016.07.111.
- [18] M. A. Shafique and E. Hato, "Formation of training and testing datasets, for transportation mode identification," *Journal of Traffic and Logistics Engineering*, vol. 3, no. 1, 2015, doi: 10.12720/jtle.3.1.77-80.
- [19] R. Medar, V. S. Rajpurohit, and B. Rashmi, "Impact of training and testing data splits on accuracy of time series forecasting in machine learning," in *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, Aug. 2017, pp. 1–6, doi: 10.1109/ICCUBEA.2017.8463779.
- [20] J. L. Alzen, L. S. Langdon, and V. K. Otero, "A logistic regression investigation of the relationship between the learning assistant model and failure rates in introductory STEM courses," *International Journal of STEM Education*, vol. 5, no. 1, Dec. 2018, Art. no. 56, doi: 10.1186/s40594-018-0152-1.
- [21] P. Ray and D. P. Mishra, "Support vector machine based fault classification and location of a long transmission line," *Engineering Science and Technology, an International Journal*, vol. 19, no. 3, pp. 1368–1380, Sep. 2016, doi: 10.1016/j.jestch.2016.04.001.
- [22] D. P. Mishra and P. Ray, "Fault detection, location and classification of a transmission line," *Neural Computing and Applications*, vol. 30, no. 5, pp. 1377–1424, Sep. 2018, doi: 10.1007/s00521-017-3295-y.
- [23] P. Ray, D. P. Mishra, and G. K. Budumuru, "Location of the fault in TCSC-based transmission line using SVR," in *2016 International Conference on Information Technology (ICIT)*, Dec. 2016, pp. 270–274, doi: 10.1109/ICIT.2016.061.
- [24] M. Arora, U. Kanjilal, and D. Varshney, "Evaluation of information retrieval: precision and recall," *International Journal of Indian Culture and Business Management*, vol. 12, no. 2, pp. 224–236, 2016, doi: 10.1504/ijicbm.2016.074482.
- [25] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.
- [26] K. Vizhi and A. Dash, "Diabetes prediction using machine learning," *International Journal of Advanced Science and Technology*, vol. 29, no. 6, pp. 2842–2852, May 2020, doi: 10.32628/cseit2173107.
- [27] J. Liu *et al.*, "Artificial intelligence in the 21st century," *IEEE Access*, vol. 6, pp. 34403–34421, 2018, doi: 10.1109/ACCESS.2018.2819688.
- [28] C. Y. J. Peng, T. S. H. So, F. K. Stage, and E. P. St. John, "The use and interpretation of logistic regression in higher education journals," *Research in Higher Education*, vol. 43, no. 3, pp. 259–293, 2002, doi: 10.1023/A:1014858517172.
- [29] Y. Cai, D. Ji, and D. Cai, "A KNN research paper classification method based on shared nearest neighbor," in *Proceedings of NTCIR-8 Workshop Meeting*, 2010, pp. 336–340.
- [30] D. Yi, S. Ji, and S. Bu, "An enhanced optimization scheme based on gradient descent methods for machine learning," *Symmetry*, vol. 11, no. 7, Jul. 2019, Art. no. 942, doi: 10.3390/sym11070942.

BIOGRAPHIES OF AUTHORS






Monalisa Panda     received the Bachelor of Technology (B.Tech.) degree in Electrical and Electronics Engineering from International Institute of Information Technology, Bhubaneswar, Odisha, India 2021. She has worked many Machine Learning Companies she has published many blogs in Medium. She is currently working as a Software Developer in Mindtree Ltd, Bangalore in Dotnet Core Technology. Her research areas of interest include Machine Learning, Data Analytics Big Data, Database Systems, Neural Networks, MongoDB, Hadoop Framework, Spark. She can be contacted at email: monalisapanda94@gmail.com.






Debani Prasad Mishra    received the B.Tech. in electrical engineering from the Biju Patnaik University of Technology, Odisha, India, in 2006 and the M.Tech in power systems from IIT, Delhi, India in 2010. He has been awarded the Ph.D. degree in power systems from Veer Surendra Sai University of Technology, Odisha, India, in 2019. He is currently serving as Assistant Professor in the Dept of Electrical Engg, International Institute of Information Technology Bhubaneswar, Odisha. He has 11 years of teaching experience and 2 years of industry experience in the thermal power plant. He is the author of more than 80 research articles. His research interests include soft Computing techniques application in power systems, signal processing and power quality. 3 students have been awarded Ph.D. under his guidance and currently 4 Ph.D. Scholars are continuing under him. He can be contacted at email: debani@iiit-bh.ac.in.



Sopa Mousumi Patro    has received the Bachelor of Technology (B.Tech.) degree in Electrical and Electronics Engineering from International Institute of Information Technology, Bhubaneswar, Odisha, India in 2021. She is currently working as a System Engineer at Infosys Limited, Mysore. Her research areas of interest include Machine Learning, Data Analytics Big Data, Database Systems, Neural Networks, MongoDB, Hadoop Framework. She can be contacted at email: smousumipatro330@gmail.com.



Surender Reddy Salkuti    received the Ph.D. degree in electrical engineering from the Indian Institute of Technology, New Delhi, India, in 2013. He was a Postdoctoral Researcher with Howard University, Washington, DC, USA, from 2013 to 2014. He is currently an Associate Professor with the Department of Railroad and Electrical Engineering, Woosong University, Daejeon, South Korea. His current research interests include power system restructuring issues, ancillary service pricing, real and reactive power pricing, congestion management, and market clearing, including renewable energy sources, demand response, smart grid development with integration of wind and solar photovoltaic energy sources, artificial intelligence applications in power systems, and power system analysis and optimization. He can be contacted at email: surender@wsu.ac.kr.