Augmented Human Research (2023) 8:1
https://doi.org/10.1007/s41133-023-00062-4

ORIGINAL PAPER

Timely Prediction of Diabetes by Means of Machine Learning Practices

Rajan Prasad Tripathi[1] · Manvinder Sharma[2] · Anuj Kumar Gupta[2] · Digvijay Pandey[3] · Binay Kumar Pandey[4] · Aakifa Shahul[5] · A. S. Hovan George[6]

Received: 24 February 2022 / Revised: 24 April 2023 / Accepted: 6 November 2023 / Published online: 9 December 2023

Abstract
The quality and quantity of medical data produced by digital devices have improved significantly in recent decades. This has led to cheap and easy data generation. There has therefore been an increased advantage in the areas of Big Data and machine learning. There is a huge application of machine leaning and artificial intelligence in health care sector. The use of machine learning to train the machine to classify the medical cases taking care of the historical data can be a boon in medical studies. In this paper, we have analyzed many machine learning algorithms and classifiers which are used to make prediction on the diabetes based on the chosen features and attributes of the dataset. The implementation of the algorithms and its performance are compared in terms of accuracy; we have also used the soft voting ensemble techniques and applied the standardized PIMA diabetes data for which the highest accuracy is achieved.

Keywords Machine learning · Diabetes · Ensemble · Soft voting · Data science

Introduction

The top ten causes of death in 2016 include diabetes. In 2016, 1.6 million people were affected by diabetes, up from fewer than 1,000,000 in 2000. HIV/AIDS was the seventh leading cause of death as shown in Fig. 1. Diabetes figures grew from the number of diabetes people in the 1980s of 108 million to 422 million in 2014; global diabetes rose from 4.7% in 1980 to 8.5% in 2014 for adults aged over 18.

By 2040, diabetes is projected to be present in 642 million people (1 in 10 people). In addition, 46.5% of diabetes patients were not diagnosed [1]. It is important to develop strategies and procedures that aid early diagnosis of diabetes, since many deaths of diabetic patients are due to late diagnosis, to reduce diabetes-related deaths.

We need advanced information technology to achieve state-of-the-art technologies for early diagnostics of diabetes, and the data mining sector is an important area for it. Data mining provides the ability to extract from a broad database repository and discover previously unknown, secret, yet interesting models. Such trends can help to [2] diagnose and determine medically.

Diabetes mellitus is one of the diseases that affect a very large human population and is often called diabetes mellitus. Diabetes [3], a very large amount, affected more than 425 million people in 2017. In the same year, about 4 million people died of diabetes and associated complications. Though 74 million people in India have suffered from diabetes, India is recognized as the "World Capital for Diabetes". If this disease has not been taken seriously and there are no major steps to diagnose and prevent it, an estimated 629 million people worldwide will be affected by diabetes by 2045 [4].

Diabetes is a high blood glucose condition that is caused if the body cannot make the required quantity of the insulin or the body is unable to use the insulin that is produced effectively. Diabetes is most commonly caused by obesity, urbanization, physics inactivity, unhealthy diet, aging and diabetes family history. When diabetes is not rightly

✉ Digvijay Pandey
digit11011989@gmail.com

1 Amity University Tashkent, Tashkent, Uzbekistan

2 Chandigarh Group of Colleges, Chandigarh, Landran, India

3 Department of Technical Education, IET, Dr. A.P.J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh, India

4 Department of Information Technology, College of Technology, Govind Ballabh Pant University of Agriculture and Technology, Pantnagar, Uttrakhand, India

5 SRM Medical College, Kattankulathur, Tamil Nadu, India

6 Tbilisi State Medical University, Tbilisi, Georgia

```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Pregnancies	2000.0	3.70350	3.306063	0.000	1.000	3.000	6.000	17.00
Glucose	2000.0	121.18250	32.068636	0.000	99.000	117.000	141.000	199.00
BloodPressure	2000.0	69.14550	19.188315	0.000	63.500	72.000	80.000	122.00
SkinThickness	2000.0	20.93500	16.103243	0.000	0.000	23.000	32.000	110.00
Insulin	2000.0	80.25400	111.180534	0.000	0.000	40.000	130.000	744.00
BMI	2000.0	32.19300	8.149901	0.000	27.375	32.300	36.800	80.60
DiabetesPedigreeFunction	2000.0	0.47093	0.323553	0.078	0.244	0.376	0.624	2.42
Age	2000.0	33.09050	11.786423	21.000	24.000	29.000	40.000	81.00
Outcome	2000.0	0.34200	0.474498	0.000	0.000	0.000	1.000	1.00

Fig. 1 Description of dataset

diagnosed or managed properly, it can cause many complications, such as cardiovascular problems, kidney diseases, blindness, and neural complications such as stroke [5]. Early diagnosis is the most important fact for effective diabetes management and related complications. Early diagnosis and the recommended daily healthy lifestyle are the most important factor [6].

Literature Review

When you open trans_jour.docx, select “Page Layout” from the.

The following describes some of the various methods used on PIMA Indian Diabetes Datasets with their results.

Rohan Bansal et al. used diabetes diagnosis KNN classifier; the attributes are selected using the PSO techniques. This method has proven to be 77 percent accurate [7]

In the case of the normalization and unconventional KNN algorithm model, i.e. the KNN class-specific classification algorithm, the preprocessing of the dataset is proposed as class-wise KNN (CKNN) methodology for diabetes classification. The accuracy of this process is 78.16% [7].

Lin Li et al. proposed one of the techniques known as weight-adjusted voting classification commonly known. This method is predictive of the accuracy of 77 percent following implementation of PIMA's Indian diabetes dataset [8].

The principle of modified extreme learning machines was used by Priyadarshini et al. to determine whether or not the patient is diabetic dependent or not on the available data. In neural networks and extreme classifier learning, the authors draw comparative conclusions.

Prema NS et al. [9] proposed to use ensemble technique on normalized PIMA Indian diabetes dataset and got efficiency of 81%.

In its analysis, Iyer [10] indicated that a forecast for diabetes should be made with the use of the Naïve Bayes algorithm. The study reported a 79.56 percent accuracy result. Throughout the classification of diabetic patients, Tarun [11] used a PCA and a support vector machine. Experimental tests have shown that while their accuracy is 93.66 percent, the previous amount can be enhanced. Kadhmi [12] suggested that, after applying a nearest K algorithm to the elimination of unwanted data, the decision tree (DT) be used to assign every data sample to its corresponding class. Han et al. [13] developed a model that uses the algorithm for the prediction of diabetes using the K-means algorithm. The model attained a 95.42% accuracy [14].

In Ref. [15], *k*-mean clustering was used for defining and removing outliers, genetic algorithm and CFS for the related extraction of characteristics, as well as for the classification of diabetic patients by *k*-nearest neighbor (KNN). Patil [16] has proposed a hybrid model of forecasting which applied *k*-means to the original dataset and then used C4.5 algorithms to construct the model for the classifier. The result was 92.38% classification precision. Anjali [17] proposed to reduce the dimension of the extracted features with neural network (NN) as a classification technique dependent upon principal component analysis. The accuracy result was 92.2% [18].

Methodology

PIMA Indian Diabetes Dataset

A list of different datasets is available for the research and implementation of ML algorithms in the UCI Machine Learning Repository. The data have been very regularly used as a primary source of machine learning datasets by researchers, students and educators. We took the PIMA Diabetes Dataset [15] for our study from this repository. This dataset is made up of 768 patients ' medical data.

There are eight attributes in each data point, and they are:

- Number of times pregnant
- Plasma glucose concentration
- Diastolic blood pressure
- Triceps skin fold thickness
- Body mass index
- 2-h serum insulin
- Diabetes pedigree function
- Age

The 9th attribute of each data point is the class variable. The outcome will be either 0 or 1 for positive or negative diabetes.

Data Cleaning

The data when found to have many missing values, these missing values create a lot of problems in the analysis, and when we train the model with the help of original dataset, having these missing values will not give good result and hence the missing values have to be taken care of; there are many methods available for cleaning the data like replacing the whole row or deleting the complete row but that would result in less number of training data which we don't want and hence we have used the mean method; we have replaced all the missing data with the mean of the values taken from other values, and hence, it has given the same kind of values and we can process further with the pipeline [16].

Algorithm:Baseline

- We normally provide training and testing results. Only at the end of the measurement and the final performance assessment should we reach the test range. Then we can set the train to train and check settings. We use the validation dataset to tune the model [17].
- High variance test issues with conventional train testing process. It means that by changing the test set the result of the prediction changes. We use the k-fold validation method in our train and validation set to solve this problem [18].
- We analyzed the data; after that, we visualized the data to understand the data more better; we plotted a pair plot and found out there were lot of outliers in the data [19]. We investigated each feature distribution and checked its skewness and kurtosis. We followed this step with feature engineering which includes the following.

Data Preprocessing

Numerical features preprocessing is different for tree and non-tree model. Usually, tree-based models do not depend on scaling. Non-tree-based models hugely depend on scaling. Most often used preprocessing are: MinMax scaler to [0,1], Standard scaler to mean = 0, and Std = 1. Then we removed the outliers.

Feature Selection

Feature selection means that we will have to select those variables or features which will give very high dependency on our target variable which is diabetes is there or not in our case. In our data the features or the attributes are automatically selected using the feature selection; the most relevant to the prediction of our test case variable will be taken up.

Feature selection methods allow you to build a predictive model in our task. It allows us to choose those feature which will give very high dependency on the target class [20].

All the redundant and irrelevant features or the columns are deleted as they can have adverse effect on the prediction accuracy.

Models and Chosen Hyperparameters

A. Logistic Regression (https://www.kaggle.com/pouryaayria/a-complete-ml-pipeline-tutorial-acu-86#5.1.Logistic-Regression)
- C: Regularization value, the more, the stronger the regularization (double).
- Regularization type: can be either "L2" or "L1". Default is "L2".

B. KNN
- n_neighbors: Number of neighbors to use by default for k_neighbors queries

C. SVC (https://www.kaggle.com/pouryaayria/a-complete-ml-pipeline-tutorial-acu-86#5.3.-SVC)
- C: The penalty parameter C of the error term.
- Kernel: Kernel type could be linear, poly, rbf or sigmoid.

D. Decision Tree (https://www.kaggle.com/pouryaayria/a-complete-ml-pipeline-tutorial-acu-86#5.4.-Decision-Tree)
- max_depth: Maximum depth of the tree (double).
- row_subsample: Proportion of observations to consider (double).
- max_features: Proportion of columns (features) to consider in each level (double).

E. AdaBoostClassifier (https://www.kaggle.com/pouryaayria/a-complete-ml-pipeline-tutorial-acu-86#5.5-AdaBoostClassifier)
- learning_rate: Learning rate shrinks the contribution of each classifier by learning_rate.
- n_estimators: Number of trees to build.

F. GradientBoosting
- learning_rate: Learning rate shrinks the contribution of each classifier by learning_rate.
- n_estimators: Number of trees to build.

Ensemble Methods

Ensemble is a technique of machine learning which combines multiple machine learning techniques in one optimal predictive model. Reduce variance, bias or improve predictions [20]. This approach makes it possible to improve predictive performance

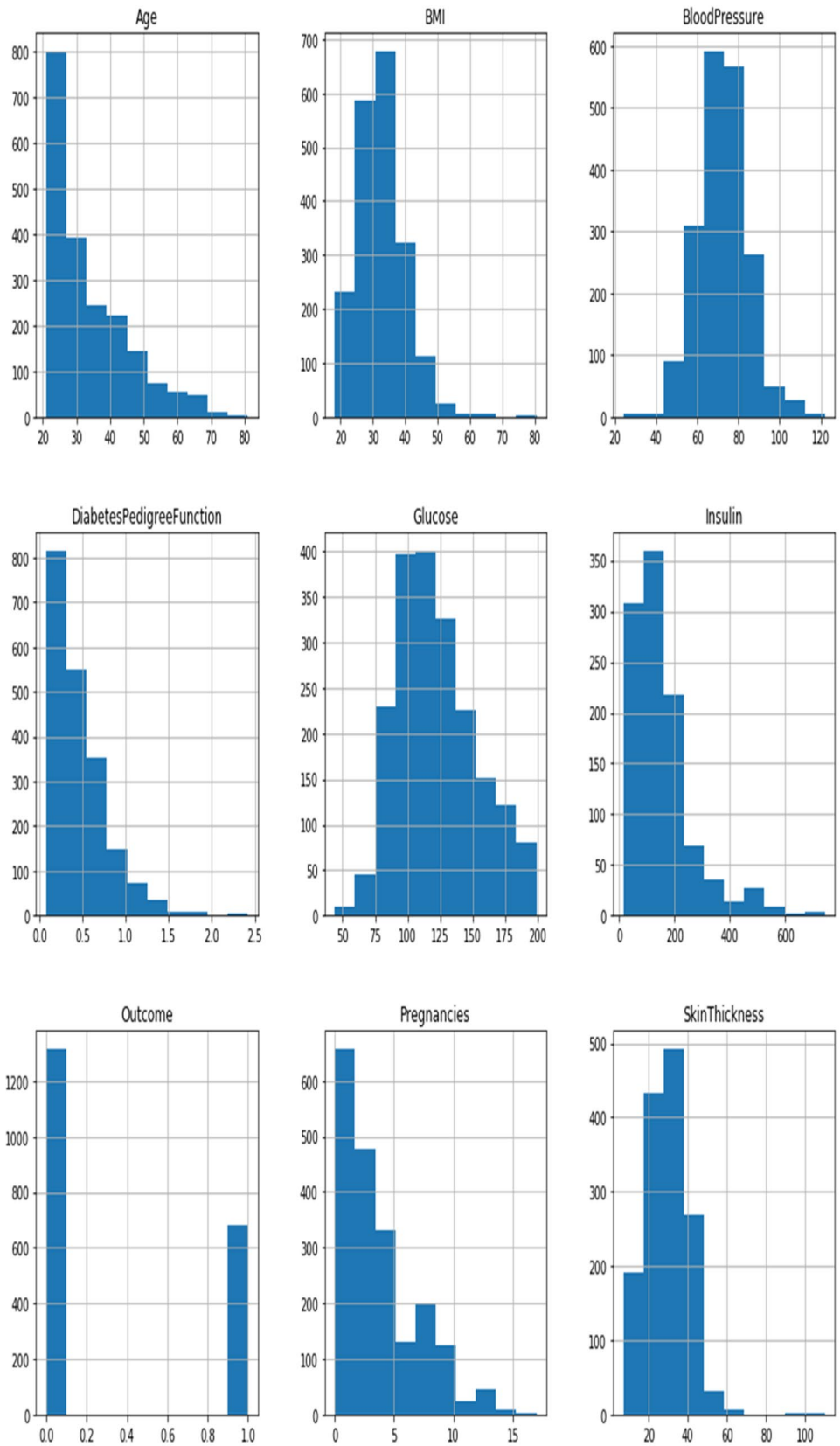

Fig. 2 Histogram of features

```
df_copy['Glucose'].fillna(df_copy['Glucose'].mean(),inplace=True)
df_copy['BloodPressure'].fillna(df_copy['BloodPressure'].mean(),inplace=True)
df_copy['SkinThickness'].fillna(df_copy['SkinThickness'].mean(),inplace=True)
df_copy['Insulin'].fillna(df_copy['Insulin'].mean(),inplace=True)
df_copy['BMI'].fillna(df_copy['BMI'].mean(),inplace=True)
```

```
df_copy.isnull().sum()
```

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Fig. 3 Data cleaning

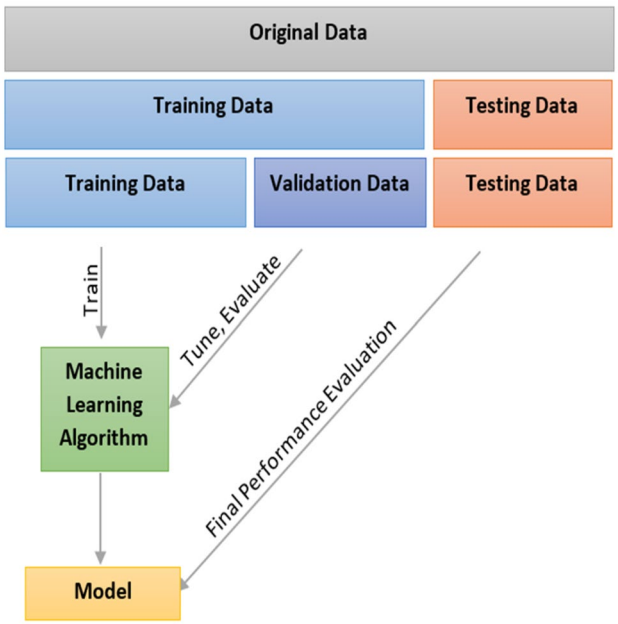

Fig. 4 The blueprint of algorithm

when compared to a single model. There are various methods of ensembling such as bagging, boosting, adaboosting, stacking, voting, averaging, etc. We have applied voting-based ensembling method on PIMA Indian diabetes dataset. The ensemble vote classifier is a meta-classifier which combines similar or conceptually different machine learning classifiers for classification through majority or plurality voting.

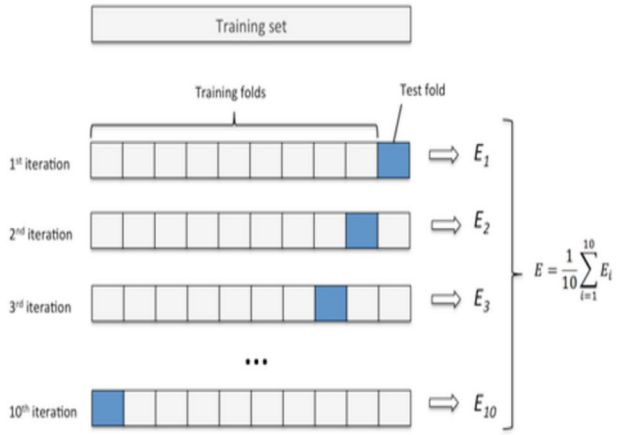

Fig. 5 Cross-validation

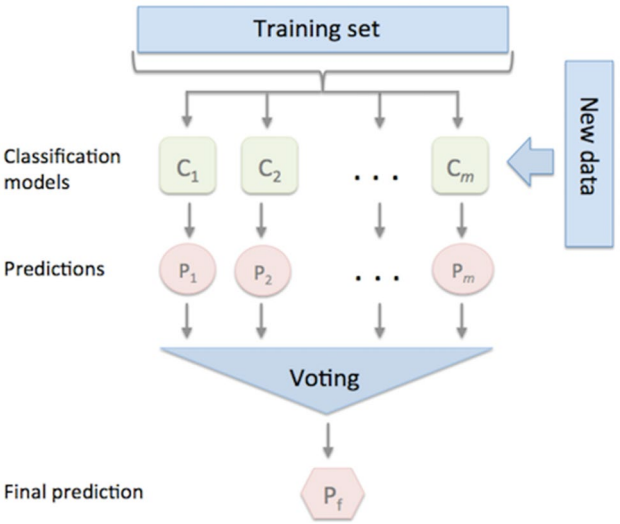

Fig. 6 Flow of ensemble method

Table 1 Various models with accuracy

Model	Accuracy	Parameters
Logistic regression	84.3	$C = 0.76$, penalty = l1
KNN	82.8	'n_neighbors': 15
SVC	84.3	'C': 1.7, 'kernel': 'linear'
Decision tree	76.5	criterion': 'gini', 'max_depth': 3, 'max_features': 2, 'min_samples_leaf': 2
AdaBoostClassifier	81.2	'learning_rate': 0.05, 'n_estimators': 150
GradientBoosting	81.2	'learning_rate': 0.01, 'n_estimators': 100
Ensemble method	82.8	Soft voting

Voting Classifier Using Python Library Scikit learn

A voting classifier is a ML model that forms on a collection of various models and forecasts an output on the basis of its highest probability of the selected class.

We pass the findings of each classifier, and our voting classifier sums all of them and predicts the output class based on the highest majority of the vote. The idea is that instead of creating different dedicated models and calculating the accuracy for each of them, we create a single model that trains all the specified machine learning model [21]; these models predict output based on their cumulative majority voting for each output class (Figs. 2, 3, 4, 5, 6).

Two Types of Votes are Supported by Voting Classifier

Hard voting: The expected performance class in hard polling is a class which is most likely to be expected by each classifier, with the most number of votes. Suppose the output class (*A*, *A*, *B*) is foreseen by three classifiers, so that most predicted A as output. A is therefore the ultimate forecast.

Soft voting: The prediction in soft voting is based on the average probability given to this class. Assume the likelihood for class $A = (0.40, 0.57, 0.63)$ and $B = (0.30, 0.42, 0.50)$ given some inputs to three models. The average is 0.5333 for class *A* and 0.4067 for class *B*. The winner is clearly class *A*.

In soft voting, class label is predicted on the predicted probabilities p for classifier [22].

$$y\hat{} \arg\max i \sum j = 1mwjpij,$$

where *wj* is the weight that can be assigned to the *j*th classifier.

We assume as per our figure a binary classification task with class labels $i \in \{0,1\}$; our ensemble could make the following prediction:

$$C1(x) \rightarrow [0.8, 0.2]$$

$$C2(x) \rightarrow [0.7, 0.3]$$

$$C3(x) \rightarrow [0.3, 0.7]$$

Using uniform weights, we compute the average probabilities:

$$p(i0|x) = (0.8 + 0.7 + 0.3)/3 = 0.6$$

$$p(i1|x) = (0.2 + 0.3 + 0.7)/3 = 0.4$$

$$y\hat{} \arg\max i[p(i0|x), p(i1|x)] = 0$$

[12]

Result

We have applied different classification techniques for PIMA Indian diabetes; the results are shown in Table 1. The data are sent to the classifier by dividing the data into 30% testing and 70% training, the accuracy of various models using cross-validation technique is shown in Table 1, and the comparative analysis is shown in Fig. 1 as well.

Conclusion

Diabetes prediction is done using various machine learning model and classifier; we have also used ensemble voting with a group Indian diabetes dataset for PIMA classifiers compared to highest consistency with different classification algorithms. We have used cross-validation on dataset with tenfold CV data which were distributed into 30% tests and 70% training. Logistic regression performed surprisingly very well 84.3% and by using ensemble voting classifier with default soft voting the accuracy came out to be 82.8%.

Acknowledgements I would like to thank the DTE.

Authors' Contributions All authors contributed.

Funding No funding.

Availability of Data and Materials The datasets used/or analyzed during the current study is available from the corresponding author on reasonable request.

Declarations

Conflict of interest There is no conflict of interest.

Consent for Publication "Not applicable".

References

1. http://www.who.int/news-room/fact-sheets/detail/diabetes Accessed 27 July 2018
2. IDF diabetes atlas-8th edition (2017) International Diabetes Federation, 2017. Available online https://diabetesatlas.org/. Accessed 15 Dec 2018
3. https://www.diabetesdaily.com/learn-about-diabetes/what-is-diabetes/how-many-people-have-diabetes/
4. Jhaldiyal T, Mishra PK (2014) Analysis and prediction of diabetes mellitus using PCA, REP and SVM. Int J Eng Technol Res (IJETR) 2(8) ISSN: 2321-0869.
5. Prabhu P et al (2011) Improving the performance of K-means clustering for high dimensional data set. Int J Comput Sci Eng 3(6):2317
6. Anjali Khandegar, Khushbu Pawar (2017) diagnosis of diabetes mellitus using PCA, neural network and cultural algorithm. Int J Digital Appl Contemp Res 5(6)

7. Kaur N, Sharma M (2017) Brain tumor detection using self-adaptive K-means clustering. In: 2017 International conference on energy, communication, data analytics and soft computing (ICECDS), pp 1861–1865. IEEE
8. Motka R, Parmarl V, Kumar B, Verma AR (2013) Diabetes mellitus forecast using different data mining techniques. In: IEEE 4th international conference on computer and communication technology (ICCCT), IEEE (2013), pp 99–103
9. Global Report on Diabetes WHO Library Cataloguing-in-Publication Data Global report on diabetes. 2016
10. Pandey BK, Mane D, Nassa VKK, Pandey D, Dutta S, Ventayen RJM, Rastogi R (2021) Secure text extraction from complex degraded images by applying steganography and deep learning. Multidisciplinary approach to modern digital steganography. IGI Global, pp 146–163
11. Kaur SP, Sharma M (2015) Radially optimized zone-divided energy-aware wireless sensor networks (WSN) protocol using BA (bat algorithm). IETE J Res 61(2):170–179
12. Madhumathy P, Pandey D (2022) Deep learning based photo acoustic imaging for non-invasive imaging. Multimed Tools Appl 81(5):7501–7518
13. PIMA Indian diabetes dataset, An open dataset (2019) UCI machine learning repository. Available online http://ftp.ics.uci.edu/pub/machine-learningdatabases/pima-indians-diabetes/. Accessed 11 Jan 2019
14. Bansal R, Kumar S, Mahajan A (2017) Diagnosis of diabetes mellitus using PSO and KNN classifier. In: 2017 international conference on computing and communication technologies for smart nation (IC3TSN), 2017, pp 32–38
15. Lelisho ME, Pandey D, Alemu BD, Pandey BK, Tareke SA (2023) The negative impact of social media during COVID-19 pandemic. Trends Psychol 31(1):123–142
16. Li L (2014) Diagnosis of diabetes using a weight-adjusted voting approach. In: 2014 IEEE international conference on bioinformatics and bioengineering, pp 320–324
17. Pandey BK, Pandey D, Wairya S, Agarwal G, Dadeech P, Dogiwal SR, Pramanik S (2022) Application of integrated steganography and image compressing techniques for confidential information transmission. Cyber Secur Netw Secur 169–191
18. Kotsiantis SB, Kanellopoulos D, Pintelas PE (2007) Data preprocessing for supervised leaning. World Acad Sci Eng Technol Int J Comput Electr Autom Control Inf Eng 1(12):4091–4096
19. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N (2017) Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 15:104–116
20. Ali R et al (2014) Prediction of diabetes mellitus based on boosting ensemble modeling. In: International conference on ubiquitous computing and ambient intelligence, part of the lecture notes in computer science book series. LNCS. vol 8867. Springer
21. Sharma M, Sharma B, Gupta AK, Pandey D (2023) Recent developments of image processing to improve explosive detection methodologies and spectroscopic imaging techniques for explosive and drug detection. Multimed Tool Appl 82(5):6849–6865
22. Goyal S, Pandey D, Singh H, Singh J, Kakkar R, Srinivasu PN (2022) Mathematical modelling for prediction of spread of corona virus and artificial intelligence/machine learning-based technique to detect COVID-19 via smartphone sensors. Int J Mode Identif Control 41(1–2):43–52