A DNN Model for Diabetes Mellitus Prediction on PIMA Dataset

Ovass Shafi Zargar[1]
Avinash Baghat[2]
Tawseef Ahmed Teli[3*]

[1,2]School of Computer Applications
Lovely Professional University
Punjab, India
[3*]Department of Higher Education
UT of Jammu and Kashmir, India
[1]owaisfour03@gmail.com
[2]avinash.baghat@lpu.co.in
[3*]mtawseef805@gmail.com

Abstract. There are several deadly diseases that have affected the common man around the globe irrespective of region and the most common among them is diabetes mellitus (DM). To predict the onset of the disease, deep learning can play a vital role. However, the data fed to the deep learning algorithm need to be free from various anomalies like outliers, missing values and inappropriate attributes etc. To enhance the data and make it adequately appropriate for the decision-making process, various pre-processing techniques are available. In this paper, a Deep Neural Network (DNN) model is developed and the effect of pre-processing techniques is shown by comparing the results of applying various pre-processing techniques to handle missing values to the PIMA dataset. The results show that the model performs better using Mean while applying Minmax and Robust scaler techniques.

Keywords: Diabetes, prediction, CNN, RNN, LSTM.

(Received November 1st,2022 / Accepted December 1st, 2022)

1 Introduction

From the year 1980 onwards diabetes mellitus has become one of the diseases that affect people of all age groups. This disease is caused when the pancreas does not work properly and an inadequate amount of insulin is secreted or when the body does not respond to insulin normally. Both cases result in an abnormally high level of glucose in the blood. Prediabetes is another condition in which the glucose level within the blood is not high enough, but it is higher than normal and an indication of the onset of diabetes in a person. A person is pre-diabetic when the range of glucose contained in blood is 100 mg/dL and 125 mg/dL during fasting. A pre-diabetes person can reduce the chances of diabetes through dieting and proper exercise. Among the patients suffering from diabetes, only 50 to 10% suffer from Type-I kind of diabetes. When the body develops resistance to insulin effect and hence does not respond to insulin, the patient is said to have developed Type-II diabetes. This is the most common diabetes that people suffer from. Diabetes affects the metabolism of the body. It does not affect a single organ but multiple organs. Diabetes narrows blood vessels and restricts blood flow and results in multiple complications like Brain Stroke, Blindness, Heart Attack, kidney malfunctioning, Nerve Disorder etc.

The traditional mechanism for diagnosis of Diabetes mellitus is measuring glucose levels in the blood using BGM and Oral Glucose Tolerance tests. With the new and recent advancements in the field of AI, ML and DL [1–3], enough efforts have been made for early diagnosis and prediction of diabetes mellitus. With the huge volume of data available in the medical industry, data quality's effect in extracting useful information becomes more important. During the application of data mining techniques on medical datasets, it should be kept in mind that these attributes are subjected to problems of missing values, noise and outliers. The reason for these problems can be unavailable information during data entry, computer-related errors, human errors, and missing relevant data due to misunderstanding or equipment malfunction. Enhancing data quality and improving the performance of prediction and diagnosis methods could be a direct consequence of the application of pre-processing techniques on datasets which could lead to accurate results.

2 Literature Review

In a research study [4], authors used many ML algorithms on PIMA to diagnose diabetes. They propose that better accuracy would require including omics data for the prediction of diabetes. The authors in [5] provide a review and analysis of the diagnosis and detection of diabetes mellitus from six different points. They further provide insight into the three challenges in the field of Diabetes mellitus, in particular detection and diagnosis. The researchers in [6] propose a method that uses breath as the sample for the detection of diabetes instead of blood or urine samples. They read the acetone level in breath and then apply the deep learning algorithm viz convolutional neural network along with support vector machine for calculating the automated features from the raw signal and then classifies the derived features. Authors in [7] carried out a research study to develop a novel diabetes detection framework using Conv-LSTM. They analysed the performance of their framework by applying the model to the PIMA dataset. For the feature extraction from the dataset, they used the Boruta algorithm and then used hyperparameter optimisation by applying the GSA to get the optimized features for the model. The developed model (Conv-LSTM) gives an accuracy of 91.38%. After the application of the cross-validation technique, the accuracy of the model improved to 97.26% and thus outperformed other deep learning models. In a research study [8], the authors give a review of the applicability of deep learning in the detection of diabetes mellitus. From the review, the author documents the importance of various deep learning methods in diabetes detection like CNN, RNN, and HAN. They highlighted the possibility to meet the challenges of early detection of diabetes with the advancement in DL techniques and the availability of a huge volume of data from the medical industry. The authors in [9] carried out a research study that describes the modern artificial intelligence techniques suitable for the diagnosis of Diabetes retinopathy that have already been described in the literature. They compare the performance of various artificial intelligence algorithms and emphasise future research in the field to address multiple challenges like medico-legal implications, clinical deployment models and ethics. Authors in [10] carried out a research study comparing various machine learning prediction models with commonly used regression models used for predicting Type II DM. In a research study [11] develops a Deep Human Leukocyte Antigen DL model for ascribing HLA genotypes. The developed model achieves better performance in terms of accuracy having a higher performance for low-frequency. The authors in [12] proposed a mechanism for designing a model using a DNN for early diagnosis of diabetes by training the attributes of the PIMA dataset with an accuracy of 98.3%. With the application of the ten-fold validation the accuracy was 97.11% with 96.25% sensitivity, and specificity of 98.08%. The work [13] proposed a framework for the learning of feature structures across sources that are the same. The authors applied the proposed model to glucose forecasting using a convolutional neural network. The results obtained show that the accuracy of the model could be enhanced by using an adversarial method to train in such a way that over-performs the existing deep learning technologies. The proposed model performs well when used with multiple datasets. The authors in [14] propose a diabetes prediction system using the dropout method and also addresses the problem of overfitting. Using the DNN, they devised a system that outperforms conventional neural network methods. The authors in [15] proposed a diabetes prediction solution called the Smart Diabetes Diagnosis System using machine learning, huge medical datasets and a large cloud of health intelligence to provide a more focused risk assessment and personal treatment schedule. The proposed system also provides patients with daily guidance for the improvement of medication. In [16], researchers investigated the diagnosis of diabetes mellitus and studied the importance of HbA1c and FPG used as input parameters. The authors [17] carried out a research study to identify the problem of automatic prediction of diabetes and provide the solution for the same using a deep learning algorithm and support vector machine for better therapeutic management. The authors achieved an accuracy of 65.1042% accuracy. Other ML and DL-based applications include [18–27].

3 Deep Learning

Deep Learning is a smaller set of AI used for data processing, and object detection as well as in a myriad of application areas including decision-making. With the help of deep learning, computers can learn without the intervention of humans. With the advancement in the digital era, there is an explosion of data in various forms and from various regions. This huge data most commonly known as big data originates from multiple sources like social networks, search engines, online entertaining agencies, e-commerce etc. the big data is easily available and sharable through fintech applications like cloud computing. Unstructured data is so huge in volume that it could take hundreds of years for humans to understand the patterns in it and to extract useful information so that it can be used for decision-making. Researchers realize the potential of AI systems, ML and DL for processing such huge volumes of data and extracting relevant information for making smart decisions. Fig. 1 provides the classification of DL architectures based on supervised and unsupervised learning.

3.1 DL Architecture

The underlying architecture behind DL is an Artificial Neural Network based on which multiple variations of the algorithms have been devised. Deep Learning is not a new concept however it shows tremendous growth in the last few decades due to the intersection of deeply layered neural networks and the availability of highly powerful GPUs that enable fast execution of Big Data.

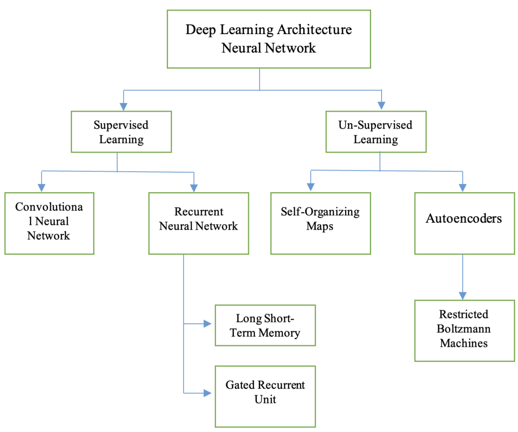


Figure 1: DL Classification

4 Dataset and DNN Model

This section provides insights into the dataset and the DNN-based model that was used in this study Firstly, the significance and types of data pre-processing are provided.

4.1 Data Pre-Processing

Data is the main resource for data analysts to make reliable and accurate decision-making. Without data, policymakers and decision-makers can't make accurate decisions. Nowadays huge volume of data is available and more and more data is generated on an everyday basis by every organization. With the ever-increasing amounts of data, it is impossible to process the data manually by humans and unravel significant information buried in the data. To process the huge repository of data, we need high-configuration computational resources. However, the data produced within the organization is not always in a form that can be directly fetched into the computer for processing. There are a number of anomalies present in the data like missing values, inappropriate values, outliers etc. if the data is analyzed without proper screening it may produce misleading results. Therefore, to make data appropriate to be used by machine learning algorithms, it needs to be pre-processed appropriately so that the results obtained from data analysis will be more accurate and reliable. Data preprocessing helps to check the quality of the data. The main techniques employed are as under:

4.1.1 Data Cleaning

It is the process of removing data from the dataset that is incomplete, incorrect or inaccurate. The various reasons for using data cleaning are the presence of missing values: The Missing values in the dataset can be replaced by mean value, median value or most probable value. Presence of Noisy data: Noisy data can be handled by techniques like Binning, Regression or clustering.

4.1.2 Data Integration

The process of combination of multiple data sources into a single dataset is known as data integration. Various issues that need to be considered during data integration include Schema Integration, Entity identification Problems and Detecting and resolving data values.

4.1.3 Data Reduction

Using data reduction the volume of data is reduced by removing irrelevant data from the dataset without affecting the quality of the dataset. This process helps in making data processing tasks less time-consuming. Various techniques applied for data reduction are Dimensionality, Numerosity reduction and Data Compression.

4.1.4 Data Transformation

Data transformation changes the format of the data to make the new data simple, easy to interpret and easy to process. Some of the methods used for data transformation are smoothing, aggregation, Discretization and Normalization.

4.2 Dataset

The PIMA dataset has been used for this study. This dataset is used for the prediction of diabetes using ML algorithms. A total of eight attributes with 768 records are present in the dataset. There are 500 records whose value for outcome attribute is 0 (non-diabetic) and 268 records have a value of 1(diabetic).

4.3 Model

A deep neural network with the architecture, shown in Fig. 2 was used to diagnose the onset of DM using the eight attributes and one outcome.

Layer (type)	Output Shape	Param #
dense_4 (Dense)	(None, 8)	72
dense_5 (Dense)	(None, 16)	144
dense_6 (Dense)	(None, 8)	136
dense_7 (Dense)	(None, 1)	9

Total params: 361
Trainable params: 361
Non-trainable params: 0

Figure 2: Model Architecture

The total number of the trainable parameter is 833 on the network containing a total of three layers. The deep neural network calculates and examines the weights of different attributes that are related to regressions of the network. Finding the best parameters in terms of having significance with respect to predicting accurately reasons to have a function known as loss function Y(θ) which is used to find the displacement of actual values from predicted values. The goal is to minimise this function. We use back-and-forward propagation to achieve this.

$$Y(\theta) = \frac{1}{k}\sum_{p=1}^{k} L(n_i^{\theta}, n_i) \tag{1}$$

where n_i^{θ} is neural network p is training set size, θ confers to model parameters, L cost function

The process of updating the parameters using backpropagation:

$$\theta = G_d(\theta) \tag{2}$$

5 Results

The above-discussed model was implemented using a jupyter notebook.

The total number of the trainable parameter is 833 on the network containing a total of three layers. The missing values were handled using the following four techniques:

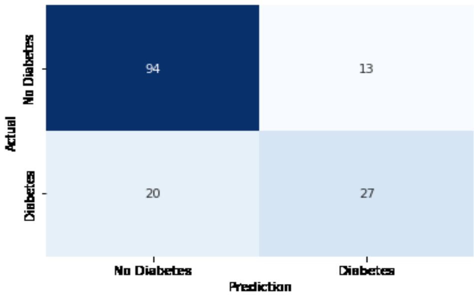
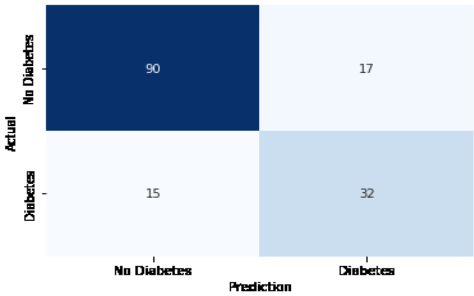
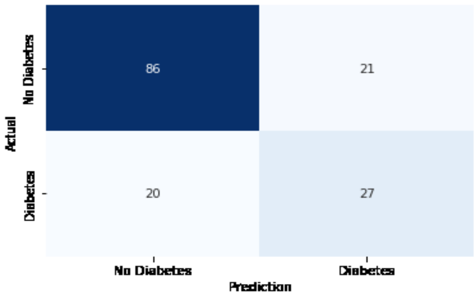
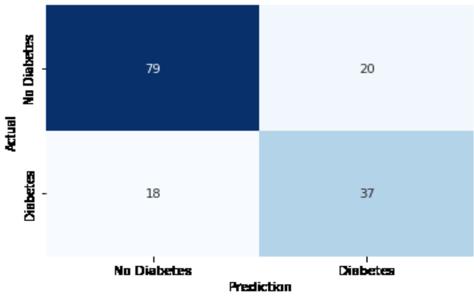
1. Mean: Missing values are replaced with mean. The mean is calculated by adding the total data points over the total number. It gives an average value of all data points.
2. Median: To handle the missing values, the median is used. The median separates the highest half from the lowest half of the data set. It is the middlemost value in data items.
3. Most Frequent: The most prominent in the entire column is used to handle the missing values. The most frequent also known as mode searches for the value that occurs most in the entire column.
4. Constant: The missing values are replaced by some constant value.

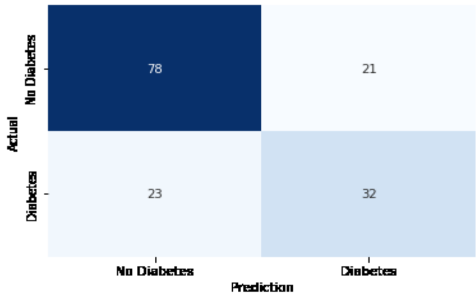
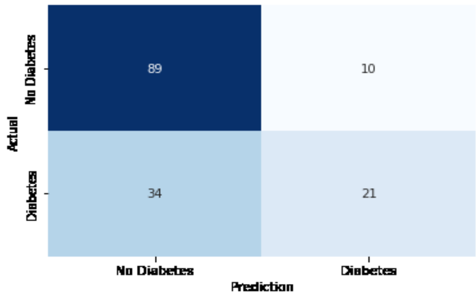
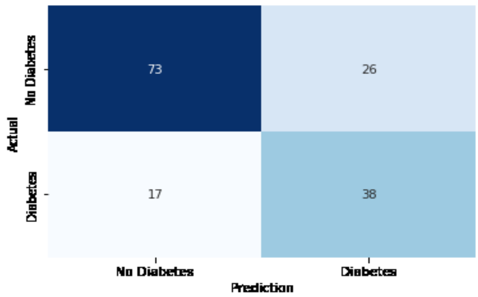
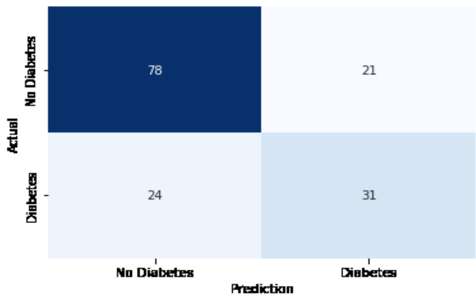
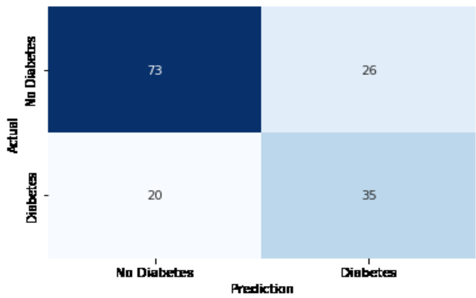
The standardisation techniques used in this experiment are as under:

1. MixMax scaler (MMS): The MinMax Scaler technique changes each value within a range of 0 and 1. Having column c, the function can be defined as follows:

$$diff\left[c\right] = \frac{\left(diff\left[c\right] - diff\left[c\right].min\left(\right)\right)}{diff\left[c\right].max\left(\right) - diff\left[c\right].min\left(\right)} \tag{3}$$

2. Standard Scaler (SS): It standardises an attribute by taking out the mean and then changing it to unit variance.

Missing Value Technique	Standardisation	Confusion Matrix
Mean	MMS	
	SS	
	RS	
Median	MMS	

	SS	
	RS	
Most Frequent	MMS	
	SS	
	RS	

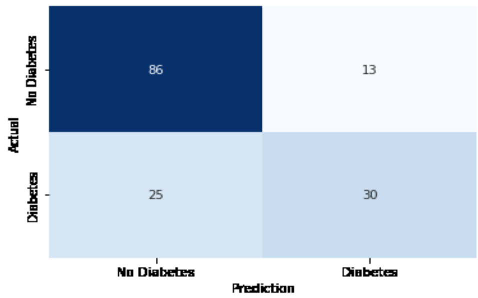
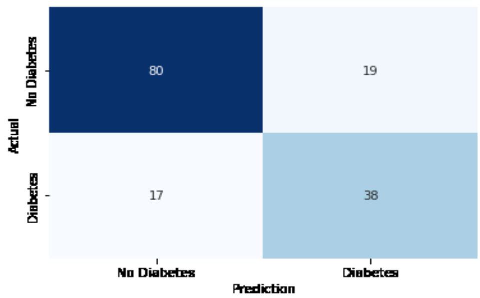
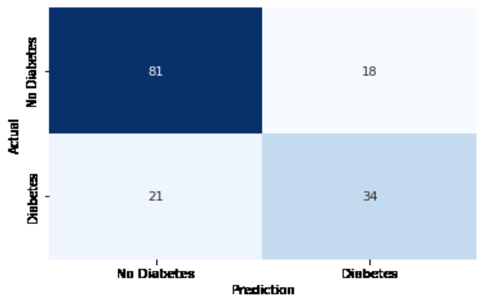
Constant	MMS	
	SS	
	RS	

Table 1: Confusion Matrices

3. Robust Scaler (RS): Robust scaler algorithms scale attributes that are robust to outliers. It uses the interquartile.

The confusion matrix obtained after implementing each technique on the PIMA dataset is enumerated in Table 1:

It may be noted that these preprocessing techniques have been employed to enhance the overall accuracy of the model. The confusion matrices give a good insight into the True Positive, False Positive, True Negative and False Negative samples during the testing phase. Mix-Max and StandardScaler give a lesser number of False Negatives for all missing value techniques and subsequently perform well than the others. This directly affects the recall and hence the Total Negative Rate (TNR). From these confusion matrices, the accuracy is calculated as summarized in Fig. 3 given below.

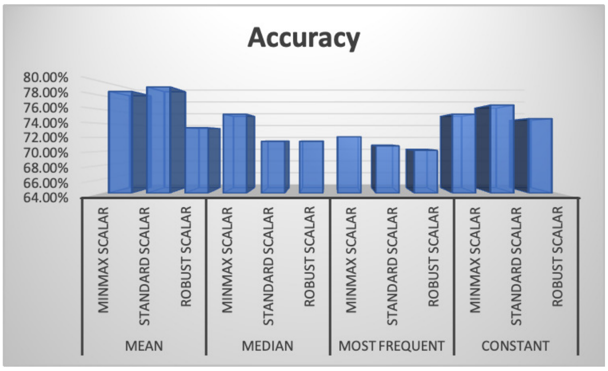


Figure 3: Accuracy

It is clear that the accuracy levels reached using min-max and standard scaler using mean are way better than any of the other techniques applied. Constant provides second best results while the median and most frequent perform the same.

6 Discussion and Conclusion

The simulations for the preprocessing and its effects on the accuracy were done using the PIMA dataset. It contains information about women only and it could very well be established that the dataset considers only gestational diabetes only. The dataset contains features that are skewed which means that the normalized data form is required. The dataset contains several missing or null entries which have been handled using the trivial techniques mean, median and most frequent etc. It is pertinent to mention here that these techniques employ some sort of bias in the dataset. Despite the limitations, the mean method integrated with the standardization technique of the MinMax scaler performs better than any other method. In future, a detailed and hybrid deep learning-based model shall be trained that encompasses techniques like missing values imputations etc for better results and accuracy.

References

[1] Ogobuchi, O. D., Vieira, S. T., Saadi, M., Rosa, R. L., and Rodríguez, D. Z. Intelligent network planning tool for location optimization of unmanned aerial vehicle base stations using geographical images. *Journal of Electronic Imaging*, 31(6):061822, 2022.

[2] Teodoro, A. A., Silva, D. H., Rosa, R. L., Saadi, M., Wuttisittikulkij, L., Mumtaz, R. A., and Rodríguez, D. Z. A skin cancer classification approach using gan and roi-based attention mechanism. *Journal of Signal Processing Systems*, pages 1–14, 2022.

[3] Teodoro, A. A., Silva, D. H., Saadi, M., Okey, O. D., Rosa, R. L., Otaibi, S. A., and Rodríguez, D. Z. An analysis of image features extracted by cnns to design classification models for covid-19 and non-covid-19. *Journal of signal processing systems*, pages 1–13, 2021.

[4] Naz, H. and Ahuja, S. Deep learning approach for diabetes prediction using pima indian dataset. *Journal of Diabetes and Metabolic Disorders*, 19:391–403, 6 2020.

[5] Chaki, J., Ganesh, S. T., Cidham, S. K., and Theertan, S. A. Machine learning and artificial intelligence based diabetes mellitus detection and self-management: A systematic review, 6 2022.

[6] Parte, R. S., Patil, A., Patil, A., Kad, A., and Kharat, S. Non-invasive method for diabetes detection using cnn and svm classifier. *International Journal of Scientific Research and Engineering Development*, 3.

[7] Rahman, M., Islam, D., Mukti, R. J., and Saha, I. A deep learning approach based on convolutional lstm for detecting diabetes. *Computational Biology and Chemistry*, 88, 10 2020.

[8] Zhu, T., Li, K., Herrero, P., and Georgiou, P. Deep learning for diabetes: A systematic review, 7 2021.

[9] Grzybowski, A., Brona, P., Lim, G., Ruamviboonsuk, P., Tan, G. S., Abramoff, M., and Ting, D. S. Artificial intelligence for diabetic retinopathy screening: a review, 3 2020.

[10] Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., and Stiglic, G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, 10, 12 2020.

[11] Naito, T., Suzuki, K., Hirata, J., Kamatani, Y., Matsuda, K., Toda, T., and Okada, Y. A deep learning method for hla imputation and transethnic mhc fine-mapping of type 1 diabetes. *Nature Communications*, 12, 12 2021.

[12] Ayon, S. I. and Islam, M. M. Diabetes prediction: A deep learning approach. *International Journal of Information Engineering and Electronic Business*, 11:21–27, 3 2019.

[13] Bois, M. D., Yacoubi, M. A. E., and Ammi, M. Adversarial multi-source transfer learning in healthcare: Application to glucose prediction for diabetic people. *Computer Methods and Programs in Biomedicine*, 199, 2 2021.

[14] Ashiquzzaman, A., Tushar, A. K., Islam, M. R., Shon, D., Im, K., Park, J. H., Lim, D. S., and Kim, J. Reduction of overfitting in diabetes prediction using deep learning neural network. volume 449, pages 35–43. Springer Verlag, 2017.

[15] Latchoumi, T. P., Dayanika, J., and Archana, G. A comparative study of machine learning algorithms using quick-witted diabetic prevention, 2021.

[16] Ahmad, H. F., Mukhtar, H., Alaqail, H., Seliaman, M., and Alhumam, A. Investigating health-related features and their impact on the prediction of diabetes using machine learning. *Applied Sciences (Switzerland)*, 11:1–18, 2 2021.

[17] Thaiyalnayaki, K. Classification of diabetes using deep learning and svm techniques. *International Journal of Current Research and Review*, 13:146–149, 1 2021.

[18] Yousuf, R., Jeelani, Z., Khan, D. A., Bhat, O., and Teli, T. A. Consensus algorithms in blockchain-based cryptocurrencies. Institute of Electrical and Electronics Engineers Inc., 2021.

[19] Teli, T. A. and Masoodi, F. Blockchain in healthcare: Challenges and opportunities. *SSRN Electronic Journal*, 2021.

[20] Teli, T. A. and Wani, M. A. A fuzzy based local minima avoidance path planning in autonomous robots. *International Journal of Information Technology (Singapore)*, 13:33–40, 2 2021.

[21] Teli, T. A., Yousuf, R., and Khan, D. A. Manet routing protocols, attacks and mitigation techniques: A review, 2022.

[22] Teli, T., Masoodi, F., and Bahmdi, A. *HIBE: Hierarchical Identity-Based Encryption*. 2021.

[23] Teli, T. A., Masoodi, F., and Yousuf, R. Security concerns and privacy preservation in blockchain based iot systems: Opportunities and challenges, 2020.

[24] Teli, T. A., Yousuf, R., and Khan, D. A. Ensuring secure data sharing in iot domains using blockchain, 2 2022.

[25] Jahangeer, S. S., Ovass, S., Majid, Z., and Ahmed, T. T. Big data and deep learning in healthcare, 2022.

[26] Shafi, O., Sidiq, J. S., Teli, T. A., and Kashmir. Effect of pre-processing techniques in predicting diabetes mellitus with focus on artificial neural network, 2022.

[27] Shafi, O., Sidiq, S. J., Teli, T. A., and Zaman, M. A comparative study on various data mining techniques for early prediction of diabetes mellitus, 11 2021.