[1]Ovass Shafi Zargar

[2]Avinash Bhagat

[3]Tawseef Ahmed Teli

A Deep Learning-Based Diabetes Diagnosis Model on PIMA Image Dataset

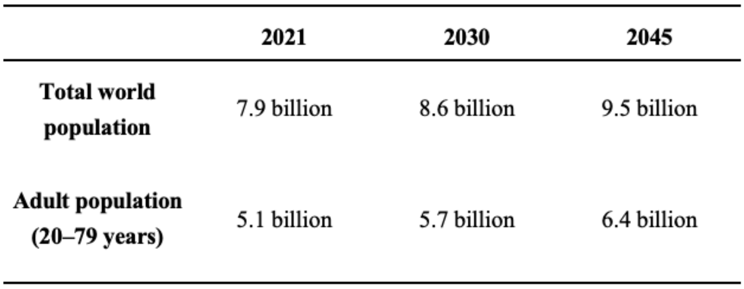
Journal of Electrical Systems

Abstract: - Deep learning is a highly useful technique for the early identification of diabetes mellitus, according to the study done by numerous authors over the past few decades. By using pre-processing techniques on the dataset to get rid of various anomalies like over-fitting, under-fitting, redundancy, missing values, and non-significant features to make it more efficient for analysis, it is possible to increase the effectiveness of deep learning algorithms for diagnosing the disease. This work addresses the global problem of diabetes by exploring a revolutionary deep-learning method for early identification. Conventional convolutional neural network (CNN) models have drawbacks when used with numerical medical datasets, like this study's PIMA Indians Diabetes Database. The article suggests a technique for transforming numerical data into visual representations depending on feature relevance to get over this obstacle. This conversion makes it possible to use strong CNN models for diabetes early diagnosis. Classifying the created diabetic images after feeding them into CNN architectures that have already been trained on VGG16 and ResNet50. The promising outcomes with an accuracy of 97.19% demonstrate the possibility of the suggested strategy for improving diabetes detection and validating the effectiveness of diabetes imaging in obtaining an early diagnosis.

Keywords: Chronic, Diabetes, Glucose, Mellitus, PIMA

I. INTRODUCTION

No matter the person's age, gender, or location, diabetes mellitus is a devastating illness that has affected people all over the world. One type of sickness that affects multiple organs is diabetes. Diabetes mellitus can lead to many consequences, including heart disease, obesity, stress, blindness, and stroke. Due to inadequate hygienic practices and poor health circumstances, the diabetes problem is particularly severe in developing nations like India, Bangladesh, Pakistan, Sri Lanka, and Indonesia. Type I diabetes and Type II are the two basic kinds of diabetes. Type II diabetes is the most prevalent type of the disease, making up over 90% of cases. The development of this disease is brought on when the body develops an inability to respond to insulin and the body's need for the hormone outweighs the pancreas' capacity to generate it, ultimately leading to an insulin deficit [1]. Total Type II diabetic cases among young people have significantly increased during the past few decades. Various AI-based approaches have been utilized for data mining to extract usable information from massive volumes of data and use that data for treating various diseases through early diagnosis and prompt treatment as a result of breakthroughs in the field of artificial intelligence. According to a study by the National Library of Science, the estimates of the total world population by 2045 are given in Table 1. It pertinently implies that the disease is going to be a bigger problem in future. Deep neural networks are capable of digesting vast amounts of data and applying expertise to make early disease diagnoses, which offers promising outcomes.



	2021	2030	2045
Total world population	7.9 billion	8.6 billion	9.5 billion
Adult population (20–79 years)	5.1 billion	5.7 billion	6.4 billion

Table 1. Diabetes estimates by 2045.

However, there are several abnormalities in the data collected from different sources, including missing numbers, outliers, and undesired features. Such abnormalities could lead to erroneous predictions and outcomes.

Different pre-processing approaches can be used to a dataset to cope with these abnormalities so that it can be used for diagnosis by deep neural networks and machine learning algorithms. One of the finest methods for

[1] Email : mtawseef805@gmail.com

determining the importance of features and choosing features is Spearman's rank correlation coefficient. The dataset's missing values can be computed using the missing value imputation method after the most significant features have been chosen from it. When the data points are dispersed in a nonlinear way, finding the association between two or more independent attributes and the dependent attribute is done using polynomial regression and then utilising that relation to impute missing values. The data augmentation technique will be applied to deal with tiny datasets for deep learning to handle under- and oversampling while efficiently training the deep neural network. Many studies [2] have emphasized the role of digital therapies in lifestyle therapy for diabetes management and underscored the significance of artificial intelligence (AI) in enabling continuous remote monitoring of patient symptoms and biomarkers. The necessity of early diabetes detection to mitigate potentially fatal consequences such as heart attacks, heart failure, and strokes is emphasized [3]. The methodologies are proposed for diabetes diagnosis which could serve as an efficient prognostic tool for healthcare professionals, paving the way for the development of automated prognostic instruments for early disease diagnosis. An extensive review of deep learning's applications in diabetes is presented in [4], focusing on glucose management, diabetes diagnosis, and diabetes-related complications diagnosis. Multiple deep learning algorithms and frameworks demonstrate state-of-the-art performance, surpassing traditional machine learning methods. However, challenges such as data accessibility and model interpretability persist, although ongoing developments and increasing data availability offer opportunities to address these issues, facilitating wider adoption of deep learning in clinical settings.

II. LITERATURE SURVEY

Various sources were searched for publications related to diabetes mellitus diagnosis including www.scoupus.com, www.springer.com, www.wos.com, www.mdpi.com, www.frontiers.com and https://ieeexplore.ieee.org/ among others, using the following query strings given in Table 2.

Table 2. Search Strings

#1 &	#2 &	#3 &	#4 &	#5
“diabetes”	“mellitus”	Artificial Intelligence		
“diabetes”	“mellitus”	“prediction”		
“diabetes”	“diagnosis”	“using”	machine	“learning”
“diabetes”	“diagnosis”	“using”	“deep”	“learning”
“Diabetes”	“PIMA”	“AI”		
Diabetes Diagnosis	“PIMA”	“Deep”	Learning	
Diabetes Diagnosis	using	Deep Learning	on	“PIMA”

Based on the PRISMA process and Systematic Literature Review, a detailed review of the current state of the art is provided. A study on diabetes patients was done to find out how common type 1 diabetes is, and the results showed that this diabetes is more common [5]. The dataset's missing values were handled using pre-processing techniques such as mean, median, and K-Nearest Neighbor [6], before attribute selection using forward selection and backward elimination and optimised selection using a genetic algorithm. Comparing the suggested model to other machine learning techniques, it performs better. The authors in [7] created the Enhance Weak Learner Model of AdaBoost, which improves AdaBoost and uses the correlation technique to quickly process features. The authors developed an EWLM model using SVM, NB, and KNN and showed that it performs better than other comparable models. Extra Trees and Random Forest are used for feature selection and scoring in an unsupervised learning method that employs Deep Neural Networks [8]. The authors perform an entropy calculation for each attribute, followed by an information gain calculation to determine the score of each attribute, and then they choose the four most important features that were used to diagnose diabetes mellitus. Several machine-learning techniques were applied to the PIMA dataset in a study for the early identification of diabetic mellitus [9]. To address missing values, records with missing values were removed, which decreased the dataset's cardinality. The dataset was utilized for training and testing a variety of ML models after the dimension reduction utilizing feature selection by application of LDA was completed. The findings demonstrate that ensemble approaches can aid in disease diagnosis. The findings of a statistical analysis study using the NB, LR, and RF algorithms on the PIMA dataset reveal that RF produces effective results when compared to the other two ML algorithms for mellitus diagnosis [10]. Voting ensemble feature selection and deep belief neural networks were used to determine the impact of pre-processing approaches like feature selection and feature imputation on deep neural networks diabetes onset [11]. According to the results, deep belief neural networks performed better for DM diagnosis when pre-processing methods were utilized. The outcome in [12] demonstrated that CGM is a superior alternative to OGTT for the diagnosis of GDM. The transformer-based TF-GDM diagnostic model was developed for the diagnosis of gestational diabetes mellitus [13]. To impute missing values at first, a matrix factorization method was used, and then a random forest model was used to pick features. When compared to other cutting-edge models, the outcome demonstrates that the suggested model provides more accurate findings for the diagnosis of GDM. to leverage IoT to better the lives of regular people who have diabetes mellitus [14]. Two datasets were combined to train the suggested system. For tracking people's diabetes problems and managing their diabetes, the IOT-fog cloud combo produces better results. Infrared imaging is utilized to determine various diseases by extracting hidden information from images to control foot ulcers in diabetic patients [15]. This non-invasive technique was suggested by the authors as a quicker way to diagnose ulcers in diabetic patients. Pre-diabetic, positive, and negative cases from the dataset were identified using a rule-based multi-class classification system to reduce the risk of Type II diabetes [16]. The feature selection process chooses the dataset's three most important features—BMI, Plasma Glucose, and Blood Pressure—and uses them with the suggested system. The findings are then compared with those of the Decision Tree, RepTree, and Logistic Regression algorithms for diabetes onset. The findings demonstrated that the suggested platform has higher accuracy than any existing comparable algorithms. In [17], the authors created a framework using machine learning methods such as RF, SVM, LR, and KNN. The author uses blood reports, gestational diabetes prediction, and retinal images to forecast diabetes. The suggested model yields typical outcomes in the identification of diabetes mellitus. Before applying machine learning techniques, it was advised to utilize data augmentation on the PIMA dataset to prevent under-sampling because of the tiny dataset [18]. When combined with augmented datasets, machine learning algorithms significantly improve their ability to forecast diabetes mellitus. When employed with the PIMA dataset in its raw form, LR performed better, according to research on the effectiveness of different ML algorithms [19]. For accurate pre-diabetic stage diagnosis, the mayfly-SVM multi-class predictive model was suggested and put into practice [20]. The proposed model has significantly improved in terms of several performance indicators. A paradigm for diabetes onset was developed in a research study [21]. Polynomial regression was used to impute missing values before training the model, and the Spearman correlation coefficient was used to select features. Grid search and repeatedly stratified K-Foldcross-validation were used to create several machine learning algorithms with hyper-parameter optimization. The random forest method produces better outcomes for the framework. The 5-layer Neural Network model with SMOTE for addressing data imbalance was integrated with the updated Adaptive Network Fuzzy Inference System to increase the cost-effectiveness of the diabetes diagnosis. The proposed model has an accuracy of 97.7\% [22] and performs effectively in the early detection of diabetes mellitus. Average Weighted Objective Distance, a brand-new diabetes prediction model, was created [23]. The dimension reduction of the dataset made possible by feature selection utilizing information gain speeds up learning. A completely automated deep-learning model was used in a study to examine the biomarkers of abdomen CT for Type II DM [24]. Muscle volume, visceral fat, liver CT attenuation, and atherosclerotic plaque biomarkers were examined. Based on the interval between the diagnosis and the CT scan, the patients were separated into groups. The logistic regression model was trained for early diabetes mellitus prediction, and the clinical parameters that were taken into consideration were sex, age, and BMI. The deep learning model achieved the best results in diagnosing DM using DNA sequence classification, according to a study of the significance of DNA sequence classification for early diagnosis of diabetes mellitus that examined large volumes of biological data and attempted to extract useful information from the dataset [25]. The impact of several pre-processing methods for early diabetes mellitus prediction was investigated [26], using a variety of normalization methods such as z-score and min-max procedures. After applying pre-processing techniques, the accuracy of the missing value imputation utilizing mean, mode, and median was significantly improved. The PIMA India dataset [27] was used. When comparing the effectiveness of several data mining techniques for the early detection of DM, Artificial Neural Networks outperform the other methods. The dataset was used by the authors in its unprocessed state without any feature selection or pre-processing. Six different facets, including dataset, feature selection, feature imputation, pre-processing, machine learning, classification, and deep learning techniques, were carried out in a review study [28], and the importance of early detection of diabetes mellitus was emphasized.

There has been a decline in the number of research publications in recent years for papers related to the PIMA dataset and the usage of deep networks as shown in Fig. 1. The reason is the smaller size of the available numeric dataset. This study mainly focuses on generating an image-based dataset on the existing PIMA dataset and using DNN for enhanced results.

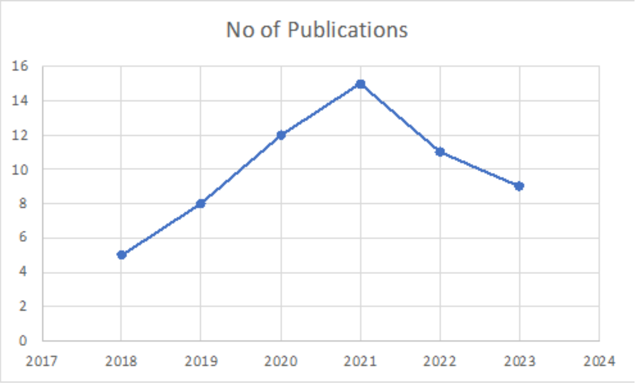


Figure. 1. Decline in DNN-based Publications.

According to the study [29], feature selection and MVI approaches combined can greatly improve the performance of classification models in the diagnosis and prediction of diabetes. The review by authors in [30] comprehensively examines various methods employed in the identification, diagnosis, and self-care of Diabetes Mellitus (DM) across six domains. The study in [31] aims to enhance prediction models' accuracy and their adaptability to diverse datasets. It proposes an improved K-means algorithm and logistic regression algorithm, supplemented by several preprocessing steps, to construct the platform. The effectiveness of the proposed framework is tested using the Waikato Environment for Knowledge Analysis toolbox and the Pima Indians Diabetes Dataset, demonstrating a 3.04% increase in prediction accuracy compared to earlier research when dataset quality is assured. The research in [32] describes a deep learning architecture-based methodology for classifying Heart Rate Variability (HRV) signals into normal and diabetic categories. The CNN and CNN-LSTM designs exhibit performance improvements of 0.03% and 0.06%, respectively. With a high accuracy rate of 95.7%, this categorization approach presents a significant opportunity for accurate diabetes diagnosis from ECG data. The work in [33] introduces a deep learning-based, highly sensitive biosensor for synchronized diabetes diagnosis. The platform, which integrates nanotechnology and electrochemical techniques, offers sensitivity and selectivity in detecting glucose molecules, enabling quick and precise diabetes diagnosis in a single test. The biosensor detects glucose levels in the range of 0.5–5 mmol per litre, surpassing current methods. The research [34] presents a deep-learning pipeline for predicting diabetes incidence. The pipeline integrates various deep learning methods, including a convolution neural network (CNN) for classification, VAE for data augmentation, and SAE for feature augmentation. Utilizing the Pima Indians Diabetes Database, the study demonstrates that training the CNN classifier alongside SAE for feature augmentation on a balanced dataset yields a notable accuracy of 92.31%. Other machine learning and deep learning have a plethora of applications which can be implemented in diabetes [35-49].

III. METHODOLOGY

The PIMA dataset that was used in this study is freely available and was obtained from Kaggle. There are nine numerical properties in all; eight of them are independent, while the ninth, called "class," is dependent. The dataset is examined closely for any missing values, which are then handled using Polynomial Regression Imputation. The complete methodology is given in Fig. 2. The PIMA dataset is pertinently converted into an Image dataset for the application of CNN-based networks.

1.1. PIMA Dataset

PIMA Indian dataset is available at the National Institute of Diabetes and Digestive and Kidney Disease (NKDDKD) used way back from 1965 onwards to study diabetes and the risks associated with diabetes mellitus. The dataset contains various diagnostic measurements and parameters that can be analysed and used for the onset and diagnosis of DM. The dataset consists of 768 tuples and 9 attributes. The dataset contains 500 records of non-diabetic persons and 268 tuples of diabetic persons. Out of 9 attributes, 8 attributes are predictive attributes and the 9th attribute is a diagnosing attribute. The dataset in its primitive form is not suitable to be used to train the model.

The dataset contains various missing values and un-normalized data. Further, some attributes in the dataset are non-significant towards a diagnosis of diabetes mellitus. The dataset must be pre-processed for the model to be used for training. For imputation of the missing values, various techniques are available viz mean, mode, median, constant etc. These methods of data imputation are simple to use but they introduce some amount of bias in the dataset that results in biased and unreliable results to be generated by the dataset. In this study, we use a more sophisticated and more reliable technique namely polynomial regression for the imputation of missing values. Post missing value imputation the next step is to score attributes as per their contribution to making a diagnosis of diabetes mellitus. Again, there are multiple techniques used for feature selection and feature importance, we used Spearman's rank correlation coefficient in this work.

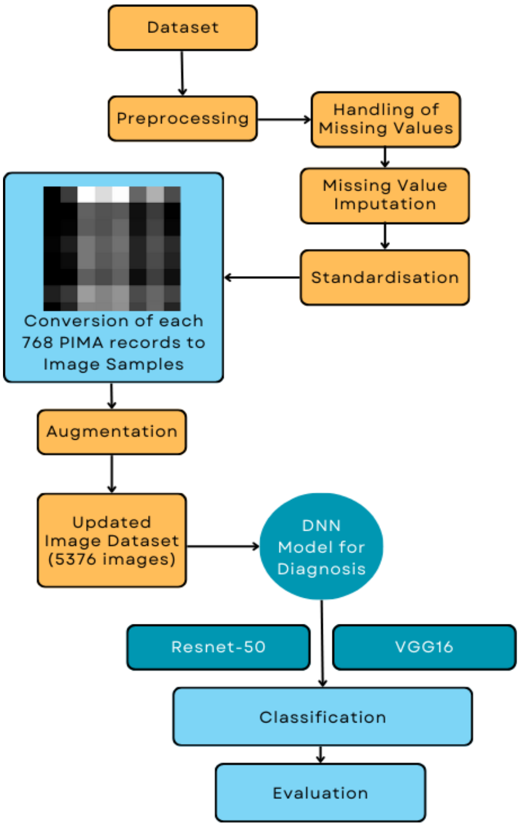


Figure. 2. Proposed Methodology.

1.1.1. Handling of Missing Values

For an experiment of an observation, the outcome depends upon the input dataset. The input data collected in an organized way gives reliable results however in most cases the input data collected in a well-controlled manner may contain some anomalies like missing/not available values. The presence of these anomalies in the dataset affects the final result. There are several ways to handle the missing values like mean, median and mode. The simplest approach is to remove records from the dataset that have missing values; however, if there are a lot of missing values in the dataset, this could lead to information loss. Using the mean, mode, or median in place of the missing values is another way to cope with missing data. Replacing missing values with these statistical observations may introduce some kind of bias in the dataset and ultimately result in a biased experiment or observation. The most commonly used missing values imputation methods are mean, KNN, Regression and Hot-deck.

1. Mean Imputation: In this method, the missing value of any feature in a particular record is replaced by the mean value of other records for that attribute. In this way, the sample size of the dataset is preserved and it is

easy to use, however, it reduces the variability in data which may lead to underestimation of standard deviation and variance. The mean imputation is calculated as follows:

$$\hat{z_i} = \bar{z}_h \qquad (1)$$

where,

$$\hat{z_i} = imputed\ value\ of\ record\ i\ and$$

$$z_h = \text{sample mean of respondent data within some class}$$

2. Regression: To deal with the problems faced in the mean imputation method, regression imputation can be used. The formulae for missing value imputation using regression are as follows:

$$\hat{z}RI(m) = p_0 + \sum_{i=1}^{2} p_i + \sum_{i=1}^{2} p_{ii} {x_i}^2 + \sum_{i<j} \sum p_{ij} x_i x_j + \varepsilon \qquad (2)$$

3. KNN imputation: The KNN missing value imputation imputes the missing value of an attribute based on the concept of feature similarity. The missing value is replaced by looking for K closest neighbours.
4. In this research study, polynomial regression is used for the imputation of missing values. It is another type of linear regression that uses an nth-degree polynomial to model the relationship between the independent variable (x) and the dependent variable (y). This method of missing value imputation is used when the two variables are related in a nonlinear fashion. The general form of polynomial regression is:

$$y = a + b_1 x + b_2 x^2 + \cdots + b_n x^n \qquad (3)$$

The main features of using polynomial regression for missing value imputation are:

1. It can fit a wider range of functions.
2. It can fit a large range of curvatures.
3. It provides a more accurate relationship between two variables.

1.2. Standardization

One essential stage in the preprocessing of data for machine learning tasks is standardization, which is sometimes referred to as feature scaling or normalization. It is the process of altering a dataset's features to have a zero mean and one standard deviation. Through this process, features become more similar and machine learning algorithms—particularly those that depend on feature scale, including support vector machines, k-nearest neighbours, and neural networks—perform better. Some of the techniques are as under:

1.2.1. MixMax

All values within the range of 0 and 1 are altered using the MinMax method. With column d, the function has the following definition:

$$diff[d] = \frac{(diff[d] - diff[d].min())}{diff[d].max() - diff[d].min()} \qquad (4)$$

1.2.2. Standard Scaler

An attribute is made more uniform by converting it from mean to unit variance after the mean has been removed.

1.2.3. Robust Scaler

Characteristics that are resilient to outliers are scaled by robust scaler algorithms. The interquartile is employed.

1.3. PIMA to Image Dataset

The text emphasizes how deep learning, in particular, CNN, is superior to conventional machine learning methods in several applications [50-51]. Applying current CNN models—which are made for 2D data, such as images—to the PIMA diabetes dataset, which is made up of numerical values, presents a problem. Existing methods use 1D

CNN models that are specially designed for this dataset. To facilitate the use of well-established CNN models for feature extraction and subsequent diabetes prediction, this work suggests transforming the raw PIMA data. This method seeks to enhance diabetes detection by utilizing deep learning capabilities.

1.3.1. PIMA Dataset

The NKDDKD originally offered the PIMA dataset which has been used to create machine-learning models for early diabetes detection. The dataset comprises nine parameters, namely: age, outcome, skin thickness, blood pressure, glucose concentration, insulin level, body mass index (BMI), diabetes pedigree function, and pregnancy. The collection has 768 entries in total. The dependent variable we are trying to forecast is the "Outcome" attribute. The result of 1 shows the presence of diabetes, whereas a value of 0 indicates no diabetes. 500 entries have a value of 0 (non-diabetic), according to the analysis of the "Outcome" attribute, while 268 samples with 1 (diabetic).

The creation of a universal machine learning model that can diagnose Type I and Type II diabetes is hampered by several issues [52]. Using datasets with inadequate records might be problematic since it can produce erroneous findings [53]. Furthermore, several research uses the PIMA dataset without performing necessary preprocessing procedures like normalization. This may impair the models' accuracy by introducing problems such as outliers, overfitting, and underfitting [54-55]. Moreover, several research works use restricted machine learning methods for diagnosis and fail to deal with missing values in the data [56]. The limited use of feature extraction algorithms is another drawback. The procedure of extracting features might be greatly enhanced by automatic deep feature extraction [57].

One of the biggest limitations in applying a deep learning model like Resnet50 is not having a large enough dataset to train the model correctly. The study intends to modify this limitation by creating an image-based dataset from PIMA.

1.3.2. PIMA Image Dataset

The methodology for converting the PIMA dataset into an image dataset is given in Fig. 3. As shown in the flowchart, after performing the necessary pre-processing including normalization on the textual dataset, the most pertinent characteristics from the numerical data are extracted using Spearman’s correlation coefficient. After data normalization, the bounds of these features are modified for the numeric-to-image conversion step. This approach, which is nonparametric and dependent on correlation, measures the statistical dependency of ranking between two variables. It looks at the level of correlation between variables and is represented by ρ. Pearson's Coefficient is seen to be a better choice for feature selection when variables show linear connections. On the other hand, feature selection uses Spearman's Rank Coefficient when variables show monotonic connections. To get Spearman's Rank Coefficient, use this formula:

$$\rho = 1 - \frac{6\sum_i b_i^2}{n(n^2-1)} \qquad (1)$$

where,

n= No of data points

b_i= variation ith value

ρ’s represent values in [+1, -1]

The idea of calculating the brightness of a particular area (cell) in the image based on the amplitude of each sample is used in the process of converting PIMA data to images.

Every sample in the PIMA dataset has an 8 × 8 picture structure.

1. *The relevant feature value's amplitude determines the colour of each cell in the first row of an 8x8 image.*
2. *The other 7 rows of the image are filled with amplitude values by arranging different pieces or features farther away and comparable elements together, it is possible to make collective use of nearby elements.*
3. *Since all of the data were previously normalized, the values of each feature fall between 0 and 1. After multiplying each feature value by 255, pictures with cells that range in brightness from 0 to (255 or max*

value in the sample record) are produced.

4. *For each of the 768 records, a total of 7 samples are generated using augmentation techniques of reflection, rotation and translation etc.*
5. *The final dataset is generated with 5376 total images.*

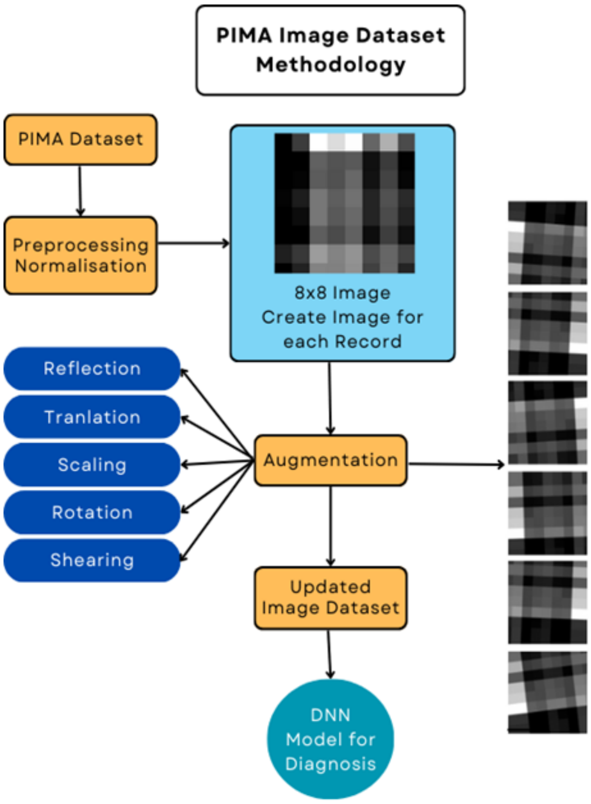


Figure. 3. Proposed Methodology.

Additionally, the application of data augmentation techniques improves the classification accuracy of deep CNN models by increasing the dataset size. From the PIMA dataset containing 768 records, a total of 768 images were obtained. Then the augmentation technique encompassing reflection, rotation, translation, scaling and shearing was employed to generate 7 samples on each sample. The final dataset contains a total of 5376 images. Table 3 provides the details about the generated dataset.

Table 3. Image Dataset

	Test (a)	**Train (b)**	**Val (c)**	
YES	190	1504	352	
NO	344	2804	352	
Total	532	4308	534	**5376**

IV. EXPERIMENTATION AND RESULTS

The experiment was done on the newly generated PIMA Image dataset using two CNN-based models, VGG16 and Resnet50. Vgg16 and Resnet50 were trained with a total of 120 epochs on pre-trained ImageNet weights. The confusion matrix for VGG16 and resnet50 are shown in Fig. 4 and Fig. 5 respectively.

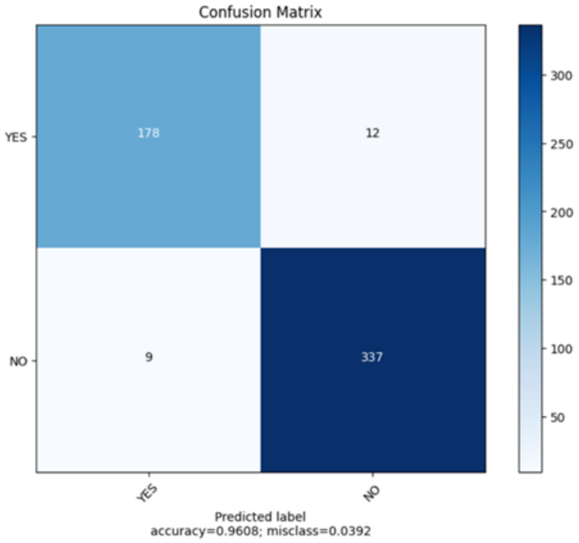


Figure 4. Confusion Matrix Vgg16.

1.4. Evaluation Criteria

The following evaluation criteria have been used to ascertain the performance of different ML methods:

$$Acc = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

$$Sen = \frac{(TP)}{(TP+FN)} \quad (2)$$

$$Spec = \frac{(TN)}{(TN+FP)} \quad (3)$$

$$F1\ Score = \frac{(2TP)}{(2TP+ FP+FN)} \quad (4)$$

The performance comparison between Vgg16 and ResNet50, and other research studies is shown in Table 3 and Table 4 respectively.

Table 4 shows that the VGG16 and ResNet50 models produce the best accuracy with rates of 96.07% and 97.19%, respectively. As demonstrated by the approach's successful classification using VGG16 and ResNet50 models, the method's results

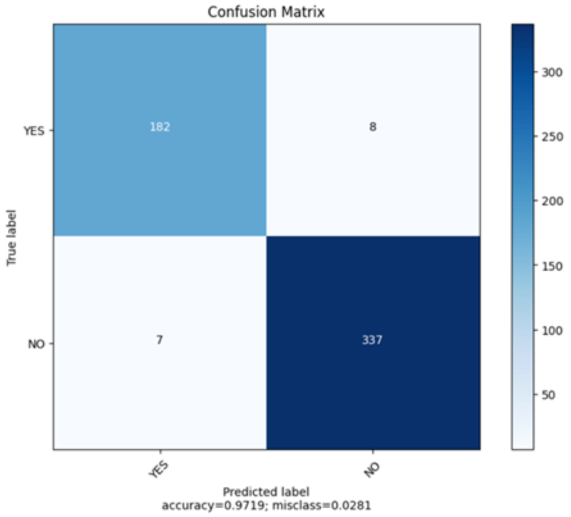


Figure 5. Confusion Matrix Resnet50.

show that turning diabetes data into an image dataset is a useful strategy.

When these results are compared to previous research using the PIMA dataset, as Table 5 illustrates, the suggested approach outperforms many previous research efforts. On the other hand, the research work creatively converted the PIMA dataset into image data, making it compatible with widely used CNN models and setting it apart from earlier methods.

Table 4. Performance of the proposed model.

Measure	VGG16	Resnet50
Sensitivity	0.9368	0.9579
Specificity	0.9738	0.9797
Precision	0.9519	0.963
Negative Predictive Value	0.9654	0.9768
False Positive Rate	0.0262	0.0203
False Discovery Rate	0.0481	0.037
False Negative Rate	0.0632	0.0421
Accuracy	**0.9607**	**0.9719**
F1 Score	0.9443	0.9604

Table 5. Comparative analysis with previous works.

Ref	Technique	Accuracy
[58]	ANN	92
[59]	SA-DNN	86.26
[32]	LSTM-CNN	95.70%
[34]	CNN-SAE	92.31%
Proposed method	VGG16 and ResNet50 with PIMA Image Dataset	96.07%, 97.19%

V. CONCLUSION

Numerous authors' studies from the past few decades demonstrate the value of deep learning as a technique for diabetes mellitus early diagnosis. Deep learning algorithms may diagnose diseases much more accurately if the dataset is improved using pre-processing techniques that handle over-fitting, under-fitting, redundancy, missing values, and inconsequential features. The methodology proposed in this study has the potential to be applied to a variety of numerical datasets. Although research based on deep learning has significantly decreased dependence on particular traits, the importance of precisely constructed structures has gained prominence. However, the approach shown in this work is important since it can also support deep and complete structures for numerical data. While the production of image data may need extra processing steps in comparison to studies that use raw data directly, this application presents opportunities to improve diabetes prediction ability. This is because different designs of CNN models may now adapt to numerical inputs. Moreover, the integration of data augmentation methodologies may be executed with ease with images related to diabetes.

REFERENCES

[1] M. Khanna, L. K. Singh, S. Thawkar, and M. Goyal, “Deep learning based computer-aided automatic prediction and grading system for diabetic retinopathy,” *Multimed. Tools Appl.*, Mar. 2023, doi: 10.1007/s11042-023-14970-5.

[2] S. Ellahham, “Artificial Intelligence: The Future for Diabetes Care,” Am. J. Med., vol. 133, no. 8, pp. 895–900, 2020, doi: 10.1016/j.amjmed.2020.03.033.

[3] B. Tymchenko, P. Marchenko, and D. Spodarets, “Deep Learning Approach to Diabetic Retinopathy Detection,” Mar. 2020, [Online]. Available: http://arxiv.org/abs/2003.02261

[4] T. Zhu, K. Li, P. Herrero, and P. Georgiou, “Deep Learning for Diabetes: A Systematic Review,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7. Institute of Electrical and Electronics Engineers Inc., pp. 2744–2757, Jul. 01, 2021. doi: 10.1109/JBHI.2020.3040225.

[5] S. O. Song *et al.*, “Prevalence and clinical characteristics of fulminant type 1 diabetes mellitus in Korean adults: A multi-institutional joint research,” *J. Diabetes Investig.*, vol. 13, no. 1, pp. 47–53, Jan. 2022, doi: 10.1111/jdi.13638.

[6] F. Zamani Boroujeni, R. Asgarnezhad, and M. Shekofteh, “Improving diagnosis of diabetes mellitus using combination of preprocessing techniques,” *Artic. J. Theor. Appl. Inf. Technol.*, vol. 15, no. 13, 2017, [Online]. Available: https://www.researchgate.net/publication/318777104

[7] P. Sornsuwit, “ENHANCE WEAK LEARNER MODEL OF ADABOOST (EWDM) FOR DIABETES MELLITUS CLASSIFICATION,” *Int. J. Innov. Comput. Inf. Control*, vol. 18, no. 4, pp. 1117–1132, Aug. 2022, doi: 10.24507/ijicic.18.04.1117.

[8] B. M. K. P, S. P. R, N. R K, and A. K, “Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier,” *Int. J. Cogn. Comput. Eng.*, vol. 1, pp. 55–61, Jun. 2020, doi: 10.1016/j.ijcce.2020.10.002.

[9] S. Simaiya *et al.*, “A novel multistage ensemble approach for prediction and classification of diabetes,” *Front. Physiol.*, vol. 13, Dec. 2022, doi: 10.3389/fphys.2022.1085240.

[10] F. A. Jaber and J. W. James, “Early Prediction of Diabetic Using Data Mining,” *SN Comput. Sci.*, vol. 4, no. 2, Jan. 2023, doi: 10.1007/s42979-022-01594-z.

[11] O. Olabanjo, M. Mazzara, and A. Wusu, “Deep Unsupervised Machine Learning for Early Diabetes Risk Prediction using Ensemble Feature Selection and Deep Belief Neural Networks Prediction of Twitter Message Deletion View project AutoReq View project Deep Unsupervised Machine Learning for Early Diabetes Risk Prediction using Ensemble Feature Selection and Deep Belief Neural Networks,” 2023, doi: 10.20944/preprints202301.0208.v1.

[12] D. Di Filippo *et al.*, “A new continuous glucose monitor for the diagnosis of gestational diabetes mellitus: a pilot study,” *BMC Pregnancy Childbirth*, vol. 23, no. 1, p. 186, Mar. 2023, doi: 10.1186/s12884-023-05496-7.

[13] H. Wang, Y. Yao, J. Zheng, D. Peng, J. Wu, and J. Wang, “Accurate prediction of gestational diabetes mellitus via a novel transformer method,” 2023, doi: 10.21203/rs.3.rs-2461259/v1.

[14] A. Pati *et al.*, “Diagnose Diabetic Mellitus Illness Based on IoT Smart Architecture,” *Wirel. Commun. Mob. Comput.*, vol. 2022, 2022, doi: 10.1155/2022/7268571.

[15] M. Rai, T. Maity, R. Sharma, and R. K. Yadav, “Early detection of foot ulceration in type II diabetic patient using registration method in infrared images and descriptive comparison with deep learning methods,” *J. Supercomput.*, vol. 78, no. 11, pp. 13409–13426, Jul. 2022, doi: 10.1007/s11227-022-04380-z.

[16] R. Karthikeyan, P. Geetha, and E. Ramaraj, “THE RULE-BASED MULTI-CLASS CLASSIFICATION MODEL PREDICTS EARLY DIABETES USING SUPERVISED MACHINE LEARNING TECHNIQUES,” 2022, [Online]. Available: https://dbdxxb.cn/

[17] K. D. Yesugade, H. V Ankam, A. A. Urunkar, P. D. Dede, and S. S. Kale, “MACHINE LEARNING BASED WEB APPLICATION FOR DIABETES PREDICTION,” JETIR, 2022. [Online]. Available: www.jetir.orgh488

[18] B. S. Ahamed, M. S. Arya, and A. O. V. Nancy, “Diabetes Mellitus Disease Prediction Using Machine Learning Classifiers with Oversampling and Feature Augmentation,” *Adv. Human-Computer Interact.*, vol. 2022, 2022, doi: 10.1155/2022/9220560.

[19] Z. Zaman, M. A. A. A. Shohas, M. H. Bijoy, M. Hossain, and S. Al Sakib, “Assessing Machine Learning Methods for Predicting Diabetes among Pregnant Women,” *Int. J. Adv. Life Sci. Res.*, vol. 05, no. 01, pp. 29–34, 2022, doi: 10.31632/ijalsr.2022.v05i01.005.

[20] R. Patil, S. Tamane, S. A. Rawandale, and K. Patil, “A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus,” *Int. J. Electr. Comput. Eng.*, vol. 12, no. 1, pp. 524–533, Feb. 2022, doi: 10.11591/ijece.v12i1.pp524-533.

[21] C. C. Olisah, L. Smith, and M. Smith, “Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective,” *Comput. Methods Programs Biomed.*, vol. 220, Jun. 2022, doi: 10.1016/j.cmpb.2022.106773.

[22] X. Wang *et al.*, “A Revised Adaptive Network-based Fuzzy Inference System Combined with Neural Network to Predict Diabetes,” 2022, doi: 10.21203/rs.3.rs-2388120/v1.

[23] Aleena Farooq, Muhammad Kamran Abid, Wasif Akbar, Hafiz Humza, and Naeem Aslam, “Type-II Diabetes Prediction by using Classification and Novel based Method (AWOD),” *J. Comput. Biomed. Informatics*, vol. 4, no. 01, pp. 152–174, Dec. 2022, doi: 10.56979/401/2022/110.

[24] H. Tallam, D. C. Elton, S. Lee, P. Wakim, P. J. Pickhardt, and R. M. Summers, “Fully Automated Abdominal CT Biomarkers for Type 2 Diabetes Using Deep Learning,” *Radiology*, vol. 304, no. 1, pp. 85–95, Jul. 2022, doi: 10.1148/radiol.211914.

[25] N. E. El-Attar, B. M. Moustafa, and W. A. Awad, “Deep learning model to detect diabetes mellitus based on dna sequence,” *Intell. Autom. Soft Comput.*, vol. 31, no. 1, pp. 325–338, 2022, doi: 10.32604/IASC.2022.019970.

[26] O. Shafi, J. S. Sidiq, T. Ahmed Teli, and K. -, “EFFECT OF PRE-PROCESSING TECHNIQUES IN PREDICTING DIABETES MELLITUS WITH FOCUS ON ARTIFICIAL NEURAL NETWORK,” 2022.

[27] O. Shafi, S. J. Sidiq, T. A. Teli, and M. Zaman, “A Comparative Study on Various Data Mining Techniques for Early Prediction of Diabetes Mellitus,” 2021.

[28] J. Chaki, S. Thillai Ganesh, S. K. Cidham, and S. Ananda Theertan, “Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6. King Saud bin Abdulaziz University, pp. 3204–3225, Jun. 01, 2022. doi: 10.1016/j.jksuci.2020.06.013.

[29] C. C. Olisah, L. Smith, and M. Smith, “Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective,” *Comput. Methods Programs Biomed.*, vol. 220, Jun. 2022, doi: 10.1016/j.cmpb.2022.106773.

[30] J. Chaki, S. Thillai Ganesh, S. K. Cidham, and S. Ananda Theertan, “Machine learning and artificial intelligence-based Diabetes Mellitus detection and self-management: A systematic review,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6. King Saud bin Abdulaziz University, pp. 3204–3225, Jun. 01, 2022. doi: 10.1016/j.jksuci.2020.06.013.

[31] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, “Informatics in Medicine Unlocked Type 2 diabetes mellitus prediction model based on data mining,” *Informatics Med. Unlocked*, vol. 10, no. August 2017, pp. 100–107, 2018, doi: 10.1016/j.imu.2017.12.006.

[32] G. Swapna, R. Vinayakumar, and K. P. Soman, “Diabetes detection using deep learning algorithms,” *ICT Express*, vol. 4, no. 4, pp. 243–246, 2018, doi: 10.1016/j.icte.2018.10.005.

[33] A. Armghan, J. Logeshwaran, S. M. Sutharshan, K. Aliqab, M. Alsharari, and S. K. Patel, “Design of biosensor for synchronized identification of diabetes using deep learning,” *Results Eng.*, vol. 20, no. September, p. 101382, 2023, doi: 10.1016/j.rineng.2023.101382.

[34] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, “Diabetes detection using deep learning techniques with oversampling and feature augmentation,” *Comput. Methods Programs Biomed.*, vol. 202, 2021, doi: 10.1016/j.cmpb.2021.105968.

[35] Ahmed Teli, T., Masoodi, F., & Yousuf, R. (2020). *Security concerns and privacy preservation in blockchain based IoT*

systems: opportunities and challenges.

[36] Teli, T. A., Masoodi, F. S., & Bahmdi, A. M. (2021). HIBE: Hierarchical identity-based encryption. In *Functional Encryption* (pp. 187–203). Springer International Publishing Cham.

[37] Teli, T. A. (2022). Ensuring Secure Data Sharing in IoT Domains Using Blockchain. In *CYBER SECURITY AND DIGITAL FORENSICS* (pp. 205–219). scrivener publishing with wiley.

[38] Shafi, O., Sidiq, S. J., Teli, T. A., & Zaman, M. (2021). *A Comparative Study on Various Data Mining Techniques for Early Prediction of Diabetes Mellitus*.

[39] Shafi, O., Sidiq, J. S., Teli, T. A., & Others. (2022). Effect of pre-processing techniques in predicting diabetes mellitus with focus on artificial neural network. *Advances and Applications in Mathematical Sciences*, *21*(8), 4761–4770.

[40] Teli, T. A., & Wani, A. (2018). Fuzzy Logic And Medicine With Focus On Cardiovascular Disease Diagnosis. *5th International Conference on Computing for Sustainable Global Development*.

[41] Ahmed Teli, T., & Masoodi, F. (2021). Blockchain in healthcare: Challenges and opportunities. *Proceedings of the International Conference on IoT Based Control Networks & Intelligent Systems-ICICNIS*.

[42] Zargar, O. S., Baghat, A., & Teli, T. A. (2022). A DNN Model for Diabetes Mellitus Prediction on PIMA Dataset. *INFOCOMP Journal of Computer Science*, *21*(2).

[43] Teli, T. A., & Yousuf, R. (2023). Deep Learning for Bioinformatics. In *Applications of Machine Learning and Deep Learning on Biological Data* (pp. 181–196). Auerbach Publications.

[44] Teli, T. A., Masoodi, F. S., & Masoodi, Z. (2023). Application of ML and DL on Biological Data. In *Applications of Machine Learning and Deep Learning on Biological Data* (pp. 159–180). Auerbach Publications.

[45] Zargar, O. S., Bhagat, A., & Teli, T. A. (2023). Feature Selection, Importance and Missing Value Imputation in Diabetes Mellitus Prediction. *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, 914–919. IEEE.

[46] Mushtaq, S., Roy, A., & Teli, T. A. (2021). A comparative study on various machine learning techniques for brain tumor detection using MRI. *Proc. of the Global Emerging Innovation Summit (GEIS-2021)*, 125–137.

[47] Zargar, O. S., Bhagat, A., Teli, T. A., & Sheikh, S. (n.d.). EARLY PREDICTION OF DIABETES MELLITUS ON PIMA DATASET USING ML AND DL TECHNIQUES. *Journal of Army Engineering University of PLA ISSN*, *2097*, 0730.

[48] Sofi, M. A., Singh, D., & Teli, T. A. (n.d.). AN INTELLIGENT APPROACH FOR DE-NOVO DRUG DISCOVERY: A SYSTEMATIC REVIEW. *Journal of Army Engineering University of PLA ISSN*, *2097*, 0730.

[49] Jabbari, A., Teli, T. A., Masoodi, F., Reegu, F. A., Uddin, M., & Albakri, A. (2024). Prioritizing factors for the adoption of IoT-based smart irrigation in Saudi Arabia: a GRA/AHP approach. *Frontiers in* Agronomy, 6, 1335443.

[50] Saleem, T.J.; Chishti,M.A. Deep learning for the internet of things: Potential benefits and use-cases. Digit. Commun. Netw. 2021, 7, 526–542.

[51] O'Mahony, N.; Campbell, S.; Carvalho, A.; Harapanahalli, S.; Hernandez, G.V.; Krpalkova, L.; Riordan, D.;Walsh, J. Deep learning vs. traditional computer vision. In Proceedings of the Science and Information Conference, Las Vegas, NV, USA, 2–3 May 2019; *pp. 128–144.*

[52] A. Saxena et al., "Data mining techniques based diabetes prediction,", IJAINN, vol. 1, no. 2, pp. 29-35, 2021. doi:10.35940/ijainn.B1012.041221.

[53] M. Alehegn, "Analysis and prediction of diabetes mellitus using machine learning," vol. 9, pp. 871-878, 2018 Algorithm. 118.

[54] J. R. Raut, "Performance evaluation of various supervised machine learning algorithms for diabetes," vol. 7, no. 8, pp. 4921-4925, 2020.

[55] M. K. Hasan et al., "Diabetes prediction using ensembling of different machine learning classifiers,", IEEE Access, vol. 8, 76516-76531, 2020. doi:10.1109/ACCESS.2020.2989857.

[56] O. Shafi et al., K. (2022), "Effect of preprocessing techniques in predicting diabetes mellitus with focus on artificial neural network" in Adv. Appl. Math. Sci., vol. 21, no. 8.

[57] A. Yahyaoui et al., “A decision support system for diabetes prediction using machine learning and deep learning techniques,” vol. 2, pp. 1-4, 2019. doi:10.1109/UBMYK48245.2019.8965556.

[58] Srivastava, S.; Sharma, L.; Sharma, V.; Kumar, A.; Darbari, H. Prediction of Diabetes Using Artificial Neural Network Approach; Springer: Singapore, 2019; pp. 679–687.

[59] Kannadasan, K.; Edla, D.R.; Kuppili, V. Type 2 diabetes data classification using stacked autoencoders in deep neural networks. Clin. Epidemiol. Glob. Health 2019, 7, 530–535.