Scientific Research Publishing

Journal of Computer and Communications, 2022, 10, 15-28
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Leveraging Pima Dataset to Diabetes Prediction: Case Study of Deep Neural Network

Pélagie Hounguè, Annie Ghylaine Bigirimana

Laboratory of Data Science and Digital Inclusion, Institut de Mathématiques et de Sciences Physiques, Université d'Abomey-Calavi, Abomey-Calavi, Bénin
Email: pelagie.hounguе@imsp-uac.org, annie.bigirimana@imsp-uac.org

How to cite this paper: Hounguè, P. and Bigirimana, A.G. (2022) Leveraging Pima Dataset to Diabetes Prediction: Case Study of Deep Neural Network. *Journal of Computer and Communications*, **10**, 15-28.
https://doi.org/10.4236/jcc.2022.1011002

Received: September 30, 2022
Accepted: October 30, 2022
Published: November 2, 2022

Abstract

Diabetes is a chronic disease. In 2019, it was the ninth leading cause of death with an estimated 1.5 million deaths. Poorly controlled, diabetes can lead to serious health problems. That explains why early diagnosis of diabetes is very important. Several approaches that use Artificial Intelligence, specifically Deep Learning, have been widely used with promising results. The contribution of this paper is in two-folds: 1) Deep Neural Network (DNN) approach is used on Pima Indian dataset to predict diabetes using 10 k-fold cross validation and 89% accuracy is obtained; 2) comparative analysis of previous work is provided on diabetes prediction using DNN with the tested model. The results showed that 10 k-fold cross-validation could decrease the efficiency of diabetes prediction models using DNN.

Keywords

Deep Learning, Artificial Intelligence, Deep Neural Network, k-Fold Cross-Validation, Diabete Mellitus

1. Introduction

Diabetes is a disorder that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin, it produces. Insulin is a hormone that regulates blood sugar. Hyperglycaemia or high blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels. Glucose is an important source of energy for the cells that make up the muscles and tissues. It is also the main source of fuel of the brain. The main cause of diabetes varies by type. But, no matter what type of diabetes you have, it can lead to excess sugar in the blood. Too much sugar in the blood can lead to serious

health problems [1].

Chronic diabetes conditions include type 1 diabetes and type 2 diabetes. Potentially, reversible diabetes conditions include prediabetes and gestational diabetes. Prediabetes happens when blood sugar level is higher than normal. However, the blood sugar level could not be high enough to be called diabetes. Moreover, prediabetes can lead to diabetes unless steps are taken to prevent it. On the other hand, gestational diabetes happens during pregnancy. Nevertheless, it may go away after the baby is born. For many authors, the new prospects for automation opened up by AI require a rethinking of the division of labor between humans and machines, as advances in this field mean that AI systems are becoming more efficient than humans in many cases such as diagnosis or predictive analysis. AI systems are outperforming doctors in detecting or estimating probabilities of disease occurrence [2]. While AI is the broad science of mimicking human abilities, Machine Learning is a specific subset of AI that trains a machine how to learn and Deep Learning uses huge Neural Networks with many layers of processing units, taking advantage of advancements in computing power and improved training techniques to learn complex patterns in large amounts of data. Deep Learning uses large Neural Networks whose neurons are connected to each other and have the ability to change their hyperparameters every time new data is updated. Thus, this technology allows computer systems to learn things on their own. In fact, Machine Learning and Deep Learning techniques have attained reliable accuracy rate in classification as compared with existing algorithms.

The early prediction of some chronic diseases such as diabetes mellitus is crucial and obtaining higher accuracy rate in diabetes prediction is decisive. Hence, researchers are introducing several Machine Learning and Deep Learning based approaches for diabetes mellitus prediction. However, those technologies require very large amount of data in order to perform better than other techniques. They are extremely expensive to train due to complex data models. The objective of this paper is to prove that DNN on pima dataset and k-fold cross-validation, decrease the efficiency of diabetes detection models. Artificial Neural Network (ANN) with "n" hidden layers replicates the input data from previous output layers.

In the remaining of the paper, the fundamentals are discussed to inquire about some useful concepts. Then, the related works are reviewed, following by the proposed model description. Subsequently, experimental and comparative analysis results are presented. Finally, discussions elements are highlited just before the conclusion.

2. Background

In this section, some useful concepts are presented to understand the performed proposal.

2.1. Machine Learning

Machine learning (ML) is a type of artificial intelligence (AI). ML uses historical

data as input to predict new output values [3]. Machine Learning has four main categories:

- **Supervised learning**: method that feeds input data as well as correct output data to the ML model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable (x) with the output variable (y).
- **Unsupervised Learning**: method in which models are not supervised using a training data set. Instead, the models find hidden patterns and ideas on their own from the given data. It can be compared to the learning that takes place in the human brain when it learns new thing.
- **Semi-supervised Learning**: is the middle ground between supervised and unsupervised learning and operates on the data that consists of a few labels. It mostly consists of unlabeled data.
- **Reinforcement Learning**: feedback-based Machine Learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty.

2.2. Deep Learning

Deep Learning is a component of Machine Learning methods based on Artificial Neural Networks with representation learning. The adjective "deep" in Deep Learning refers to the use of multiple layers in the network as shown in Figure 1.

There are many kinds of Deep Learning architectures such as Deep Neural Networks, Deep Belief Networks, Deep Reinforcement Learning, Recurrent Neural Networks, Convolutional Neural Networks and Transformers.

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other [4]. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods, filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics. Figure 2 describes how Convolutional Neural Networks work.

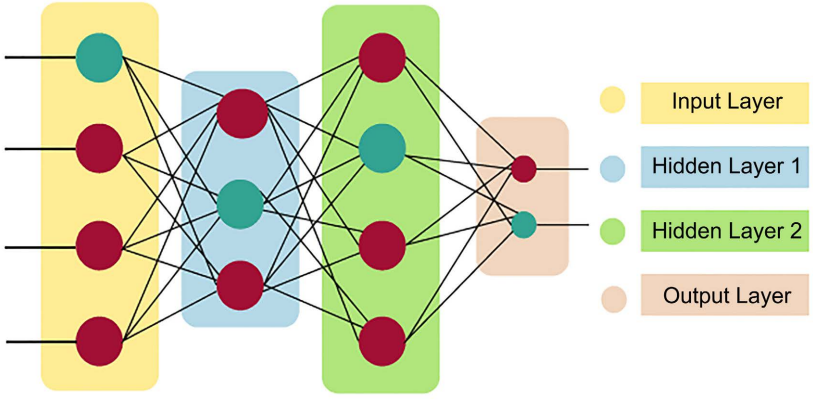

Figure 1. Architecture of deep neural network.

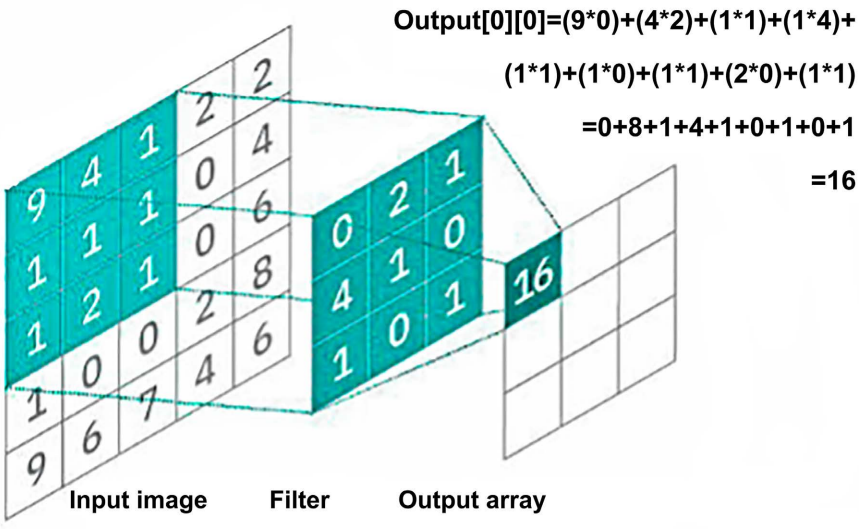

Figure 2. Architecture of convolutional neural network.

2.3. k-Fold Cross-Validation

Cross-validation randomly partitions the training data into folds. The algorithm uses 10 samples by default if the dataset has not already partitioned. To split the dataset into a different number of samples, you can use the partition and sample component and specify the number of samples to use. Cross-validation is an essential tool in the Data Scientist toolbox. It allows to better use available data. When building a Machine Learning or Deep Learning model using some data, data are often divided into training and validation/test sets. The training set is used to train the model, and the validation/test set is used to validate it on data it has never seen before. The classic approach is to do a simple 80% - 20% split, sometimes with different values like 70% - 30% or 90% - 10%. In cross-validation, researchers can do more than one split. It can be 3, 5, 10 or any k number of splits. Those splits are called folds, and there are many interesting strategies to create these folds [5]. An example is given in **Figure 3**.

3. Related Works

In order to provide a deep analysis of existing works, a systematic research was done, collection of articles are selected and summarized by focusing on two axes: diagnosis of diabetes using Deep Learning algorithms and early diagnosis of diabetes using Machine Learning algorithms.

3.1. Deep Learning Based Approaches

Deep Neural Network (DNN), an unsupervised learning approach, is used in [6] for a prediction on the Pima Indian diabetes dataset. The model achieved 98.16% accuracy. Deep learning can outperform shallow networks [7].

Authors in [8] proposed a strategy for the diagnosis of diabetes that uses a Deep Neural Network by training its attributes by cross-validation at five and ten times. The accuracy obtained is 98.35%.

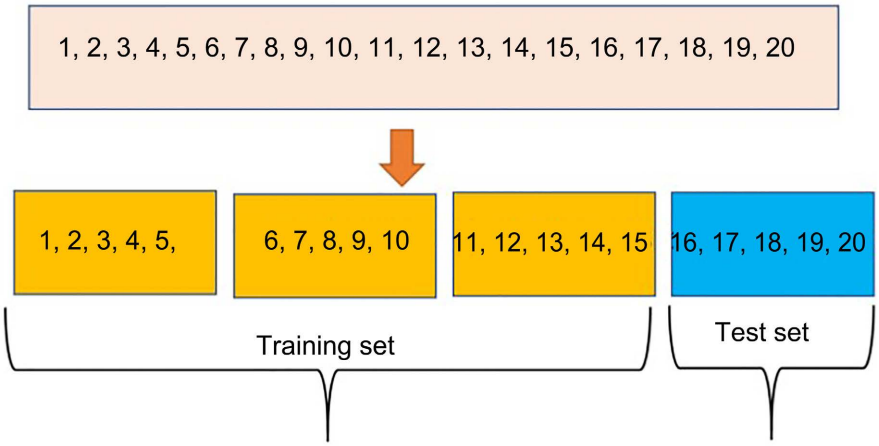

Figure 3. k-fold cross-validation.

Moreover, authors in [9] proposed a model that uses the binary cross-entropy loss function in addition to a number of parameters. The prediction model that uses Deep Neural Network on the PIMA database gives performance that is about 99.41%.

In [10], the criteria of authors were to minimize the error function in the training of the neural network using a Neural Network model. The test accuracy is 87.3%.

Besides, researchers in [11] collected electrocardiograms (ECGs) from 20 people with diabetes and a group of people lying in a stable posture for 10 minutes. Heart rate data are extracted from the ECG signals for the detection of diabetes, so the authors used 5-layer CNN, LSTM and 5-fold cross-validation and obtained 95.7% accuracy.

Authors in [12] used VRC (Heart Rate Variability) as data and CNN-LSTM (LSTM = Long Short Term Memory) to automatically detect the anomaly. 5-fold cross-validation and CNN are used and gave an accuracy of 93.6% while the CNN-LSTM combination gave the maximum accuracy of 95.1%.

Moreover, an approach is proposed [13], based on convolutional long-term memory (CLSTM), diabetes classification model was developed and compared i with existing methods on the Pima Indians Diabetes Database (PIDD). The result obtained by the CLSTM model is higher than the other methodologies 96.8%.

In addition, authors in [14] present a methodology for diabetes prediction using various Machine Learning algorithms using the PIMA. DL and DT dataset provide promising accuracy (98.07%).

Finally, authors in [15] presented DeepCare, an end-to-end dynamic Deep Neural Network that reads medical records, stores the history of previous illnesses, infers the current disease state and predicts future medical outcomes. The performance obtained is 79%.

3.2. Machine Learning Based Approaches

Early diagnosis of diabetes has been a topic in which Machine Learning algo-

rithms have played a large role before the rise of Deep Learning.

Machine learning (ML) algorithms and Neural Network (NN) methods are used in [16]. Authors found that the model with logistic regression (LR) and Support Vector Machine (SVM) works well for diabetes prediction with an accuracy of 88.6%. Six learning-based classification methods were implemented and tested on a dataset collected through online and offline questionnaires with 18 diabetes-related questions. The same algorithms were also applied to the PIMA database. The experimental results show that the accuracy of Random Forest for our dataset pima is 94.10%, which is the highest among others.

Many Machine Learning algorithms are trained on many datasets in [17]. It appears that SVM outperforms the other algorithms with 98.6% accuracy.

On other hand, researchers in [18] proposed a new approach to efficiently predict diabetes from the medical records of Indian Pima patients. The modified J48 classifier was used to increase the accuracy rate of the data mining procedure. The WEKA data mining tool was used as an API of MATLAB to generate the modified J-48 classifiers. The experimental results showed a significant improvement over the existing J-48 algorithm. It was proved that the proposed algorithm can achieve 99.87% accuracy.

The soft voting ensemble classifier proposed in [19] uses the set of three Machine Learning algorithms, namely random forest, logistic regression and Naïve Bayes, for classification, taking accuracy as evaluation criteria. The proposed ensemble approach gives the highest accuracy with 79.04% on the PIMA diabetes dataset.

In addition, the objective of authors [20] is to establish the best Machine Learning classifier for the application of wafer defect detection. The experiments proved that the logistic regression classifier is the best classifier for detection with an accuracy of 86.9%.

Three Machine Learning classification algorithms, namely Decision Tree, SVM and Naive Bayes, are used in [21]. The experiments are performed on the Pima Indians Diabetes Database (PIDD). The results obtained, show that Naive Bayes outperforms the other algorithms with the highest accuracy of 76.30%. These results are verified using ROC (Receiver Operating Characteristic) curves.

Besides, authors in [22] provide a comparative analysis of different machine learning models to reach the most supporting decision for diagnosing heart disease with better accuracy as compared to existing models. The comparative study has proven that the XGB is the most suitable model due to its superior prediction capability to other models with an accuracy of 91.6% and 100% on two different heart ailments datasets, respectively.

Moreover, authors in [23] provide the comparative analysis of different Machine Learning algorithms for diagnosis of different diseases. It brings attention towards the suite of Machine Learning algorithms and tools that are used for the analysis of diseases and decision-making process accordingly.

Futhermore, authors in [24] applied and compared Machine Learning techniques initially to predict the outcome of TB therapy. After feature analysis, six

algorithms including decision tree (DT), Artificial Neural Network (ANN), Logistic Regression (LR), Radial Basis Function (RBF), Bayesian Networks (BN), and Support Vector Machine (SVM) are developed and validated.

Finally, in [25] significant attributes selection was done via the principal component analysis method. Their findings indicate a strong association of diabetes with body mass index (BMI) and with glucose level, which was extracted via the Apriori method. Artificial Neural Network (ANN), Random Forest (RF) and K-means clustering techniques were implemented for the prediction of diabetes. The ANN technique provided a best accuracy of 75.7%, and may be useful to assist medical professionals with treatment decisions.

4. Proposed Model

Among all the approaches that use the same dataset Pima Indian dataset for training and testing their model, the Deep Neural Network performed better. These different methods (PIDD, ECG, HRV, etc.) used for diagnosing diabetes, have the clinical advantage of not requiring a blood sample for the diagnosis of diabetes, as these methods can diagnose diabetes without pain.

The hope is that as the number of layers increases, Neural Networks will learn more and more complicated, abstract things that correspond more and more to the way of a human reasoning. The architecture of the proposed model is shown in Figure 4.

Following is the description of the main modules of the proposed architecture.

4.1. Preprocessing

The unrelated information will be removed by applying preprocessing. Pima

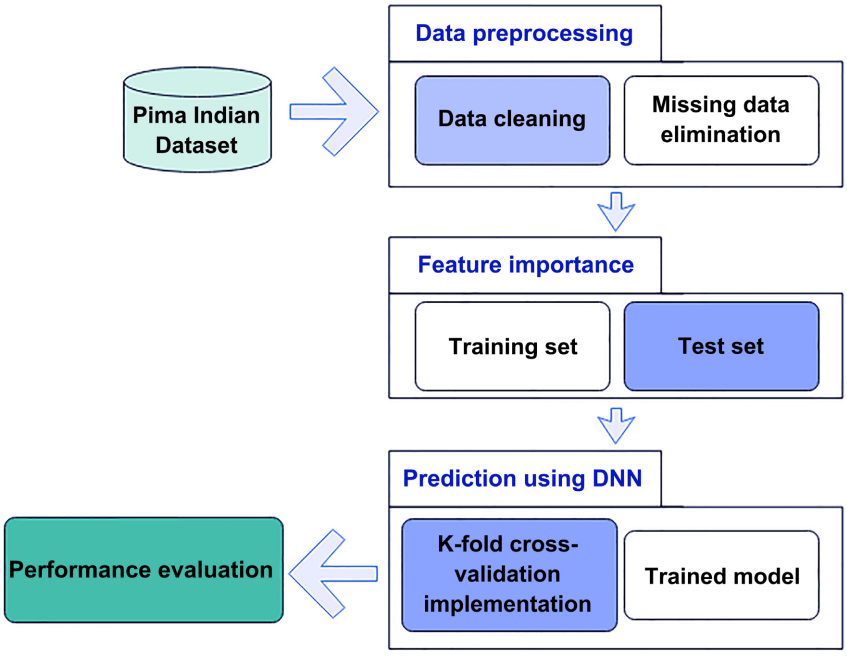

Figure 4. Architecture of the proposed model.

Indian Dataset (PID) consists of noise and empty entries, therefore preprocessing steps is used to reduce noise by eliminating redundant, empty, or any ambiguous data in the dataset. The dataset consists of 768 occurrences with nine attributes of features. It supports classification in a binary format that consists of 0 and 1 for diabetes positive and diabetes negative respectively. Diabetes positive is to identify persons with diabetes and diabetes negative is to identify persons without diabetes. In 768 in-stances, there are many entries with 0.

Multicollinearity is a condition where a predictor variable correlates with another predictor. Although multicollinearity doesn't affect the model's performance, it will affect the interpretability. If multicollinearity is not removed, it will be impossible know how much a variable contributes to the result. Figure 5 shows the correlation matrix of our model.

4.2. Feature Selection

Feature selection phase improves the Deep Learning process and increases the predictive power of algorithms by selecting the most important variables and eliminating redundant and irrelevant features.

4.3. k-Fold Cross-Validation

When a model is trained using all of the data in a single, it could not give the best performance accuracy. To resist, this k-fold cross-validation that helps us to

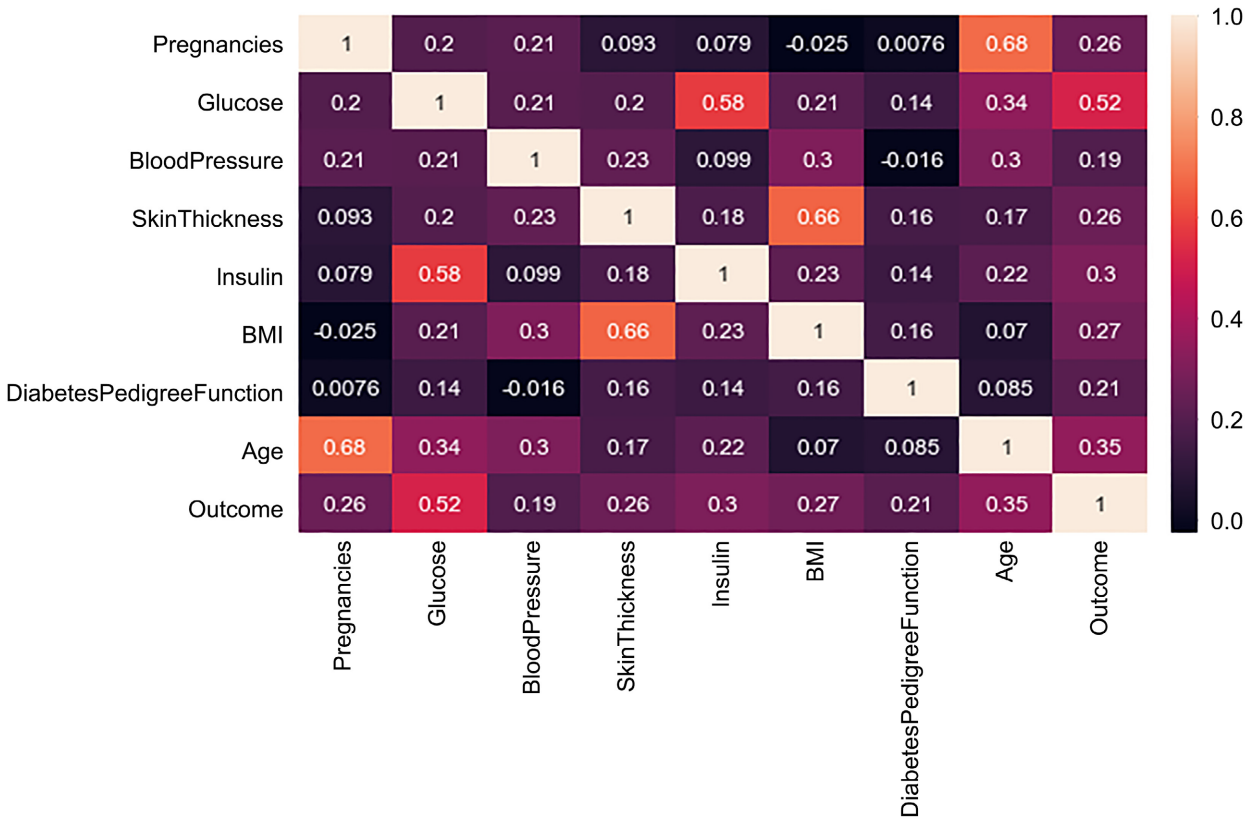

Figure 5. Correlation matrix.

build the model, is a generalized one. To achieve this k-Fold cross-validation, the dataset is divided into two sets: training and testing, with the challenge of the volume of the data. Here Test and Train dataset supported building model and hyperparameter assessments. The model has been validated multiple times based on the value assigned as a parameter and which is called k and it should be an INTEGER.

Based on the k value, the dataset would be divided, and train/testing will be conducted in a sequence way equal to k time. In this work k = 10.

4.4. Pima Indian Dataset

Pima Indian Diabetes dataset consists of 768 tuples of instances with nine feature attributes. The diabetes dataset is suitable for binary classification models, where it has only binary values like 0 and 1, *i.e.*, 0 for diabetes negative and 1 for diabetes positive. Out of nine (9) feature attributes from the dataset, only four (4) are considered for classification based on the feature Importance attribute score. The description of the PID dataset is shown in Figure 6.

5. Experimental Results and Comparative Analysis

5.1. Experimental Results

The early prediction of diabetes is crucial and obtaining higher accuracy rate in diabetes prediction is decisive. All experiments are run on GPU enabled TensorFlow with Keras. Figure 7 shows the feature importance of this model.

In experimental analysis, the pima indian dataset have been divided between 80% - 20% for training and testing purpose. We obtained 89% of accuracy on 10 k-fold cross-validation. The results of comparative analysis showed that DNN could be able to give best accuracy for diabetes prediction on pima indian dataset but using k-fold cross-validation on DNN model and pima indian dataset could decrease the accuracy of model.

Through the confusion matrix, presented in Figure 8, the performance of the DNN classifier is validated. Accuracy is the prior most metric used to know the performance and effectiveness of the model.

```
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
```

Figure 6. Description of PIMA Indian dataset.

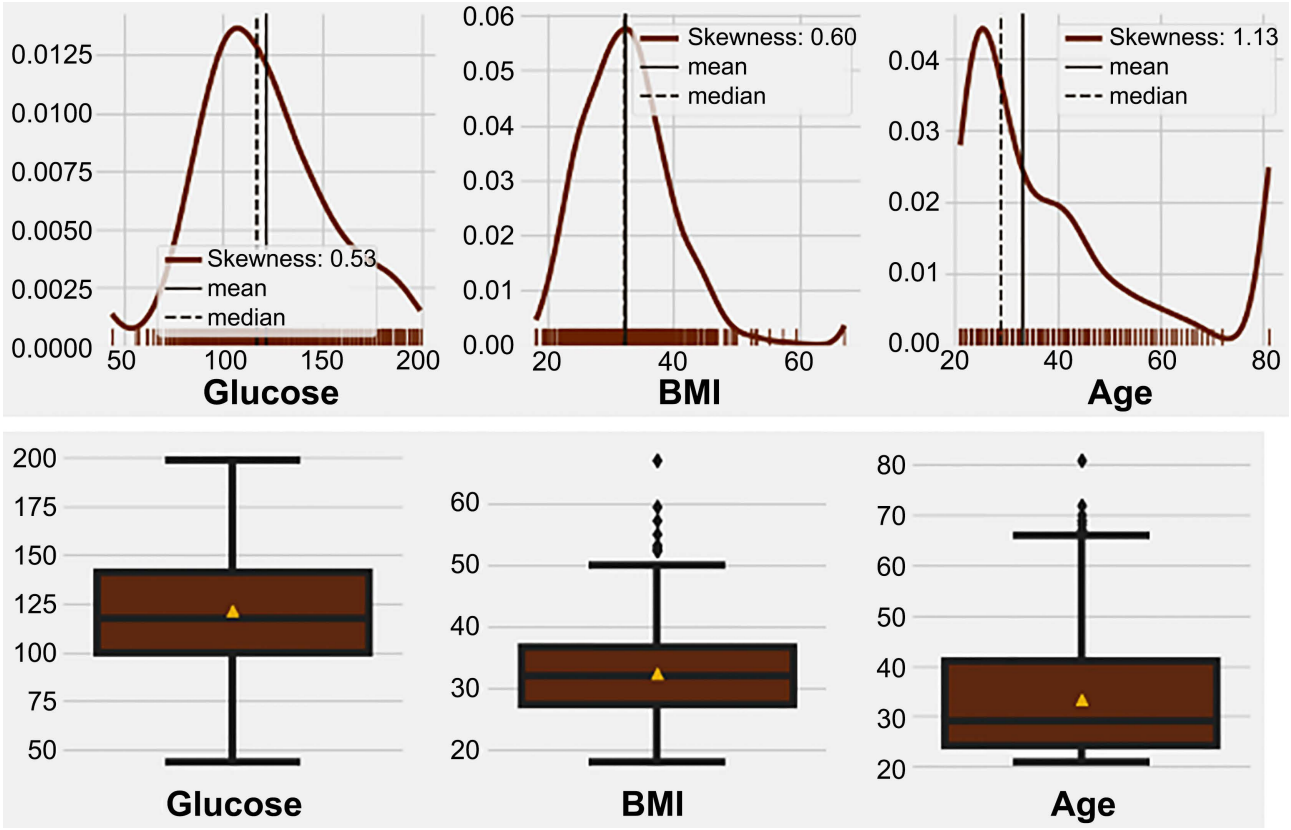

Figure 7. Feature importance.

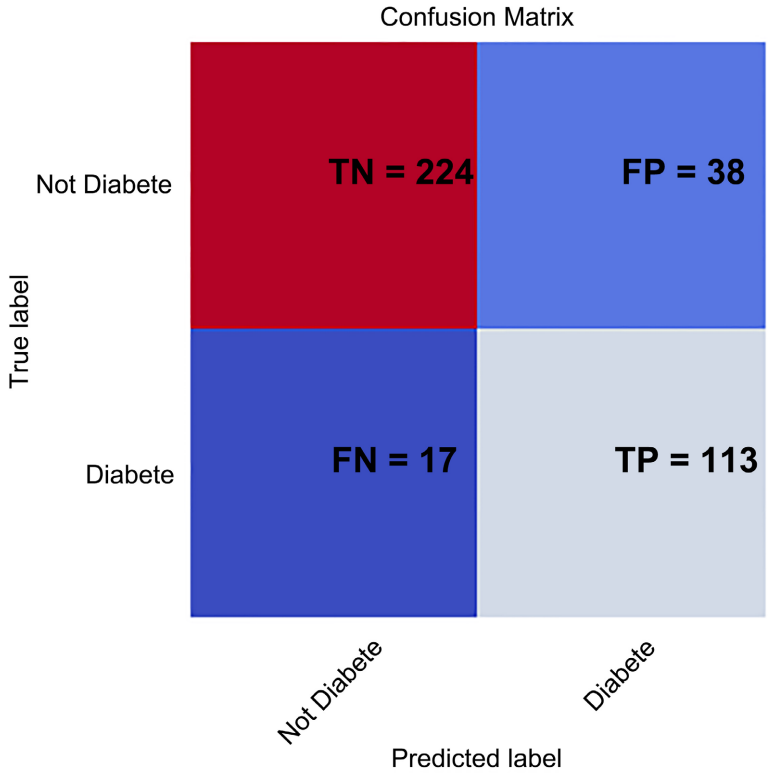

Figure 8. Confusion matrix.

The formula for calculating accuracy is (TP + TN)/(TP + FP + FN + TN) or all true positive and true negative cases divided by the number of all cases.

TN—True Negative, TP—True Positive, FP—False Positive, FN—False Negative.

- True Negative—Predicted negative, if person not affected by diabetes.
- True Positive—Predicted positive, if person affected by diabetes.
- False Positive—Predicted positive even if person not affected by diabetes.
- False Negative—Predicted negative even if person affected by diabetes.

5.2. Comparative Analysis

Our main objective is to compare DNN models which are built on Pima Indian Dataset and the model that we proposed in this work and prove the impact of 10 k-fold cross-validation. Figure 9 and Figure 10 present the results of comparative study of five models on diabetes prediction using DNN on pima dataset. Table 1 gives precisions on performance metrics values for each model.

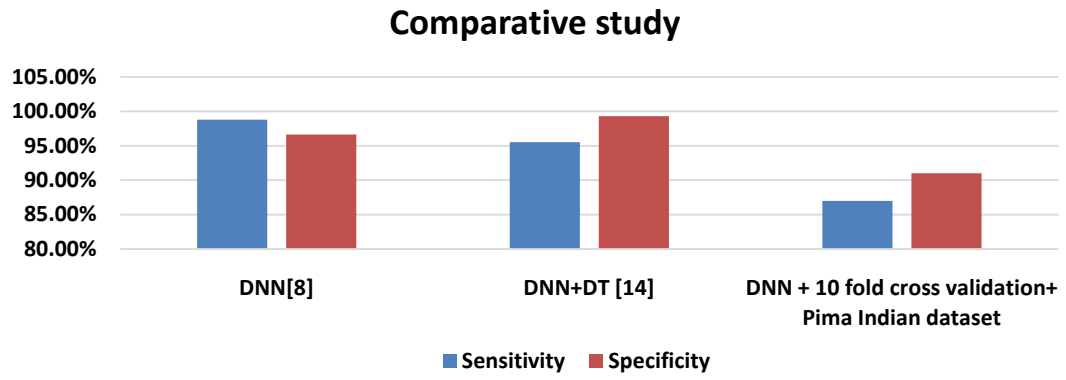

Figure 9. A comparative study with sensitivity and specificity metrics.

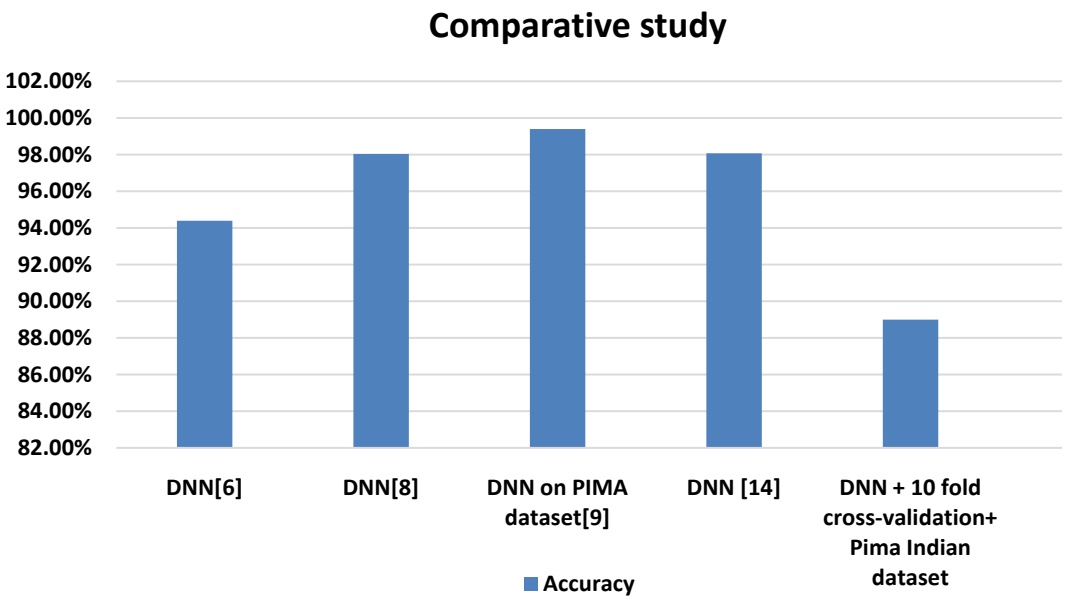

Figure 10. A comparative study with accuracy results.

Table 1. A Comparative analysis of DNN in terms of sensitivity, specificity and accuracy.

Models	Results
DNN [6]	Accuracy: 94.39%
DNN [8]	Sensitivity: 98.80%. Specificity: 96.64% Accuracy: 98.04
DNN on PIMA dataset [9]	Accuracy: 99.4%
DNN + DT [14]	Sensitivity: 95.52% Specificity: 99.29% Accuracy: 98.07%
DNN + 10 fold cross-validation + Pima Indian dataset	Sensitivity: 87% Specificity: 91% Accuracy: 89%

On **Figure 9**, the highest accuracy is seen on DNN + DT [14] but the highest sensitivity is in DNN [8]. However, specifivity and sensitivity of our proposed model, are less important on **Figure 9**. On other hand, **Figure 10** shows that DNN on PIMA dataset [9] is the most performant model with the highest accuracy between the five models. Our proposed model tested with k-fold cross-validation is the less efficient.

6. Discussions

Many authors who have worked on diabetes detection using Deep Learning have achieved good results. However, our concept is implemented on Pima Indian Diabetes Dataset (PIDD), Deep Neural Network and 10 k-fold cross-validation to evaluate the model. In classification tasks, the distribution of classes in the database can be unbalanced. k-fold cross-validation helped to better use available data of pima dataset. At the end of the procedure of k-fold corss validation, k performance scores are obtained, one per block. The mean and standard deviation of the k performance scores can be calculated to estimate the bias and variance of the validation performance. Results showed that 10 fold cross-validation can decrease the performance on models which are built into DNN and PIDD. In this study, comparative analysis of previous works on diabetes prediction using DNN, is provided in order to demonstrate how 10 k-fold cross-validation and DNN could decrease the performance of diabetes prediction models.

7. Conclusion

Diabetes is a world health problem that is approaching epidemic proportions globally. It's a challenge for scientists, doctors, medical and experts to predict the disease in early stages. The main reason for this is due to lack of awareness among underdeveloped and developing countries. Predicting the disease at an early stage and proper medication can save the life of a person. Deep Learning models diagnose all types of diseases with accurate results. In this paper, we did

a comparative analysis of different works on diabetes prediction using DNN. The results showed that diabetes detection using PIMA Indian dataset with k-fold cross-validation on pima could decrease the efficiency of the model with respect to using a model.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] World Health Organization (2022) Diabetes: Keys Facts. https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] Moustafa, Z. (2020) Évolutions de l'Intelligence Artificielle: Quels enjeux pour l'activité humaine et la relation Humain-Machine au travail? *Activites*, 1-39. https://doi.org/10.4000/activites.4941

[3] Machine Learning. Java T Point. https://www.javatpoint.com

[4] Pankajray (2021) Convolutional Neural Network (CNN) and Its Application—All You Need to Know. https://medium.com/analytics-vidhya/convolutional-neural-network-cnn-and-its-application-all-u-need-to-know-f29c1d51b3e5

[5] Dima, S. (2022) 5 Reasons Why You Should Use Cross-Validation in Your Data Science Projects. https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-project-8163311a1e79

[6] Mhaskar, H.N., Pereverzyev, S.V. and van der Walt, M.D. (2017) A Deep Learning Approach to Diabetic Blood Glucose Prediction. *Frontiers in Applied Mathematics and Statistics*, **3**, Article 14. https://doi.org/10.3389/fams.2017.00014

[7] Bala, M.K.P., Srinivasa, P.R., Nadesh, R.K. and Arivuselvan K. (2020) Type 2: Diabetes Mellitus Prediction Using Deep Neural Networks Classifier. *International Journal of Cognitive Computing in Engineering*, **1**, 55-61. https://doi.org/10.1016/j.ijcce.2020.10.002

[8] Islam, I.A. and Milon, M.I. (2019) Diabetes Prediction: A Deep Learning Approach. *International Journal of Information Engineering and Electronic Business*, **11**, 21-27. https://doi.org/10.5815/ijieeb.2019.02.03

[9] Zhou, H., Myrzashova, R. and Zheng, R. (2020) Diabetes Prediction Model Based on an Enhanced Deep Neural Network. *EURASIP Journal on Wireless Communications and Networking*, **2020**, Article No. 148. https://doi.org/10.1186/s13638-020-01765-7

[10] Pradhan, N., Rani, G., Dhaka, V.S. and Poonia, R.C. (2020) Diabetes Prediction Using Artificial Neural Network. In: Agarwal, B., Balas, V.E., Jain, L.C., Poonia, R.C. and Sharma, M., Eds., *Deep Learning Techniques for Biomedical and Health Informatics*, Academic Press, Cambridge, 327-339. https://doi.org/10.1016/B978-0-12-819061-6.00014-8

[11] Naz, H. and Ahuja, S. (2020) Deep Learning Approach for Diabetes Prediction Using PIMA Indian Dataset. *Journal of Diabetes & Metabolic Disorders*, **19**, 391-403. https://doi.org/10.1007/s40200-020-00520-5

[12] Swapna, G., Soman, K.P. and Vinayakumar, R. (2018) Automated Detection of Diabetes Using CNN and CNN-LSTM Network and heart Rate Signals. *Procedia*

Computer Science, **132**, 1253-1262. https://doi.org/10.1016/j.procs.2018.05.041

[13] Chowdary, P.B.K. and Kumar, R.U. (2021) An Effective Approach for Detecting Diabetes Using Deep Learning Techniques Based on Convolutional LSTM Networks. *International Journal of Advanced Computer Science and Applications*, **12**, 519-525. https://doi.org/10.14569/IJACSA.2021.0120466

[14] Pham, T., Tran, T., Phung, D. and Venkatesh, S. (2017) Predicting Healthcare Trajectories from Medical Records: A Deep Learning Approach. *Journal of Biomedical Informatics*, **69**, 218-229. https://doi.org/10.1016/j.jbi.2017.04.001

[15] Khanam, J.J. and Foo, S.Y. (2021) A Comparison of Machine Learning Algorithms for Diabetes Prediction. *ICT Express*, **7**, 432-439. https://doi.org/10.1016/j.icte.2021.02.004

[16] Tigga, N.P. and Garg, S. (2020) Prediction of Type 2 Diabetes Using Machine Learning Classification Methods. *Procedia Computer Science*, **167**, 706-716. https://doi.org/10.1016/j.procs.2020.03.336

[17] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I. (2017) Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, **15**, 104-116. https://doi.org/10.1016/j.csbj.2016.12.005

[18] Kaur, G. and Chhabra, A. (2014) Improved J48 Classification Algorithm for the Prediction of Diabetes. *International Journal of Computer Applications*, **98**, 13-17. https://doi.org/10.5120/17314-7433

[19] Kumari, S., Kumar, D. and Mittal, M. (2021) An Ensemble Approach for Classification and Prediction of Diabetes Mellitus Using Soft Voting Classifier. *International Journal of Cognitive Computing in Engineering*, **2**, 40-46. https://doi.org/10.1016/j.ijcce.2021.01.001

[20] Mat Jizat, J.A., Abdul Majeed, A.P.P., Ahmad, A.F., Taha, Z. and Yuen, E. (2021) Evaluation of the Machine Learning Classifier in Wafer Defects Classification. *ICT Express*, **7**, 535-539. https://doi.org/10.1016/j.icte.2021.04.007

[21] Sisodia, D. and Sisodia, D.S. (2018) Prediction of Diabetes Using Classification Algorithms. *Procedia Computer Science*, **132**, 1578-1585. https://doi.org/10.1016/j.procs.2018.05.122

[22] Allah, E.M.A., El-Matary, D.E., Eid, E.M. and El Dien, A.S.T. (2022) Performance Comparison of Various Machine Learning Approaches to Identify the Best One in Predicting Heart Disease. *Journal of Computer and Communications*, **10**, 1-18. https://doi.org/10.4236/jcc.2022.102001

[23] Fatima, M. and Pasha, M. (2017) Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications*, **9**, 1-16. https://doi.org/10.4236/jilsa.2017.91001

[24] Kalhori, S.R.N. and Zeng, X.-J. (2013) Evaluation and Comparison of Different Machine Learning Methods to Predict Outcome of Tuberculosis Treatment Course. *Journal of Intelligent Learning Systems and Applications*, **5**, 184-193. https://doi.org/10.4236/jilsa.2013.53020

[25] Talha, M.A., Muhammad, A.I., Yasir, A., Abdul W., Safdar I., Talha I.B., Ayaz H., Muhammad A.M., Muhammad M.R., Salman I. and Zunish A. (2019) A Model for Early Prediction of Diabetes. *Informatics in Medicine Unlocked*, **16**, Article ID: 100204. https://doi.org/10.1016/j.imu.2019.100204