Deep Neural Network to Predict Diabetes: A Data Science Approach

Mafas Raheem

Abstract*: *Diabetes has become a famous and lethal disease among the low and medium-income countries. People could not overcome this deadly abnormal condition due to the current lifestyle, food habit and the genetic transmittance. Medical practitioners provide advice to prevent the diabetic condition and medications to control as this disease does not have a permanent cure. However, the detection of the disease is being a tidy process and deployment of machine learning predictive models to conduct smart diagnosis/detection is vital in the healthcare domain nowadays. Though several machine learning models were built in this regard, deploying a Deep Neural Network seems less focused. Therefore, a Deep Neural Network model was built with the support of complete preprocessing, class balancing, normalization, feature selection process and hyper-parameter tuning using the cross-validated searching technique. The model achieved 88% of accuracy and 0.88 ROC score and standing out as a promising predictive model in diagnosing/detecting diabetes.

Keywords*: *diabetes, healthcare, predictive modelling, deep neural network, optimization

I. INTRODUCTION

Diabetes refers to reduce the ability of the human body to regulate the level of glucose in the blood [1]. Diabetes is a chronic, metabolic disease which occurs with the elevated blood glucose (or blood sugar) level, hence ultimately leads to other health problem or serious damage to the heart, blood vessels, eyes, kidneys and nerves. Meanwhile, the long-term effects may result in coma, renal failure and retina failure, pathological destruction of pancreatic beta cells, cardiovascular disease, sexual dysfunction, joint failure, weight loss, ulcer, and pathogenic effects on immunity. There is no long-term cure, can only be done with control and prevention. Type 2 diabetes is considered as the most common among adults, which occurs when the body shows insulin resistance or doesn't produce enough insulin due to defects in Pancreases Beta-cell. About 422 million people are suffering from diabetes worldwide among which the majority are found in low and medium-income countries [2]. Further, 1.6 million deaths occur due to diabetes each year where the number of cases and the prevalence of diabetes have been steadily increasing over the past few decades worldwide [2].

Data mining techniques have been applied widely in healthcare for effective management as well as diagnosis [3]. Intense research is being carried out to build a machine learning predictive model that could learn from historical data and deliver smart diagnosis/detection of diabetes. An effective predictive model has some advantage over human diagnosis where it might be affected by human error, fatigue, stress etc. For example, a machine learning algorithm can learn from complex data, the performance of a machine is consistent compare to a human at full employment condition.

In the past, several predictive models have been implemented for predicting diabetes using various machine learning algorithms. In this line, most of the researchers used algorithms such as Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), k-Nearest Neighbor (k-NN) and Naive Bayes (NB) to build predictive models. However, the deployment of an Artificial Neural Network (ANN) or Multi-Layer Perceptron (MLP) or Hybrid form of Artificial Neural Network or Deep Neural Network (DNN) seems rare in this regard. The main objective of this paper is to build an optimized Deep Neural Network (DNN) model to predict diabetes using historical data. This is an end to end work that addresses optimality in hyperparameter tuning and model parameter selection using suitable optimization techniques. The paper has been structured chronologically where the survey of the past research works is covered in section 2. Section 3 covers the methodology followed in this study and Section 4 describes the dataset, preprocessing, class balancing and the experimental results. Finally, the conclusion including the future work is discussed in Section 5.

II. LITERATURE REVIEW

The application of data mining techniques has become more rigorous in the healthcare domain from the recent past [3]. Some examples of the application of diagnosis include breast cancer [4] [5], chest disease [6], skin disease [7] and categorize smoking patterns of older adults [8]. A Cooperative Coevolutionary Model (COVNET) for Evolving Artificial Neural Networks with the accuracy of 80.1% was implemented to predict different diseases including diabetes using Pima Indian data set [9]. Consequently, a hybrid neural network that includes Artificial Neural Network (ANN) and Fuzzy Neural Network (FNN) was implemented with an accuracy of 84.2% to detect diabetes [10]. A multilayer perceptron (MLP), Radial Basis Function (RBF) and General Regression Neural Network (GRNN) was built and the highest accuracy was obtained by GRNN which was 80.21% [11].

Manuscript received on Januar y15, 2021.
Revised Manuscript received on February 09, 2021.
Manuscript published on March 30, 2021.
* Correspondence Author

Mafas Raheem*, School of Computing, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia. Email: rmafas@gamil.com

Reference [12] built a Multi-Layer Perceptron (MLP) with preprocessing techniques applied such as missing values handling using median and feature selection while tested the models using k-folds (2,4,5,10) cross-validations and MLP classifier gave the highest accuracy of 78.7% (k = 4).
An Artificial Neural Network (ANN) was built by reference [13] where the model achieved 86% accuracy with preprocessing done, but no specific indication of class balancing while specifying single tuning. Reference [14] was built an Artificial Neural Network with preprocessing techniques applied and obtained 74.74% accuracy on Pima Indian Diabetes dataset.In a study conducted by reference [15], a Recursive General Regression Neural Network (R-GRNN) Oracle was built with cross-validation and achieved the accuracy, AUC, and sensitivity at 81.14%, 86.03%, and 63.80% respectively. Reference [16] built a deep learning model using Restricted Boltzmann Machine (RBM) with feature selection process using Random Forest and normalized through min-max normalization and obtained the accuracy of 85.5% on the Pima Indian Diabetes data set. On the other hand, reference [17] proposed an approach of Feedforward Neural Network called Extreme Learning Machine (ELM) for single hidden layer feedforward neural network (SLFNS) which randomly chooses the input weights and analytically determines the output weights SLFNS. With 20 number of neurons, the model achieved 76.54% of accuracy score. Further, a Multilayer Neural Network (MLNN) structure which was trained by Levenberg-Marquart (LM) Algorithm and Probabilistic Neural Network (PNN) structure was built where the MLNN with LM model obtained 62% accuracy (conventional valid), while PNN obtained 78.13% accuracy (conventional valid) [18]. Reference [19] implemented a Neural Network (NN) model combined with Self Organised Maps (SOM) and Principal Component Analysis (PCA) for clustering and noise removal respectively. The experiment resulted in 92.28% accuracy score. Based on the literature reviews, many research works were found with the ANN implementation on Pima Indian Diabetes dataset where mentions of the complete preprocessing, class balancing, feature selection and optimization methods were not significant. Furthermore, no evidence found for handling the zero lower level exists in the input variables. In this paper, the most suitable preprocessing and optimization techniques in finding the most suitable hyperparameters were used to build a more lucrative DNN predictive model.

III. METHODOLOGY

The healthcare domain has started using the machine learning models widely to get support in predicting the specific diseases among which diabetes is in the frontline. The health practitioners strongly believe the support of data mining and predictive modelling in predicting this disease is highly crucial.

A. Data set

The Pima Indian Diabetes dataset was obtained from Kaggle which consists of 9 features and 768 records and the details are tabulated in Table- I.

Table- I: Data set description

Name	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)
Outcome	Class variable (0 or 1)

B. Pre-Processing

A dataset usually contains noise such as missing values, outliers and so on. Many predictive machine learning algorithms are sensitive to these noises and may mislead the final results of the model. Therefore, suitable strategies should be used to clean the data before building the most profitable predictive model. Further, the class balancing should be done in the target variable to overcome the biases in model building.

C. Deep Neural Network (DNN)

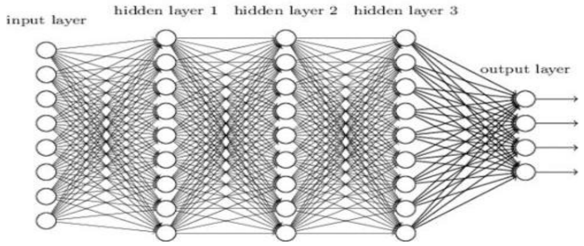


Fig. 1.Deep Neural Network [20]

A DNN is a form of an artificial neural network (ANN) with multiple hidden layers between the input and output layers which always consist of the same components: neurons, synapses, weights, biases, and functions as depicted in Fig. 1. Building a deep neural network as a predictive model is usually a crucial process where it is usually expected to give better accuracy values. At the same time, fine-tuning the deep neural network on the specific hyperparameters seems critical in building a more profitable and unbiased predictive model.

IV. EXPERIMENTS

The Deep Neural Network model was built according to the following process as depicted in Fig. 2.

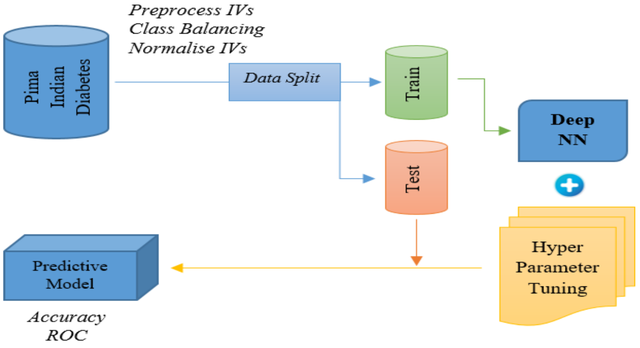


Fig. 2.The process flow diagram

A. Preprocessing

According to the dataset, there were no missing values found. However, there were many 'zero' values found in certain variables which could not be accepted and those values were replaced by missing due to the following reasons.

- BloodPressure: Living person cannot have 0 blood pressure
- Glucose Level: Even fasting people won't have 0 glucose level
- Skin Thickness; Normal people skin thickness can't be less than 10mm
- BMI: Should not be zero unless serious sick
- Insulin: Very rare case that people will have 0 insulin.

The reasons were drafted based on general biological knowledge. The missing values were then imputed using median values. The data normalization/scaling is always expected to improve the performance of the neural network models. Further, the variable "Insulin" was found with outliers and found by comparing the *mean* and the *standard deviation*. In this line, the input variables were normalized before splitting it as train (80%) and test (20%).

B. Feature Selection

Feature selection is yet another important process in building a promising predictive model. The input variables better be checked and selected with the most significant variables even though the variables seem important towards predicting the chosen target. In this line, the *VarianceThreshold* function from Sklearn was used to detect the significance of the input variables and validated using the correlation heat map as given in Fig. 3.

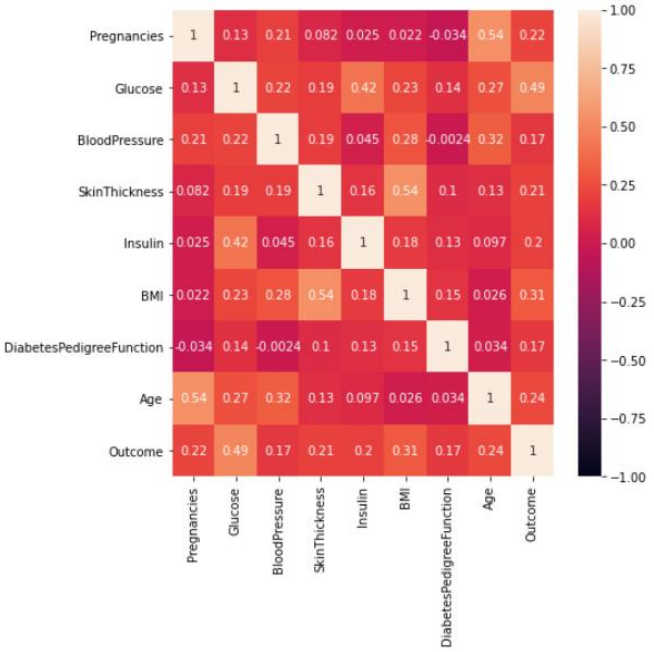
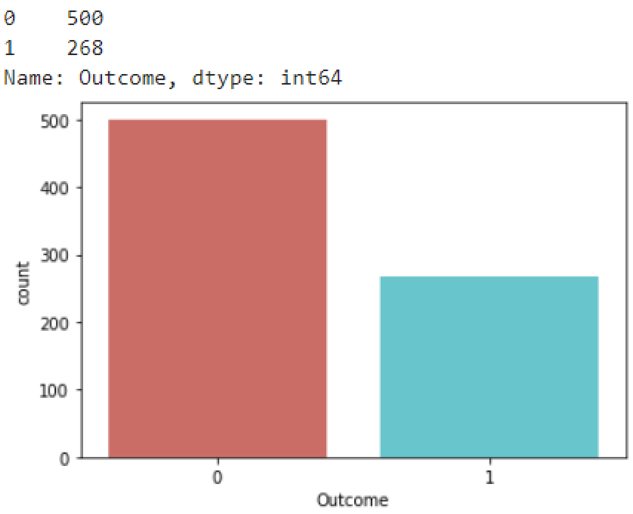


Fig. 3.Correlation heat map

Based on the correlation, it was found that no high correlated input variables were found in the dataset and no negative correlated input variables with the target were also found in the dataset. Therefore, it was decided to include all the input variables to build the DNN model.

C. Class balancing

As far as predictive modelling is concerned, an imbalanced class data is a general occurrence and usually gives a serious impact on the final results.

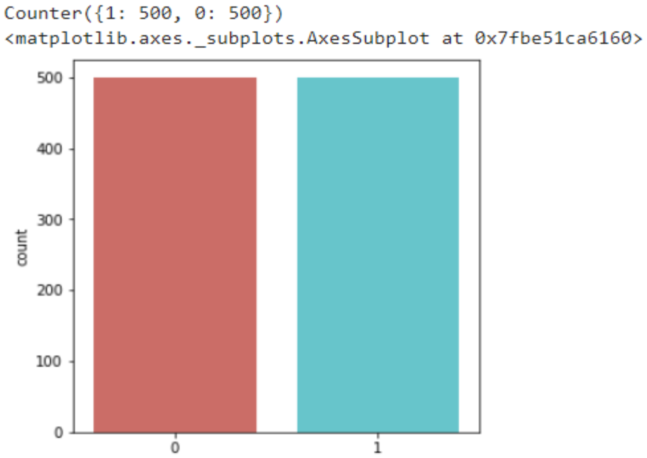


```
0    500
1    268
Name: Outcome, dtype: int64
```

Fig. 4: Imbalanced class – Target variable

Consequently, one of the two techniques such as over-sampling and under-sampling is generally applied by the researchers in balancing the class. Mostly, the over-sampling techniques are widely applied to balance the class data where the most critical information will be contained during the processing activity via this technique. The target variable in this data set is named as "Outcome" and found imbalanced as shown in Fig. 4.

An over-sampling technique named Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the target variable where the samples usually added synthetically by using the nearest neighbour as shown in Fig. 5.



```
Counter({1: 500, 0: 500})
<matplotlib.axes._subplots.AxesSubplot at 0x7fbe51ca6160>
```

Fig. 5: Balanced class – Target variable

D. Normalization

The raw data can have two main issues such as dominant features and outliers which could hinder the learning of machine learning algorithms. The data normalization is an important operation which either transforms or rescales the raw data to get uniform contributions from each feature [21].

The data set chosen was not found with any outliers, hence normalized to get uniform contributions to train the DNN and to get more effective results.

E. DNN with Hyper-parameter tuning

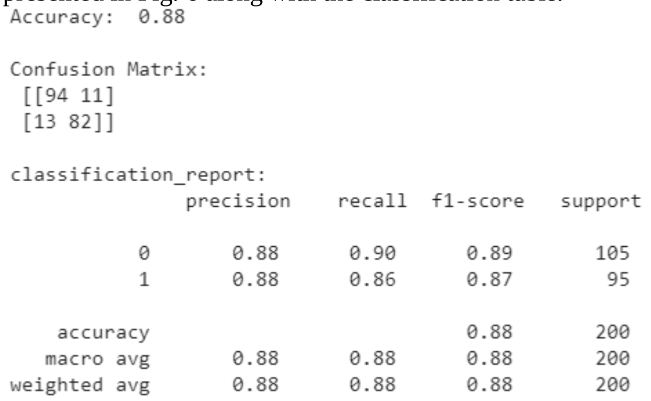
Deep Neural Networks are generally built with the default hyperparameters. The accuracy score of the DNN with the default parameters highly depends on the data set and its contribution towards the learning. However, tuning the suitable hyperparameters along with the suitably rescaled data would give a better output which could be acceptable in any domain. In this line, the architecture of the DNN and the hyper-parameters were taken into consideration to build and tune the model. Table- II showcases the finalised architecture and the hyper-parameters of the DNN model. The model obtained a good accuracy score based on the tuning of the hyper-parameters.

Table- II: DNN Architecture and Hyper-Parameters

Architecture	Input layer = 1 (input_dim = 8)	
	Hidden Layer (HL) = 3	
	Neurons	HL 1 = 300
		HL 2 = 200
		HL 3 = 100
	Output Layer = 1 (Neurons = 1)	
Hyper-parameters		
kernel_initializer	lecun_uniform	
activation	Hidden Layer	relu
	Output Layer	sigmoid
optimizer	adam	
loss	binary_crossentropy	
metrics	accuracy	
learn_rate	0.2	
momentum	0.6	
dropout_rate	0.3	
weight_constraint	3	

The hyper-parameters were set and searched using GridSearchCV where the final score was obtained via cross-validation, thus the final accuracy score can be reliable as it is cross-validated using the test data.

The final DNN model gave a promising accuracy score as presented in Fig. 6 along with the classification table.



```
Accuracy:  0.88

Confusion Matrix:
 [[94 11]
 [13 82]]

classification_report:
              precision    recall  f1-score   support

           0       0.88      0.90      0.89       105
           1       0.88      0.86      0.87        95

    accuracy                           0.88       200
   macro avg       0.88      0.88      0.88       200
weighted avg       0.88      0.88      0.88       200
```

Fig. 6.Accuracy Score & Classification report of DNN

Fig. 7 depicts the confusion matrix of the classifier and Fig. 8 depicts the Receiver Operating Characteristics (ROC) curve with the index of 0.88.

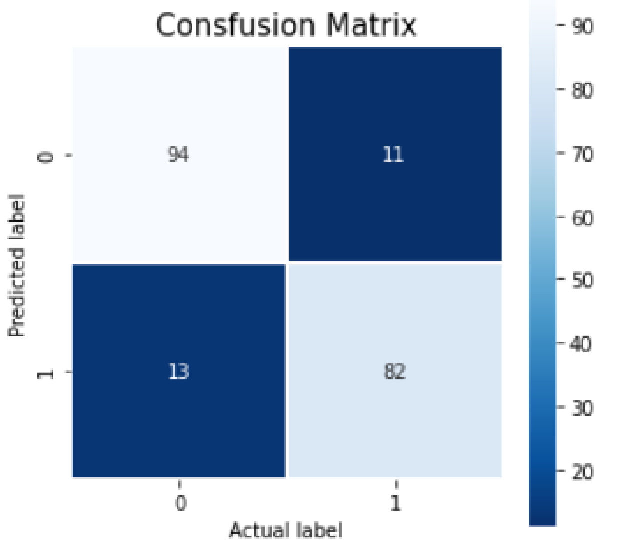


Fig 7: Confusion matrix

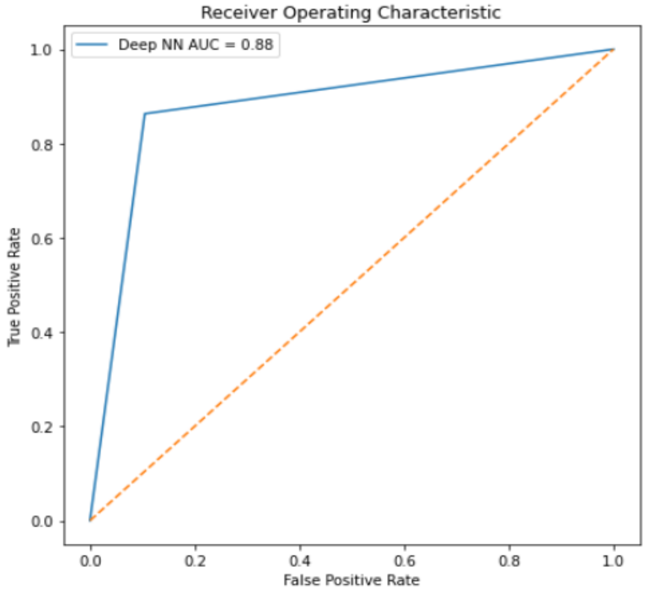


Fig. 8: ROC Curve

V. CONCLUSION

ANN is a black box technique where the algorithm can approximate any function, hence understanding its architecture won't reveal any insights on the structure of the function being approximated. Further, ANN does not provide an easy interpretation of the results except the final score depending on the nature of the model. In this line, building a DNN with many hidden layers and tuning it to obtain a promising results in the healthcare domain is challenging too. As the title expresses, a predictive model using DNN was built with optimisation techniques supported by suitable preprocessing strategies. Though the process is time consuming in searching the effective hyper-parameters using the GridSearchCV and tuning the model, the results seem promising and reliable as it is cross-validated. The obtained accuracy score (88%) seems promising than the past works gathered via the literature review. The results are tabulated in Table- III.

Table- III: Accuracy score comparison - PIMA Indian Diabetes data

Models	Accuracy Score (%)
COVNET [9]	80.10
ANN + FNN [10]	84.21
GRNN [11]	80.21
MLP [12]	78.70
ANN [13]	86.00
ANN [14]	74.74
R-GRNN [15]	81.14
RBM [16]	85.50
ELM + SLFNS [17]	76.54
MLNN + LM [18]	62.00
FNN [18]	78.13
ANN + PCA + SOM [19]	92.28
DNN [this research]*	**88.00**

* *Input variables preprocessing*
Input variables normalization
Target Variable class balancing
Complete Hyper-parameter tuning

According to Table 3, the DNN model with optimization achieved an accuracy score of 88%. However, more research can be done with the increasing number of records which may offer a room to get a more accurate model. Further, the interest of building a more accurate predictive model can be given towards the deep learning architectures to which a high number of records is a mandatory requirement.

REFERENCES

1. K. Kaul, M. T. Joanna, I. A. Shamim, M. K. Eva, and C. Rakesh, “Introduction to Diabetes Mellitus. Diabetes: An Old Disease, a New Insight,” *Springer New York CY - New York*, 2012.
2. WHO. (2021). Diabetes [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1 [Accessed: 2 January 2021].
3. I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J. F. Chang, and L. Hua, “Data mining in healthcare and biomedicine: A survey of the literature,” *Journal of Medical Systems*, 36(4), Aug 2012, pp. 2431-48.
4. J. Tyrer, S. W. Duffy, J. Cuzick, “A breast cancer prediction model incorporating familial and personal risk factors,” *Stat Med*, 15 Apr 2004, 23(7), pp. 1111-1130.
5. Er. Orhan, N. Yumuşak and F. Temurtas, “Chest diseases diagnosis using artificial neural networks,” *Expert Systems with Applications*, 37, 2010, pp. 7648-7655.
6. C. Chang, and C. Chen, “Applying decision tree and neural network to increase quality of dermatologic diagnosis,” *Expert Systems with Applications*, 2009, 36, pp. 4035-4041.
7. S, Moon, S, Kang, W. Jitpitaklert, and S. B. Kim, “Decision tree models for characterizing smoking patterns of older adults,” *Expert Systems with Applications*, 2012, 39, pp. 445-451.
8. H. S. Ruchlin, “An Analysis of Smoking Patterns among Older Adults,” *Med Care*, 1999, 37, pp. 615 – 619.
9. J. Muñoz-Pérez, C. Martínez, and N. García-Pedrajas, “COVNET: A cooperative coevolutionary model for evolving artificial neural networks,” *Ieee Transactions on Neural Networks*, 2003, 14(3), pp. 575-596.
10. H. Kahramanli, and N. Allahverdi, “Design of a hybrid system for the diabetes and heart diseases,” *Expert Systems with Applications*, 2008, 35, pp. 82-89.
11. K. Kayaer, and T. Yildirim, “Medical diagnosis on Pima Indian diabetes using general regression neural networks,” *Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing*, 2003.
12. R. Ahuja, S. C. Sharma, and M. Ali, “A Diabetic Disease Prediction Model Based on Classification Algorithms,” *Annals of Emerging Technologies in Computing (AETiC)*, 2019, 3(3), pp. 44-52.
13. K. Harleen, and V, Kumari, “Predictive modelling and analytics for diabetes using a machine learning approach,” *Saudi Computer Science*, 2018, 12, pp. 1-6.
14. S. Bashir, U. Qamar, and F. H. Khan, “IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework,” *Journal of Biomedical Informatics*, 59, 2016, pp. 185-200.
15. D. Bani-Hani, P. Patel, and T. Alshaikh,(2019). An Optimized Recursive General Regression Neural Network Oracle for the Prediction and Diagnosis of Diabetes. Global Journal of Computer Science and Technology. 2019, 19(2). pp. 1-11.
16. R. Sushant, H. Balaji, N. Ch. S. N, Iyengar and D. C. Ronnie, “Optimal Predictive analytics of Pima Diabetics using Deep Learning,” *International Journal of Database Theory and Application*, 2017, 10(9), pp. 47-62.
17. H. Guang-Bin, Z. Qin-Yu and S. Chee, “Extreme learning machine: A new learning scheme of feedforward neural networks,” *IEEE International Conference on Neural Networks - Conference Proceedings*, 2004, 2. pp. 985-990.
18. H. Temurtas, N. Yumusak and F. Temurtas, “A comparative study on diabetes disease diagnosis using neural networks,” *Expert Systems with Applications*, 2009, 36, pp. 8610-8615.
19. M. Nilashi, O. Ibrahim, M. Dalvi, H. Ahmadi, and L. Shahmoradi, “Accuracy Improvement for Diabetes Disease Classification: A case on a Public Medical Dataset,” *Fuzzy Information and Engineering*, 2017, 9(8), pp. 245-257.
20. KDnuggets. 2021. *Deep Neural Networks*. [Online] Available at: https://www.kdnuggets.com/2020/02/deep-neural-networks.html.
21. S. Dalwinder, S. Birmohan, “Investigating the impact of data normalization on classification performance,” *Applied Soft Computing*, 2020, 97.

AUTHORS PROFILE

Mafas Raheem - Data Scientist | Business Analyst Mafas is an academic specialized in the field of Data Science & Business Analytics with nearly 15 years of academic & industry experience. He holds an MSc in Data Science & Business Analytics and a Master of Business Administration degree and reading his PhD in the area of Machine Learning/Text analytics. Currently, he works as an academic at the Asia Pacific University of Technology & Innovation, Malaysia. His research areas are business intelligence, visual analytics, predictive analytics, text analytics & sentiment analysis in various domains. He has published a significant number of journal articles in the area of data analytics and machine learning.

.