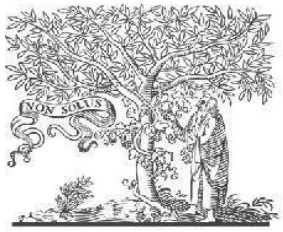
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 8th Aug 2024. Link

https://ijiemr.org/downloads.php?vol=Volume-13& issue=issue08

DOI: 10.48047/IJIEMR/V13/ISSUE 08/14

Title Diabetic Disease Prediction in Women using Deep Learning Techniques

Volume 13, ISSUE 08, Pages: 107 - 112

Paper Authors
Dr. G.V. Ramesh Babu , N. Vamsi Krishna

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper as Per UGC Guidelines We Are Providing A ElectronicBar code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Diabetic Disease Prediction in Women using Deep Learning Techniques

Dr. G.V. Ramesh Babu
Associate Professor, Department of Computer Science, Sri Venkateswara University, Tirupati
gvrameshbabu74@gmail.com
N. Vamsi Krishna
Master of Computer Applications,
Sri Venkateswara University, Tirupati
vamsikrishna911719@gmail.com

Abstract
In India irrespective of age most of the people are suffering from diabetics. A deep learning neural network is designed for early prediction of type-2 diabetics. Early prediction of disease helps the patient to take precautions from the life risk parameters. With the advancement of technology, medical fields takes the help of the artificial intelligent systems to have best predictions about various diseases and for recommending the medicines timely based on various parameters. The proposed system works better than the ensemble mechanism and gives more accurate results. The system also proved that K- Cross Fold validation with more splits.

Keywords: CNN, Heatmaps, Activation Functions, Cross Fold Validation

Introduction

Now days, Diabetes has become most common problem among any age group people. There are 3 types of diabetes namely Type-1 DM, Type-2 DM and Gestational diabetes and most of the surveys reported that Type-2 DM occurs more than other two types. The main reason for Type-2 DM is obesity and lack of physical exercises. If Type-2 DM is not identified or not treated for a long time, it may cause severe problems that may lead to life risk situation for a patient[4]. The persons suffering with Type-2 diabetes will have higher glucose levels. Applying machine learning approaches helps the doctors to diagnose the patients at their early stages and saves them from severe complications.

Literature Survey:

[1] Park J and Edington had implemented a model based on neural network-NN using the health risk management and got the merit from HRA. From past 3 years on wards like 1996-1998 I have been taken the initiative to generate a new proposed algorithm based on diabetic and I have been measuring so many patients' data (6142). Here in this I have proposed a methodology using the Sequential multi-layer perceptron- SMLP it has a benefit quality of back propagation and it is an flexible model with time varying inputs along with sequential data as well as the obtained predicted data with at most probability at the time of training using the multivariate logistic function in the embedding system has been proposed. As per the consecutive years, this paper gets the time sensitive data with the association of risk to predict that we have to make some changes based on that and here the repentance of the diabetics has been correspondingly rises with time period. This model outperforms and baseline method classification and as well as regression method which can gain the accuracy of 86.04 % for test data. Here the

result obtained for time sensitive data is predicted.

[2] Wang and Chen had implemented a model based on KNN, Random forest and here, we developed a batch learning algorithm called Xtreme Gradient Boosting algorithm to prevent that risk of chronic diseases and comparing to SVM and Random Forest as well as K Nearest Neighbor approach is used for current model with predictions. The combo is perfect. In Xicheng Area, Beijing, simple random sampling method and are snowball sampling they were utilized for conducting the survey. Personal information, daily habits, fitness level, plus family history are all part of a structured questionnaire. A total of 380 early part and older persons took part in the study. The models were then training and the illness was obtained. The 10 fold with a cross-validation each and every sample risk with the index is calculated. Tests were conducted to evaluate the data. We tested the above-mentioned widely used ML methods and discovered that XGBoost outperformed them all. The best analysis with such an overall accuracy is 0.89 as well as the surface with the recipient's and its hypothesis, the AUC curve (area under the curve) we got is 0.9182. As a result, XGBoost is better due because of its structure. Having superior predicted accuracy plus general capabilities when compare to other algorithms at time of forecasting the main hurdle of chronic diseases, is beneficial with preventing the management and the disease. Diabetes will be a future problem.

[3] Park J and Edington had a model based on SMLP in our previous study, we used reversed processing and learnt structure to extract the aspects of diabetes health with risk development in Sequence Multi Layered Perception. Here the risks trajectory and as conditional mixture models were used to evaluate the time-varying danger progress. The total risk decrease, and also the prediction, remained steady throughout time, as well as a higher body mass index has seems with the health behavior risk factor that most accurately predicts the development of diabetes. Excessive BMI obesity, hypertension (BP), hyperlipidemia, and a fat bastard diet were all important factors in the initial forecast. The changes in BMI, stress, blood pressure, cholesterol, especially fatty food consumption caused differences in trajectory across time. We used to implement SMLP on such a test sample of an employee from a huge manufacturing organization where earlier workplace health promotion was launched to see if it is useful for identifying prediabetes in 1984. As a result, there was a possibility of sensitivity of 71 percentages. Before that there is an issue with the mapping the data and it take the large time to compile. By mixing the target variables with some glucose tolerance test the with predicted data and model will be able to implement and it shows the effectiveness of this paper.

[4] Zhou and H Myrzashova had implemented a model based on DNN here diabetes is becoming one of the world largest most frequent and dangerous disease it has some serious complications. In early stage it has been detected to receive the treatment as well as prevent all the conditions which is happening. Not only proposed methods have been used to forecast the incidence of diabetes. In future, we have to face so many problems based on this disease and we can also be determined what kind of diabetes that a people has. Based on that we have to give the significant changes in treatment procedures between which types of diabetes they have, this strategy will benefit in producing the great treatment to the patients. This methodology usually created using the hidden units the deep neural network it has used dropout regularization to less per by changing the method into such classification methods has some issue. We tweaked a some parameters that have been using a binary method for cross-entropy nonlinear function has been created with large number of accuracy using deep neural network forecasting models. This experiments introduces that the efficiency of developed Deep Learning method. The patients with that type data has been trained with sample data set and gives the best

accuracy percentage of 94.02, and where as the data set which is taken from Pima Indians diabetes data set. Ithas high accuracy of 99.41 percent at time of training. The experiments tell us that our implemented model outperforms current methods.

Table 1: Existing Systems Comparisons

S.No	Author Name	Algorithms Used	Merits	Demerits	Accurate Percentage
1	Park J and Edington	ANN,SMLP	Comparing with real time existing data	Unable to crack the hidden input patterns.	
2	Wang and Chen	SVM, KNN, RF XGBoost	Avoid overfitting the data and it has some spare trained data	Preventing the occurrences of the data.	0.89%
3	Park J and Edington	SMLP	Multiple data can be taken and be predicted.	Here the point percentage has been risk taken.	
4	Zhou and H Myrzashova	RNN,CNN,SVM,DNN	Sequence of process has been executed and depending on one output to next input.	Here it has used a common variants of data.it can only say that the person has diabetics but can't say what stage he/she has.	94.02%

Materials and Methods:

The proposed paper collects the data from kaggle about diabetic patients[5] , which consists of 8 attributes and 1 class attribute which represents whether a person is diabetic or not. The dataset contains information about 768 people. Among 768 people, 268 people are diabetic and 500 people are non-diabetic. The attributes are represented as described in the Table 2 [4]

Table 2: Meta Data about Diabetic Patients

S.no	Attribute Name	Description	Range
1	Pregnancy	Number of times a participant is pregnant	0–17
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	0–199
3	Blood pressure	It consists of Diastolic blood pressure (when blood exerts into arteries between heart)(mm Hg)	0–122
4	Skin Thickness	Triceps skinfold thickness (mm).It concluded by the collagen content	0–99
5	Insulin	2-Hour serum insulin (mu U/mL)	0–846
6	BMI	Body mass index (weight in kg/(height in m)^2)	0–67.1
7	Diabetes pedigree Function	An appealing attributed used in diabetes prognosis	0.078–2.42
8	Age	Age of participants	21-81
9	Outcome	Diabetes class variable, 1 represent the patient is diabetic and 0 represent patient is not diabetic	1/0

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

The dataset contains 8 attributes, but all these attributes may not be important. To find the relation among the attributes, we can find the correlation between the attributes using heatmap and we can visualize the data as shown in Figure1. The correlation mechanism helps in finding the relation among various attributes. It plays an vital role in feature selection phase. The popular coefficient is "Pearson coefficient", which always generates the value in between -1 to +1. If the method produces positive value close to +1 then those attributes are very strongly related and if it generates 0 then those attributes are not correlated. In the PIMA dataset, the system observed that Glucose, Blood pressure, BMI and Age are the attributes which are positively correlated with all other attributes. So, these attributes might be considered as the important attributes.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

Figure 1: Correlation between Diabetes Attributes

Based on the class label "Outcome", the proposed system decided to supply supervised machine learning classification algorithms for predicting the Type-2 diabetes. The proposed system considered most popular 3 Algorithms. They are: Decision Tree Algorithm [7], represents all the attributes as internal nodes and external nodes represent class label. The algorithms generate decision rules to predict the outcome. The sample representation for the PIMA dataset is shown in the Figure2. The algorithm calculates the average of the first node i.e., pregnancy, which is approximately 4. Like this, for all the attributes the mean is calculated and rules are generated.

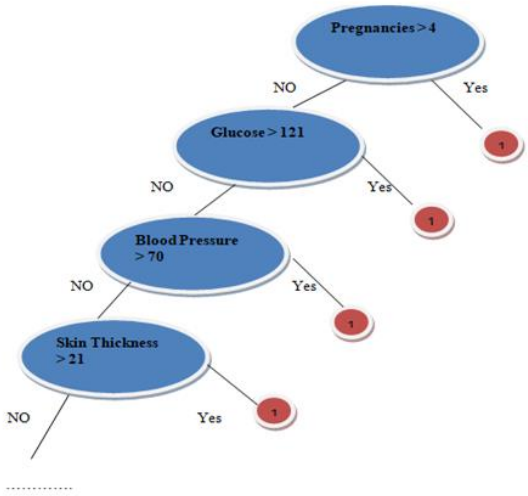


Figure 2: A Sample Construction of Decision Tree for Diabetic Dataset

The proposed system also considers KNN [9] algorithm, the k value is considered as 5, now the system calculates the Euclidean distance by considering all the attributes of training data with the values, which the

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

system want to predict as shown in the Equation-1.

$$= \sqrt{(6-1)^2 + (148-85)^2 + (72-66)^2 + (35-29)^2 + (0-0)^2 + (33.6-26.6)^2 + (0.627-0.351)^2 + (50-31)^2}$$

$$= \sqrt{5^2 + 63^2 + 6^2 + 6^2 + 0^2 + 60.2^2 + 0.276^2 + 19^2}$$

$$= \sqrt{361}$$

=19Equation (1)

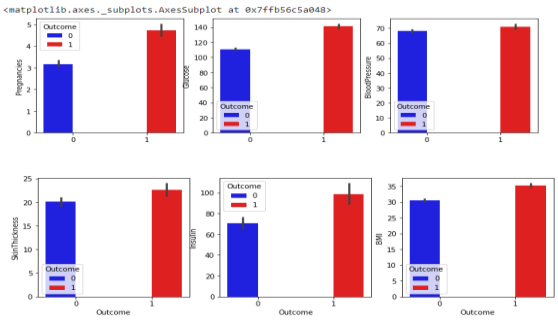
Like this, the Euclidean distance is calculated for all the records in the dataset, out of which the first five nearest neighbors are considered to determine the class label of the predicted value. Similarly, the proposed system also considers Naïve Bayesian algorithm [8], calculates the likelihood probability for all the attributes. Since all the attributes are numerical data, the probability of the data is based on the mean calculation. The calculation part can be represented as shown in the Table 3.

Table 3: Likelihood Calculation for Attribute

Attribute Name	Outcome	Mean
Pregnancies	1	3.28
	0	4.86

Proposed System:

Preprocessing: All the attributes are numerical data types, but based on the analysis graphs as shown in the Figure 3, it is observed that most of the attributes are having the values "0". So to preprocess these values, the proposed system uses the median strategy, because most of the attributes have skewed data. To normalize the data, the system has applied standard scalar technique.



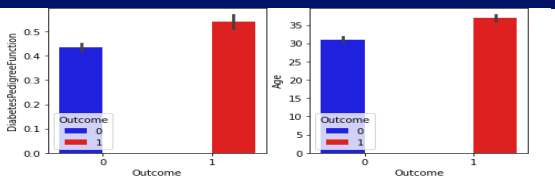


Figure 3: Bar Graph Analysis on all the Attributes

Ensemble Classification Algorithm: The purpose of ensemble algorithm is to identify the weak classifiers among the traditional algorithms and then combine them to improve the accuracy. Ensemble algorithms are of two types. They are Hard Ensemble and Soft Ensemble. The Proposed system combines Decision Tree, Naïve Bayesian, KNN and XGBoost with 5-folded cross validation. It is observed hard ensemble performance better than Soft ensemble algorithm. Hard ensemble models calculate the sum of the models and consider the majority as class label. The Soft ensemble models calculate the sum of the probabilities of the models and consider the majority among them. XGBoost is also an ensemble mechanism, which can speed up the process, it is based on the decision tree algorithm. The major goal of the boosting algorithms is to minimize the error of the previous models.

ACNN Classification Algorithm: The proposed system designed a convolution network for early prediction of the diabetes among the pregnant ladies. The convolution neural network contains 3 dense layers, with 512 units of neurons to perform the dot product of the inputs and their corresponding weights and summed with bias, to optimize the model and finally, element wise it performs the activation function. The dense layer has major advantage it takes inputs input from all the previous layers. It also add dropout layer, to overcome the problem of overfitting. Ensemble algorithms also deal with this problem but it makes the model complicated, so the regularization of neural networks with the help of probability can help the system to simply address the issue. It also applies 5-folded Cross validation and 10 –folded Cross validation, which randomly

splits the dataset into "4" train and "1" test datasets. Each execution will generate a model and all these models are used to estimate the accuracy of the system. The accuracy of the system is described in the Figure-3.

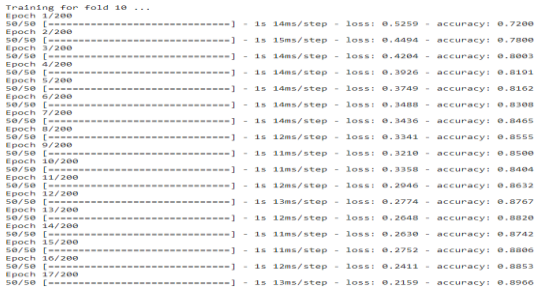


Figure 4: Accuracy of 10-Folded Convolution Neural Network

Conclusion: The proposed system is compared with traditional and ensemble algorithms based on classification evaluation parameters and observed values are tabulated in the Table 4.

Table 4: Comparison of Classification Algorithm with different folds

Parameters/Algorithm	**Proposed CNN**	**Proposed CNN**
	5-Folded	**10-Folded**
Accuracy	95.05	99.55
F1-Score	79.00	80.00
Recall	79.00	80.00
Precision	79.00	80.00
Sensitivity	85.04	84.92
Specificity	65.95	70.36

The proposed system has achieved more accuracy than the ensemble algorithms and it is satisfied all the hyper parameter optimization techniques.

References:

[1] Park J, Edington DW. A sequential neural network model for diabetes prediction. Artif Intell Med. 2001 Nov;23(3):277-93. doi: 10.1016/s0933-3657(01)00086-0. PMID: 11704441.

[2] Wang L, Wang X, Chen A, Jin X, Che H. Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model. Healthcare (Basel). 2020 Jul 31;8(3):247. doi: 10.3390/healthcare8030247. PMID: 32751894; PMCID: PMC7551910.

[3] Park J, Edington DW. Application of a prediction model for identification of individuals at diabetic risk. Methods Inf Med. 2004;43(3):273-81. PMID: 15227557.

[4] Zhou, H., Myrzashova, R. & Zheng, R. Diabetes prediction model based on an enhanced deep neural network. *J Wireless Com Network* **2020,** 148 (2020). https://doi.org/10.1186/s13638-020-01765-7

[5] Pima Indians Diabetes Database, https://www.kaggle.com/uciml/pima-indians-diabetes-database

[6] Enoch A. Frimpong et al 2021 J. Phys.: Conf. Ser. 1734 012026

[7] Panhalkar, A.R., Doye, D.D. A novel approach to build accurate and diverse decision tree forest. Evol. Intel. (2021). https://doi.org/10.1007/s12065-020-00519-0

[8] Gupta, D., Choudhury, A., Gupta, U. et al. Computational approach to clinical diagnosis of diabetes disease: a comparative study. Multimed Tools Appl (2021). https://doi.org/10.1007/s11042-020-10242-8

[9] Kose U., Deperlioglu O., Alzubi J., Patrut B. (2021) A Hybrid Medical Diagnosis Approach with Swarm Intelligence Supported Autoencoder Based Recurrent Neural Network System. In: Deep Learning for Medical Decision Support Systems. Studies in Computational Intelligence, vol 909. Springer, Singapore. https://doi.org/10.1007/978-981-15-6325-6_7

[10] Deepti Sisodia, Dilip Singh Sisodia,Prediction of Diabetes using Classification Algorithms,Procedia Computer Science,Volume 132,2018,Pages 1578-1585,ISSN 1877-0509,https://doi.osrg/10.1016/j.procs.2018.05.122.