Deep Unsupervised Machine Learning for Early Diabetes Risk Prediction using Ensemble Feature Selection and Deep Belief Neural Networks

Olusola Olabanjo * , Ashiribo Wusu , Manuel Mazzara

Posted Date: 12 January 2023

doi: 10.20944/preprints202301.0208.v1

Keywords: Deep belief network; Diabetes; Prediction; Risk Factors; Deep Learning

Article

Deep Unsupervised Machine Learning for Early Diabetes Risk Prediction using Ensemble Feature Selection and Deep Belief Neural Networks

Olusola Olabanjo [1,2,*], Ashiribo Wusu [3] and Manuel Mazzara [4]

[1] Department of Mathematics, Morgan State University, MD, USA
[2] Department of Computer Science, Lagos State University, Lagos, Nigeria
[3] Department of Mathematics, Lagos State University, Lagos, Nigeria
[4] Institute of Software Development and Engineering, Innopolis University, Innopolis 420500, Russia
* Correspondence: olola57@morgan.edu

Abstract: Diabetes mellitus is a popular life-threatening disease and patients may gradually have started suffering from other diabetes-causing diseases such as heart attacks, stroke, hypertension, blurry vision, blindness, foot ulcer, amputation, kidney damage and other organ failures before diagnosis. Early detection can help reduce the fatality of this disease. Deep learning models have proven very useful in disease detection and computer-aided diagnosis. In this work, we proposed a deep unsupervised machine learning model for early detection of diabetes using voting ensemble feature selection and deep belief neural networks (DBN). Dataset was obtained from an online repository containing responses of prediagnosed patients to direct questionnaires administered in Sylhet Diabetes Hospital in Sylhet, Bangladesh. The dataset was preprocessed and preprocessed. Features were reduced using the ensemble feature selector. The DBN model was pretrained and tuned to obtain optimal performance. The model was also compared with other models with no multiple hidden layers. The DBN performed at its relative best with F1-measure, precision and recall of 1.00, 0.92 and 1.00 respectively. We conclude that DBN is a useful tool for an unsupervised early prediction of Type II diabetes mellitus.

Keywords: deep belief network; diabetes; prediction; risk factors; deep learning

I. Introduction

Diabetes mellitus (otherwise often referred to as diabetes) remains one of the popular life-threatening diseases which affects relatively 500 million people worldwide. It is a chronic disease associated with a high blood sugar level in a human's body [1,2]. The pancreas is an organ in the human body that produces a special hormone known as insulin [3]. Insulin is released by the pancreas into the bloodstream, aiding in the transport of glucose into the cells [4]. Diabetes is a condition in which the pancreas is unable to make insulin or in which the body is unable to use insulin as it should. Due to its relatively long asymptomatic phase, its early detection has been receiving massive attention from both medical and non-medical scientists. Diabetes mellitus is known to manifest in two types: Type I and Type II [5]. The former occurs when the pancreatic beta cells are mistakenly attacked by the immune system and the body produces too little – or none at all – insulin while in the case of the latter, the body does not produce enough or becomes actively resistant to insulin. The third, but not so common type of diabetes is gestational diabetes; the case of which a woman becomes diabetic during pregnancy due to hormonal changes. Diabetes mellitus is known to exhibit symptoms such as polyuria, polydipsia, polyphagia, sudden weight loss (usually Type I [6]), weakness, obesity (usually Type II [7]), delayed healing, visual blurring, itching, irritability, genital thrush, partial paresis, muscle stiffness, alopecia, etc.

The alarming fatality of this popular disease is evident from the facts that 85% of diabetic patients were from low- and middle-income countries and that its clinical detection takes so long that

 doi:10.20944/preprints202301.0208.v1

patients may gradually have started suffering from other diabetes-causing diseases such as heart attacks, stroke, hypertension, blurry vision, blindness, foot ulcer, amputation, kidney damage and other organ failures [8,9]. These symptoms set in due to the number of years (7-12) the disease has gone without notice or treatment. In fact, the degree of severity of its manifestation and associated complications correlates with its detection period. This makes early diagnosis, early commencement of treatment as well as early awareness of patient's risk factors to contribute to the reduction of its prevalence globally, thereby beneficial in terms of the patient's health and expenditure [10]. Identification of risk and protective factors is a key component in diseases which are incurable, confusable and takes a long time to manifest [11]. These factors promote awareness, prevents the disease, influence people's lifestyles towards avoiding the disease, fosters effective prevention and suggests routines that serve as positive countermeasures.

Several statistics- and machine-learning-based studies are being conducted daily to predict and diagnose diabetes [12–15]. The advent of technology has revolutionized many sectors including healthcare and medical technologies. It helps in the improvement of services offered to patients and serves as an efficient and effective measures of treating, diagnosing, service delivery, information handling, administration etc [16,17]. In the recent past, machine learning models have centered on the use of supervised deep learning and classical machine learning models for the prediction and determination of the Type-II diabetes risk factors. However, in this study, we propose an unsupervised approach to this study. A deep belief neural network is proposed to determine the risk factors of Type-II diabetes and the impact of ensemble feature selection was measured.

Structurally, the next section discusses some major works done in relation to the study conducted in this paper. The section which follows the related works shall discuss methods in terms of the methodology, data and evaluation techniques of the proposed DBN model. The following sections present and discuss the results as obtained in the model; then we conclude the paper by highlighting major discoveries in this research and charts future directions for the subject matter.

II. Related Works

The application of machine learning models in predicting diabetes' risk factors have gained wide scholarly attention in the recent decades and this can be attributed to sophistication in compute devices and state-of-the-art machine learning algorithms. In this section, we zoom on various machine learning tools and algorithms that have been developed for the prediction of Type-II Diabetes mellitus and their accuracies are also discussed. Many notable risk assessment tools have been proposed, developed or/and deployed for a non-evasive determination of diabetes risk factors, some of which are Latin-America-FINDRISC (LAFINDRISC) [18], Risk Test by American Diabetes Association (RTADA) [19], Leicester Practice Risk Score [20], Test2Prevent [21]. These and many more have proven to be effective screener for assessing the risk of undiagnosed diabetes. The accuracies and applicable reliability of these tools are difficult to quantify because of the absence of Fasting Plasma Glucose (FPG) data or other related data. In terms of the core machine learning engine, many classical and state-of-the-art machine learning (ML) models have been proposed for a non-evasive early prediction of undiagnosed diabetes. They include, but are not limited to, individual models such as Artificial Neural Networks (ANNs) [22], k-Nearest Neighbors (kNN) [23,24], Linear Regression [25], Logistic Regression [26], Naïve Bayes [27], Random Forests (RF) [28], Decision Trees (DT) [29], Support Vector Machine (SVM) [30] among others. The maximum accuracy obtained for these classical models was 97.9%. Table 1 focuses on the accuracies and implementation details of some of the deep learning models which have been proposed for early risk factor detection of diabetes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2023 doi:10.20944/preprints202301.0208.v1

Table 1. Compressed Summary of Some Deep Learning Models for a Non-Evasive Risk Prediction of Diabetes.

Ref	Techniques/ML Models	Methodology	Major Outcomes	Data Sources
[31]	Denosing AE	Normalization, training (704,587), validation (5000) and testing (76,214)	Performance was measured using AUC (0.907)	Mount Sinai Data Warehouse (ICD-9)
[32]	Modified Long Short-Term Memory (LSTM), Attention pooling layer	training, validation and testing: 2/3, 1/6 and 1/6 respectively from 53,208 admissions	Study produced maximum accuracy of 79%	EHR data from hospital patients
[33]	Restricted Boltzmann machine (RBM) and Recurrent Neural Network (RNN)	Feature selection, Min-Max normalization, train (80%), test (20%)	Sensitivity and precision: 90.66%, 75% respectively	PID Data from the UCI Repository
[34]	Modified 1-D CNN and FC layer	The data for training and testing: 15 samples, 10 samples; leave-one out cross-validation	AUC of Type I-Diabetes, Type II - Diabetes, healthy subjects: 0.9659, 0.9625, 0.9644	Breath samples collected by MOS sensors with 1000-sec intervals
[35]	CNN, LSTM, and SVM	Heart rate variability (HRV) data from 71 ECG datasets. 5 fold cross-validation was used.	Validation accuracy of 95.7% was obtained.	ECG data sampled at 500Hz from 40 subjects
[36–40]	Deep Multi-Layer Perceptron (DMLP)	Train-test split, data transformation, k-fold cross validation, normalization, feature selection	Maximum accuracy 88.41%, maximum AUC 84.13%, Sensitivity 87.92%, f1 Score 0.808.	PID, Practice Fusion Dataset and HER dataset of
[41]	Deep Belief Network	Min-max normalization; feature selection by PCA; pre-training for RBMs; supervised fine-tuning	Sensitivity: 100%, F1 score: 0.808	Practice Fusion dataset (9948 patients, ICD-9)

The review of related works shows that:

i. There are machine-learning-assisted decision support and diagnostic tools (although not widely used or accepted in medical practice) for a non-evasive risk prediction of diabetes in medical patients.
ii. Deep Learning models have not been used for diabetes risk prediction as much as classical models such as SVM, kNN and regression models.
iii. The impact of feature engineering on the results of these models needs to be more widely studied.
iv. Ensemble feature selection method has not been applied yet to diabetes risk prediction.
v. Deep belief neural networks and other unsupervised deep learning methods for diabetes risk prediction need more attention.

These and many more form the major justifications for the works conducted in this paper.

III. Methods

a. Dataset

The diabetes dataset used in this study was obtained from a publicly available online repository. It contains the response obtained from 520 subjects (who recently became diabetic or are currently

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2023 doi:10.20944/preprints202301.0208.v1

showing symptoms of diabetes) using a direct questionnaire. This was released by Sylhet Diabetes Hospital of Sylhet, Bangladesh. It consists of the age, sex, Boolean response to each diabetes-related question and the class to which each person belongs after medical diagnosis (Positive or Negative). There are 16 attributes for each subject under consideration, the summary of which is presented in Table 2.

Table 2. Description of Data Features.

SNAttributes	Datatype	Yes (as 1)	No (as 0)
1. Age	20 ≤ Age ≤ 100		
2. Sex	Male and Female	Male (328)	Female (192)
3. Polyuria	Yes/No	258	262
4. Polydipsia	Yes/No	233	287
5. Sudden Weight Loss	Yes/No	217	303
6. Weakness	Yes/No	305	215
7. Polyphagia	Yes/No	237	283
8. Genital Thrush	Yes/No	116	404
9. Visual Blurring	Yes/No	233	287
10. Itching	Yes/No	253	267
11. Irritability	Yes/No	126	394
12. Delayed Healing	Yes/No	239	281
13. Partial Paresis	Yes/No	224	296
14. Muscle Stiffness	Yes/No	195	325
15. Alopecia	Yes/No	179	341
16. Obesity	Yes/No	88	432
17. Class	Positive/Negative	Positive (320)	Negative (200)

The age distribution of the data is given in Figure 1. It shows that our data is normally distributed.

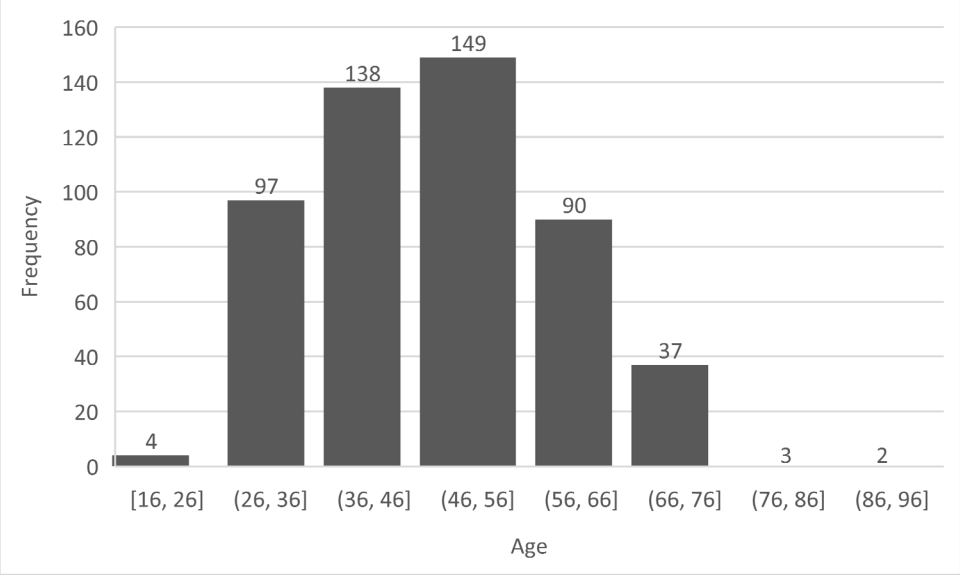

Figure 1. Frequency Distribution of Ages of Subjects.

b. Model Development Workflow

The proposed model development comprises of the preprocessing, ensemble feature selection with the final voting, the DBN pretraining and the finetuning backpropagation for classification.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2023 doi:10.20944/preprints202301.0208.v1

Performance analysis is used to measure the level of satisfaction and confidence accrued to the proposed model. These stages are diagrammatically represented in Figure 2.

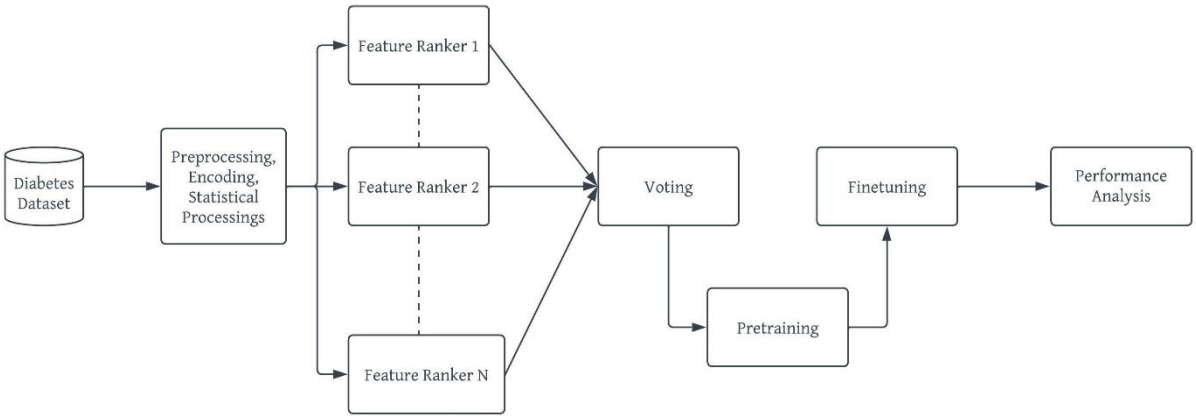

Figure 2. Design Workflow of the Proposed DBN Model.

vi. Preprocessing: This stage ensures that the diabetes dataset to be used is well prepared for the machine learning task [42]. This stage ensures the quality of the dataset in terms of noise and duplicate removal, outlier detection and processing, encoding for a numerical representation of categorical and nominal variables [43]. In the diabetes dataset, all corresponding Yes/No and Male/Female values were replaced with 1/0 respectively. The ages were encoded from 0 – 7 based on the categorization specified in Figure 1 and these values were normalized using the Min-Max normalization [44] to prevent the age column from outweighing other columns during prediction, thereby reducing bias. The output of this stage is a ready dataset for further analyses and model pre-training.

vii. Ensemble Feature Selection: This study uses the ensemble dimensionality reduction framework to select the best feature set for the developed deep learning model while removing redundant features from the dataset. This will avoid misfits, either overfitting or underfitting as well as reduce the curse and complexity of multidimensionality [45,46]. The ensemble selection leverages on the individual strengths of each candidate feature selection method to find the best feature vectors for the deep learning models. The output of this stage is a "project" or subset of the original dataset.

viii. Building, Pretraining and Finetuning the DBN Model: This step comprises of the actual stacking of Restricted Boltzmann Machines (RBMs) [47] to form a deep net and training. DBN is a generative-graph multi-layered model. The process in which the model is used to predict either in a supervised or unsupervised manner is known as pre-training. Each of the deep – hidden – layers is trained as RBMs. The first stage of training DBN is to train layers sequentially from the bottom visible (observed) layer features. This input layer contains D number of units, where D is input sample dimension. This input layer is fully connected with hidden layers. Each Hidden layer consists of N number of RBM. The output layer consists of one unit which defines the class. The final phase, called fine tuning is to train the second layer based on the results from pre-training step. Finally, the entire hidden layers are learned same way till final hidden layer is reached. The Figure 3 outlines the architecture of model pre-training proposed for our study.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2023 doi:10.20944/preprints202301.0208.v1

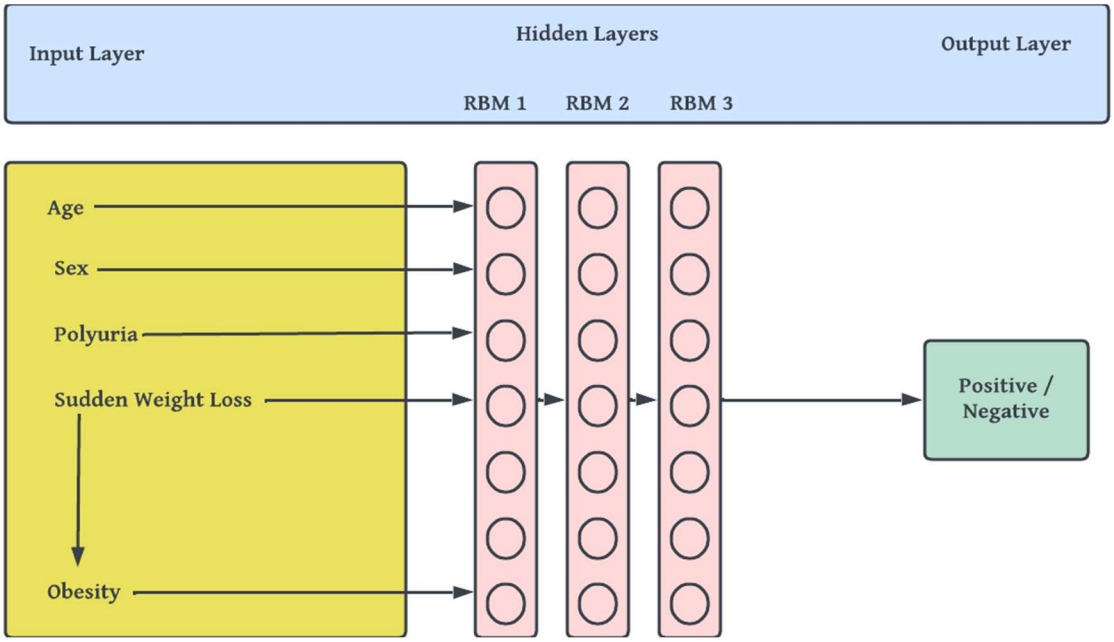

Figure 3. Architecture of Proposed System.

There features in the input layer are an output of the voting ensemble feature selection procedure containing nine features from the possible sixteen features. There are three hidden layers in our DBN model. The output layer is the class to which each instance in our dataset is classified into (Positive/Negative).

ix. Performance Analysis: Our proposed DBN model for diabetes risk prediction was assessed using F1-Measure, Precision and Recall, where

$$Recall, R = \frac{TP}{TP + FN}$$

$$Precision, P = \frac{TP}{TP + FP}$$

$$F1 = \frac{2 \times recall \times precision}{Recall + Precision}$$

where TP is True Positive, FP is False Positive and FN is False Negative as all obtained from the confusion matrix of the result.

IV. Results and Discussion

In this study, we developed a voting ensemble feature selection method which consisted of Chi-Square (CS), Mutual Information Gain (MIG) and Variance Threshold (VT) methods. Top ten methods were selected and prepared to run in the DBN pretraining for the prediction of diabetes mellitus. The parameters were tuned to achieve the most optimal accuracy obtainable. The top ten feature sets were also passed through five benchmark models (KNN, Linear SVM, Logistic Regression, Decision Trees and Random Forests) for performance comparison. We also performed correlation analysis by plotting the correlation matrix in order to determine prior to modeling if there is any overfitting. The categorical nature of the diabetes dataset required that Spearman correlation be used and not Pearson [48]. The result of the voting ensemble feature selection process is given in Table 3. The ensemble voting screened out age, itching and obesity as possible early predictors of diabetes mellitus with obesity having the lowest rank by our three feature rankers. The feature sets which were voted by our stack selectors are sex, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, irritability, delayed healing, partial paresis, muscle stiffness and alopecia. These were then prepared for the pretraining of our DBN model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2023 doi:10.20944/preprints202301.0208.v1

Table 3. Results of Voting Ensemble Feature Selection Method on the Dataset (× means disqualified, while ✓ means qualified, ✓✓ means strongly qualified).

SN	Attributes	Chi Square	Mutual Information Gain	Variance Threshold	Voting
1.	Age	×	✓	×	×
2.	Sex	✓	✓	✓	✓✓
3.	Polyuria	✓	✓	✓	✓✓
4.	Polydipsia	✓	✓	✓	✓✓
5.	Sudden Weight Loss	✓	✓	✓	✓✓
6.	Weakness	✓	✓	✓	✓✓
7.	Polyphagia	✓	✓	✓	✓✓
8.	Genital Thrush	×	✓	✓	✓
9.	Visual Blurring	✓	×	✓	✓
10.	Itching	✓	×	×	×
11.	Irritability	×	✓	✓	✓
12.	Delayed Healing	✓	×	✓	✓
13.	Partial Paresis	✓	✓	×	✓
14.	Muscle Stiffness	✓	✓	✓	✓✓
15.	Alopecia	✓	✓	✓	✓✓
16.	Obesity	×	×	×	×

In the tuning of our model, experimentally selected the values of our parameters and the best for our model and data were identified. Tuning is a crucial stage to avoid fitting problem. For instance, the choice of the number of hidden layers was carefully selected before too small results in underfitting while too large results in overfitting. In this study, we used Rectified Linear Unit (ReLU) as our hidden activation functions, with three hidden layers with 250, 250, 500 as the total number of hidden units in the neural networks, Sigmoid as our input activation function with 20 RBM epochs, 100 batch size and a global learning rate of 0.06.

In our experimental setup, the performance of the deep model was tested in three ways: all features in the original dataset, all qualified (including strongly qualified) features, and the strongly qualified features only. Table 4 shows the results of our various experiments with our DBN model compared with other classical classification models.

Table 4. Performance Analysis of Proposed DBN Model. Compared with Some Classification Models

	F1-Measure	Recall	Precision
Full (16) Features			
Deep Belief Networks	0.87	0.66	0.80
Decision Tree	0.72	0.62	0.72
Random Forest	0.79	0.76	0.65
Logistic Regression	0.86	0.59	0.67
Support Vector Machine	0.66	0.86	0.58
k-Nearest Neighbors	0.72	0.74	0.72
All Qualified (13) Features			
Deep Belief Networks	0.92	0.88	0.88
Decision Tree	0.86	0.69	0.61
Random Forest	0.77	0.72	0.70
Logistic Regression	0.77	0.72	0.78
Support Vector Machine	0.89	0.69	0.68
k-Nearest Neighbors	0.89	0.66	0.80
Strongly Qualified (8) Features			

 doi:10.20944/preprints202301.0208.v1

Deep Belief Networks	1.00	0.92	1.0
Decision Tree	0.86	0.84	0.88
Random Forest	0.69	0.77	0.87
Logistic Regression	0.77	0.69	0.91
Support Vector Machine	0.77	0.88	0.83
k-Nearest Neighbors	0.86	0.88	0.91

Results obtained in this study show that sex, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, muscle stiffness and alopecia are the strongest indicators in the dataset while age, itching and obesity are deemed by the voting ensemble model to have to significant contribution to the diabetes status of the patients. Also, our proposed DBN model performed best when tested with strongly qualified features ad least when all the sixteen features were used. Although it is not in all cases that deep models would outperform models with one or no hidden layer, however our study showed that the DBN model outperformed the classical classification models in terms of average F1-Measure, recall and precision. This study finds its significance in the fact that the deep learning model developed in this work can assist medics and patients in creating awareness on the early predictors of diabetes mellitus. One rather shocking discovery in this study is the fact that even though diabetes affects older people the more, our feature rankers disqualified it as a possible threat of diabetes. Early detection of diabetes is advantageous in the sense that it can help shape lifestyle, dietary and sleeping patterns. Studies have also shown that early and intensive intervention, not only prevents beta-cell dysfunction but also informs on the potential associated cardiovascular risk factors before reaching the blood glucose thresholds currently set for diagnosing Type II diabetes. It has also been established in literature that early treatment combined with metformin-vildagliptin provides relevant improvements in long-term glycaemic control and can positively affect the disease's progression. Hence, the importance of this study [49–51].

V. Conclusion

In this study, we proposed and successfully implemented a deep belief network model, a class of multilayer deep learning models with three RBM layers as the hidden layers. The performance (vis-à-vis F-measure, recall and precision) of the model was tested using data collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor. The effectiveness of dimensionality reduction was also measured using voting ensemble feature selection comprising of Mutual Information Gain, Variance Threshold and Chi Square. We also implemented five classical machine learning models to benchmark the performance of our model. The proposed model can be reconstructed and reoptimized for prediction of other forms of diseases using similar dataset.

References

1. Preston, E.V., et al., Climate factors and gestational diabetes mellitus risk–a systematic review. Environmental Health, 2020. 19(1): p. 1-19.
2. Wang, P., et al., Seasonality of gestational diabetes mellitus and maternal blood glucose levels: evidence from Taiwan. Medicine, 2020. 99(41).
3. Boiko, M., R. Ovchinnikova, and A. Shabrina. THE ROLE OF HORMONES IN THE HUMAN BODY. in Человек. Общество. Культура. Социализация. 2019.
4. Hauge-Evans, A.C., SUGAR, DOGS, COWS, AND INSULIN—THE STORY OF HOW DIABETES STOPPED BEING DEADLY. Frontiers for young minds, 2021. 9.
5. Padhi, S., A.K. Nayak, and A. Behera, Type II diabetes mellitus: A review on recent drug based therapeutics. Biomedicine & Pharmacotherapy, 2020. 131: p. 110708.
6. Eizirik, D.L., L. Pasquali, and M. Cnop, Pancreatic β-cells in type 1 and type 2 diabetes mellitus: different pathways to failure. Nature Reviews Endocrinology, 2020. 16(7): p. 349-362.
7. Padhi, S., M. Dash, and A. Behera, Nanophytochemicals for the treatment of type II diabetes mellitus: a review. Environmental Chemistry Letters, 2021. 19(6): p. 4349-4373.

8. Lee, K.W., et al., Neonatal outcomes and its association among gestational diabetes mellitus with and without depression, anxiety and stress symptoms in Malaysia: A cross-sectional study. Midwifery, 2020. 81: p. 102586.
9. Yang, Q.-Q., et al., The association between diabetes complications, diabetes distress, and depressive symptoms in patients with type 2 diabetes mellitus. Clinical nursing research, 2021. 30(3): p. 293-301.
10. Kopitar, L., et al., Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific reports, 2020. 10(1): p. 1-12.
11. Gadekallu, T.R., et al., Early detection of diabetic retinopathy using PCA-firefly based deep learning model. Electronics, 2020. 9(2): p. 274.
12. Yang, H., et al., New perspective in diabetic neuropathy: from the periphery to the brain, a call for early detection, and precision medicine. Frontiers in endocrinology, 2020. 10: p. 929.
13. Sungheetha, A. and R. Sharma, Design an early detection and classification for diabetic retinopathy by deep feature extraction based convolution neural network. Journal of Trends in Computer Science and Smart technology (TCSST), 2021. 3(02): p. 81-94.
14. Tofte, N., et al., Early detection of diabetic kidney disease by urinary proteomics and subsequent intervention with spironolactone to delay progression (PRIORITY): a prospective observational study and embedded randomised placebo-controlled trial. The lancet Diabetes & endocrinology, 2020. 8(4): p. 301-312.
15. Hasan, D.A., et al. Machine Learning-based Diabetic Retinopathy Early Detection and Classification Systems-A Survey. in 2021 1st Babylon International Conference on Information Technology and Science (BICITS). 2021. IEEE.
16. Ben-Israel, D., et al., The impact of machine learning on patient care: a systematic review. Artificial Intelligence in Medicine, 2020. 103: p. 101785.
17. Peiffer-Smadja, N., et al., Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clinical Microbiology and Infection, 2020. 26(5): p. 584-595.
18. Bernabe-Ortiz, A., et al., Diagnostic accuracy of the Finnish Diabetes Risk Score (FINDRISC) for undiagnosed T2DM in Peruvian population. Primary care diabetes, 2018. 12(6): p. 517-525.
19. Boulton, A.J., et al., Comprehensive foot examination and risk assessment: a report of the task force of the foot care interest group of the American Diabetes Association, with endorsement by the American Association of Clinical Endocrinologists. Diabetes care, 2008. 31(8): p. 1679-1685.
20. Gray, L., et al., Implementation of the automated Leicester Practice Risk Score in two diabetes prevention trials provides a high yield of people with abnormal glucose tolerance. Diabetologia, 2012. 55(12): p. 3238-3244.
21. Coetzee, A., et al., The prevalence and risk factors for diabetes mellitus in healthcare workers at Tygerberg hospital, Cape Town, South Africa: a retrospective study. Journal of Endocrinology, Metabolism and Diabetes of South Africa, 2019. 24(3): p. 77–82-77–82.
22. El_Jerjawi, N.S. and S.S. Abu-Naser, Diabetes prediction using artificial neural network. International Journal of Advanced Science and Technology, 2018. 121.
23. NirmalaDevi, M., S.A. alias Balamurugan, and U. Swathi. An amalgam KNN to predict diabetes mellitus. in 2013 IEEE international conference on emerging trends in computing, communication and nanotechnology (ICECCN). 2013. IEEE.
24. Alehegn, M., R.R. Joshi, and P. Mulay, Diabetes Analysis and Prediction Using Random Forest, KNN, Naïve Bayes And J48: An Ensemble Approach. Int. J. Sci. Technol. Res, 2019. 8(9): p. 1346-1354.
25. Brown, G.C., et al., Quality of life associated with diabetes mellitus in an adult population. Journal of Diabetes and its Complications, 2000. 14(1): p. 18-24.
26. Tabaei, B.P. and W.H. Herman, A multivariate logistic regression equation to screen for diabetes: development and validation. Diabetes Care, 2002. 25(11): p. 1999-2003.
27. Parthiban, G., A. Rajesh, and S. Srivatsa, Diagnosis of heart disease for diabetic patients using naive bayes method. International Journal of Computer Applications, 2011. 24(3): p. 7-11.
28. Xu, W., et al. Risk prediction of type II diabetes based on random forest model. in 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). 2017. IEEE.
29. Al Jarullah, A.A. Decision tree discovery for the diagnosis of type II diabetes. in 2011 International conference on innovations in information technology. 2011. IEEE.

30. Kumari, V.A. and R. Chitra, Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 2013. 3(2): p. 1797-1801.
31. Miotto, R., et al., Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports, 2016. 6(1): p. 1-10.
32. Pham, T., et al., Predicting healthcare trajectories from medical records: A deep learning approach. Journal of biomedical informatics, 2017. 69: p. 218-229.
33. Tripathi, G. and R. Kumar. Early prediction of diabetes mellitus using machine learning. in 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO). 2020. IEEE.
34. Parte, R., et al., Non-invasive method for diabetes detection using CNN and SVM classifier. International journal of research in engineering, science and management, 2019. 2: p. 659-661.
35. Swapna, G., K. Soman, and R. Vinayakumar, Diabetes detection using ecg signals: An overview. Deep Learning Techniques for Biomedical and Health Informatics, 2020: p. 299-327.
36. Hu, J., et al., Raman spectrum classification based on transfer learning by a convolutional neural network: Application to pesticide detection. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2022. 265: p. 120366.
37. Al-Smadi, M., et al., A transfer learning with deep neural network approach for diabetic retinopathy classification. International Journal of Electrical and Computer Engineering, 2021. 11(4): p. 3492.
38. Spänig, S., et al., The virtual doctor: an interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes. Artificial intelligence in medicine, 2019. 100: p. 101706.
39. Nguyen, B.P., et al., Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Computer methods and programs in biomedicine, 2019. 182: p. 105055.
40. Ryu, K.S., et al., A deep learning model for estimation of patients with undiagnosed diabetes. Applied Sciences, 2020. 10(1): p. 421.
41. Prabhu, P. and S. Selvabharathi. Deep belief neural network model for prediction of diabetes mellitus. in 2019 3rd international conference on imaging, signal processing and communication (ICISPC). 2019. IEEE.
42. Zelaya, C.V.G. Towards explaining the effects of data preprocessing on machine learning. in 2019 IEEE 35th international conference on data engineering (ICDE). 2019. IEEE.
43. Deshmukh, D.H., T. Ghorpade, and P. Padiya. Improving classification using preprocessing and machine learning algorithms on NSL-KDD dataset. in 2015 International Conference on Communication, Information & Computing Technology (ICCICT). 2015. IEEE.
44. Patro, S. and K.K. Sahu, Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462, 2015.
45. Saeys, Y., T. Abeel, and Y. Van de Peer. Robust feature selection using ensemble feature selection techniques. in Joint European conference on machine learning and knowledge discovery in databases. 2008. Springer.
46. Seijo-Pardo, B., et al., Ensemble feature selection: homogeneous and heterogeneous approaches. Knowledge-Based Systems, 2017. 118: p. 124-139.
47. Zhang, N., et al., An overview on restricted Boltzmann machines. Neurocomputing, 2018. 275: p. 1186-1199.
48. Bonett, D.G. and T.A. Wright, Sample size requirements for estimating Pearson, Kendall and Spearman correlations. Psychometrika, 2000. 65(1): p. 23-28.
49. Gómez-Peralta, F., et al., When does diabetes start? Early detection and intervention in type 2 diabetes mellitus. Revista Clínica Española (English Edition), 2020. 220(5): p. 305-314.
50. Gilmer, T.P. and P.J. O'Connor, The growing importance of diabetes screening. 2010, Am Diabetes Assoc. p. 1695-1697.
51. Sabariah, M.M.K., S.A. Hanifa, and M.S. Sa'adah. Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART). in 2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA). 2014. IEEE.