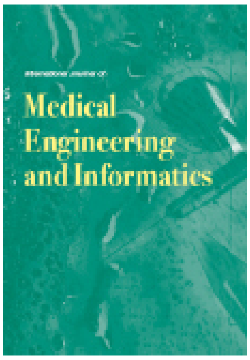


International Journal of Medical Engineering and Informatics

ISSN online: 1755-0661 - ISSN print: 1755-0653
https://www.inderscience.com/ijmei

Risk stratification of cardiovascular disease in type 2 diabetes using LDA and CNN for clinical decision management - a multi-centre study in eastern India

Suparna Dutta, Saswati Mukherjee, Medha Nag, Sujoy Majumdar, Ghanshyam Goyal

DOI: 10.1504/IJMEI.2021.10043803

Article History:

Received:	13 May 2021
Accepted:	18 September 2021
Published online:	22 December 2023

Risk stratification of cardiovascular disease in type 2 diabetes using LDA and CNN for clinical decision management – a multi-centre study in eastern India

Suparna Dutta*

Machine Intelligence Unit,
Indian Statistical Institute,
Kolkata, India
Email: sdutta184@gmail.com
*Corresponding author

Saswati Mukherjee

School of Eduation Technology,
Jadavpur University,
Kolkata, India
Email: swati.mukherjee@jadavpuruniversity.in

Medha Nag

School of Engineering Entrepreneurship,
IIT Kharagpur, India
Email: medhanag93@gmail.com

Sujoy Majumdar

GD Hospital & Diabetes Institute,
Kolkata, India
Email: sujoypinky@gmail.com

Ghanshyam Goyal

ILS Hospital,
Kolkata, India
Email: drgsgoyal@hotmail.com

Abstract: Approximately 72.9 million patients of type 2 diabetes mellitus (T2DM) in India are at a potential risk of cardiovascular diseases (CVDs), strokes and peripheral gangrene. CVD is a major cause of disability and death is one of the major areas of risk severity stratification study. Unlike well-known prediction score models of CVD herein, a unique assessment deep learning model is proposed to stratify the cardiovascular events in different risk grades in T2DM individuals This risk assessment tool can aid clinicians in decision management of CVD risk severity. It is a retrospective cross-sectional observational study that stratifies risks using linear discriminant analysis (LDA) and convolution neural network (CNN). Class separability feature of LDA helps to achieve optimal performance. The model is externally validated in a cohort of 4,719 individuals with T2DM to assess performance heterogeneity across different settings.

Keywords: convolution neural network; CNN; linear discriminant analysis; LDA; support vector machine; SVM; risk stratification; cardiovascular disease; CVD; type 2 diabetes mellitus; T2DM; India.

Reference to this paper should be made as follows: Dutta, S., Mukherjee, S., Nag, M., Majumdar, S. and Goyal, G. (2024) ‘Risk stratification of cardiovascular disease in type 2 diabetes using LDA and CNN for clinical decision management – a multi-centre study in eastern India’, *Int. J. Medical Engineering and Informatics*, Vol. 16, No. 1, pp.1–14.

Biographical notes: Suparna Dutta received both her BTech and MTech degree in Information Technology from St. Thomas’ College of Engineering and Technology, and Government College of Engineering and Ceramic Technology, Kolkata respectively. She is currently working in a research project Intelligent Analysis of Urban Data at Indian Statistical Institute.

Saswati Mukherjee received both her MTech and PhD in Information Technology from Jadavpur University and working as an Assistant Professor in the School of Education Technology, Jadavpur University. She has several publications in national and international journals, conference proceedings and book chapters.

Medha Nag received her MTech degree in Electrical Engineering from Jadavpur University and currently she is working as a research scholar at School of Engineering Entrepreneurship, IIT Kharagpur.

Sujoy Majumdar is an endocrinologist and diabetologist at G.D. Hospital and Diabetes Institute with over 29 years of experience in diabetic foot, myocardial infarction, dyslipidemia and diabetes obesity.

Ghanshyam Goyal is a diabetologist and diabetic foot specialist in the Department of Internal Medicine and Diabetology at ILS Hospitals. He started the first Indian Diabetic Educator Project in collaboration with Project HOPE, to educate and train nurses, dieticians, and paramedics about foot care and management.

1 Introduction

As of 2017, India has emerged as one of the leading countries in type 2 diabetes mellitus (T2DM), with 72.9 million people suffering from T2DM and an overall diabetes prevalence of 7.3% (Carracher et al., 2018). $HbA_1C > 9\%$ puts the vast majority of Indian T2DM patients at an increased risk of microvascular disorders in addition to cardiovascular diseases (CVDs) such as stroke, and peripheral vascular diseases (Young et al., 2008). Individuals with T2DM have a 2- to 4-fold increased risk for CVD, and 1.5 to 3.6-fold increase in mortality (Bertoluci and Rocha, 2017; Einarson et al., 2018). However, preventive clinical measures have significantly controlled the risk of CVD incidence; further supportive therapeutic strategies for individualised treatment remain a challenge among primary care clinicians (Kaasenbrood et al., 2016).

Contemporarily, three types of methods are adopted to predict the progression of chronic diseases, namely statistical methods, machine learning algorithms, and network approach. Statistical methods such as regression models predict the mortality rate of stroke patients, adverse kidney events, and early detection of diabetes (Lee et al., 2013; McKown et al., 2017; Krishnan et al., 2013). Some clinical scoring tools predict the future risk of CV events from long-term observational studies such as Framingham risk score, UKPDS risk engine, QRISK, SCORE, and DECODE (Conroy et al., 2003; Artigao-Rodenas et al., 2013; Hippisley-Cox et al., 2017, 2008). However, these scoring tools have been validated mostly for a non-Indian population with the exception of QRISK3 which considered non-resident Indians residing in UK. Various machine learning algorithms such as, PCA, LDA, and Gaussian mixture model (GMM) were applied that predict risks of heart disease (Chen et al., 2015; Giri et al., 2013). A product-unit neural networks (PUNN)-based classification (Benali et al., 2019) predicts diabetes and evolutionary algorithm estimates the network weights. Ensemble classifiers detects the onset of diabetes (Tama and Rhee, 2018). Macronutrient intake-based obesity and diabetes risks are predicted using descriptive statistics and logistic regression (Pattabiraman et al., 2020). Logistic regression and bayes algorithm are applied in coronavirus detection (Khanday et al., 2020). Wavelet transform and support vector machines (SVMs) are applied for MR-brain image classification (Khalil et al., 2021). Multi-layer perceptron and PCA are used for malignant melanoma detection (Mukherjee et al., 2020). In other studies (Wu et al., 2019; Ashraf et al., 2019; Zhang et al., 2017; Liu et al., 2019; Ricciardi et al., 2020), convolution neural network (CNN), heterogeneous CNN (HCNN) predict heart and coronary artery diseases, and cardiac arrhythmia using demographics, medication, and laboratory features. The progression of CVD in T2DM patients is studied using a network approach (Khan et al., 2018; Hossain et al., 2019, 2020). In literature, most of these predictive models can either predict the onset of different comorbid diseases or can present a score-based CV risk prediction in non-Indian T2DM patients. Thus, it is necessary to develop a prediction model for CVD risk stratification for resident-Indian T2DM patients. In our study, we use the terms method and model interchangeably.

1.1 Our contributions

The need to stratify the resident-Indian T2DM individuals into different CVD risk groups inspired us to develop a deep learning-based prediction model which can be used as a risk assessment tool by the clinicians for decision management. The proposed LDACNN model has the following components.

- The model combines linear discriminant analysis (LDA) with CNN to predict the risk-severity of CV events. The class separability feature of LDA maximises the spread between the output classes followed by CNN which gives optimum classification of non-Gaussian data.

- The multi-centre cross-sectional study was conducted by collecting two retrospective cohorts of type 2 diabetes groups with or without a baseline CVD from two different hospitals in eastern part of India (Kolkata) with different available risk factors.
- An user-interface was designed and integrated with the LDACNN model to facilitate the data-entry of the clinical profile of T2DM patients to stratify the risk-severity.
- Further, two machine learning models are implemented, namely, 1D CNN and CNNSVM for performance comparison with the proposed LDACNN model.

2 Materials and methods

This multi-centric retrospective randomised observational study was conducted on resident-Indian T2DM patients at hospitals providing tertiary care in Kolkata, India during 2010–2019. Large dataset is used to train and test the prediction model. An independent validation dataset was used to confirm the robustness of the proposed model.

2.1 Study population – train and test dataset

In this study, we collected the health data from two different hospitals in Kolkata, a city with a population of over 10 million people of diverse ethnicity and heterogenous community. Two relevant research datasets were obtained from the medical records of T2DM patients after removing ambiguous and missing data. The first dataset denoted as S_1 had 8,441 records count of single out-patient-department (OPD)-visit patients of GD Hospital & Diabetes Institute, Kolkata, and the second dataset denoted as S_2 comprised of 5,152 records of the ILS Hospital, Kolkata. Identifiers of all patients were removed to maintain confidentiality, and all patients were assigned a unique identification number in the database. Both datasets were randomly split into train and test subsets in the ratio 8:2. The inclusion criteria is adult patients ($\geq$25 years) with or without baseline CV events; demographics, diabetes span, contributing risk factors such as hypertension, dyslipidaemia, smoking, and comorbidities such as established CV disease, CKD, and ischemic foot ulcers. Patients who were pregnant or showed acute medical emergency such as sepsis, acute cardiovascular events like myocardial infarction and stroke were excluded.

2.2 Risk predictors in model building

The role of laboratory markers, comorbidities, and associated risk factors helps to understand the domains of T2DM severity (Bertoluci and Rocha, 2017). In this multi-centric study, the relevant severity domain as suggested by the clinical experts help us identify the risk predictors and form the input feature set for the train and test datasets. Based on the data availability, datasets S_1 and S_2 consists of 19 and 21 nonlinear risk predictors respectively out of which a few predictor variables are shown in Table 1.

Table 1 Predictor variables of datasets S_1 and S_2

Predictor variables	*Dataset* S_1	*Dataset* S_2
Age, gender, BMI and family history of DM	+	+
Systolic blood pressure (SBP) and diastolic blood pressure (DBP)	+	+
Fasting blood sugar (FBS) and glycated haemoglobin (HbA$_1$C)	+	+
High-density lipoprotein cholesterol (HDL-c)	+	+
Low-density lipoprotein cholesterol (LDL-c)	+	+
Triglycerides (TG), past CVD and ischemic foot ulcer	+	+
Smoking*	+	-
Past stroke**	-	+
Past chronic kidney disease (CKD)**	-	+

Notes: *smoking records not available in ILS hospital; **past stroke, and past CKD records not available in GD Hospital; S_1: patients' records of GD Hospital & Diabetes Institute; S_2: patients' records of ILS Hospital.

2.3 Study population – external validation dataset

We validated the risk assessment prediction tool in an external cohort of 4,719 patients collected from SVS Marwari Hospital, Kolkata, India for the period (2006–2013) with no overlap with the train and test datasets, that is, the datasets were different in terms of geography, institution, and/or time. Similar inclusion and exclusion criteria were applied to this cohort. The external validation dataset S_3 had 19 predictors which posses contiguity of homogeneity to dataset S_1.

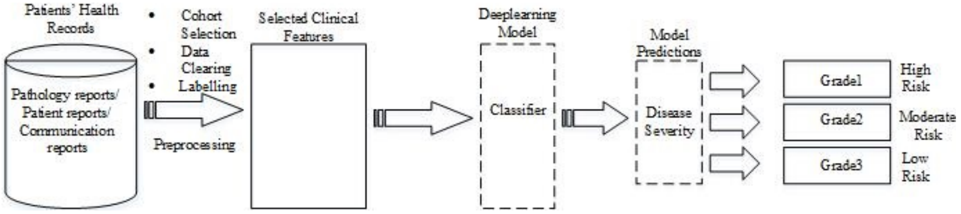
2.4 Data pre-processing

Preprocessing of the data was conducted as follows.

- Missing value calculation – The values of the missing parameters such as HDL-c, LDL-c, TG, and total cholesterol (TC) are computed using the Friedewald formula (Knopfholz et al., 2014) [LDL-c = TC – HDL-c – (TG/5)], limitations were considered as applicable.
- Assigning predictor variables – Presence/absence of risk factors such as smoking, family history, comorbidities and ischemic foot ulcer were assigned the values 1 and 0 respectively. The remaining predictor variables for both the datasets were continuous values.
- Encoding categorical data – In our feature set, we used the age-gender risk score (Grundy et al., 1999) as the predictor variable, primarily because of two reasons. First, encoding categorical feature (gender) as a number (male = 1, and female = 0) would propagate weight- bias in feature processing. This would eventually affect the risk stratification accuracy as both the variables supposed to contribute equal weights. Second, it is necessary to estimate the risk score for each age range while considering the gender (Booth et al., 2006). The age-gender risk score not only provide a perspective of a patient's overall risk status but also helps in computing the risk severity.The work-flow for data processing and model development is shown in Figure 1.

In this study, we explored the mean value of patients' demographics, and major risk factors for the two cohorts. We observed that in dataset S_1, the mean of BMI, systolic blood pressure, HbA$_1$C, HDL-c, LDL-c, TG, past CVD has marginal difference in compare to dataset S_2 with an exception of family history and diabetes span whose mean values are higher in dataset S_2.

Figure 1 Deep learning work-flow for data processing and model development



2.5 *CV risk stratification*

The major CV risks in diabetes patients as per 2019 European Society of Cardiology (ESC) guidelines are smoking, diabetes span, family history, hypertension, dyslipidaemia (HDL-c < 40 or LDL-c > 160 or TG > 150), BMI, and HbA$_1$C > 9. Cardiovascular risk stratification is recommended with the presence or absence of multiple risk factors, macrovascular complications and diabetes spans such as (duration $\geq$10 years or <20 years) (Cosentino et al., 2020). In our study, patients were categorised in different CV risk grades such as higher risk (presence of 3 or more major risk factors and duration $\geq$20 years), lower risk (duration <10 years without other risk factors), or moderate risk (duration $\geq$10 years without target organ damage, and any additional risk factors).

3 Methods

Once the sample study was complete and features were defined, we considered developing a deep learning method coupled with a machine learning algorithm as the proposed prediction model named as LDACNN. Two more prediction models, 1D CNN, and CNNSVM are also implemented for comparison. Each of the multi-centric datasets (S_1 and S_2) is considered for evaluating the three prediction models. Finally, we validated the three prediction models on the external dataset S_3 to ensure model robustness.

3.1 *Linear discriminant analysis*

LDA is a transformation technique that maximises the between-class scatter and minimises the within-class scatter (Liu et al., 2019). In the proposed method LDACNN, we applied LDA before employing the CNN classifier so to optimally organise the nonlinear data with maximum linear separability between the output classes. Thus, we obtained a new feature space with maximum variance between the different categories of risk grades (output classes) and minimum variance within each categories of risk grades.

For inter-class and intra-class variance, the mean of different risk grades categories and data points of each categories is calculated.

Application of LDA introduced a few overlapped classes due to nonlinear data. So, CNN classifier was applied followed by LDA for accuracy. Based on the input, LDA generates two components as there are three different categories of risk grades. The two components are combined with the original input of size ($1 \times m$), changing the dimension to ($1 \times n$). The combined data is fed to the convolution layer followed by a dropout of 50% to avoid overfitting. After linking with the flatten layer followed by the fully-connected layer and softmax layer, the output is the discrete probabilities of the three types of risk-grades.

3.2 Convolution neural network (1D CNN)

CNN is a class of deep neural networks, commonly applied to classification and recognition problems that performs feature extraction and selection simultaneously (Zhang et al., 2017). After applying LDA, CNN classifier stratified the patients in different risk-grades. Feature extraction is done by convolution layers and pooling layers. Convolutional layers are convolved using multiple filters to generate feature maps given in equation (1).

$$x_i^l = f\left(\sum_{j \varepsilon M_i} x_i^{l-1 * k_{ji}^l} + b_i^l\right) \tag{1}$$

where l is the number of layers, x_j^{l-1} represents the input of l^{th} layer, k_{ji}^l is the kernel from the j^{th} neuron at the layer $l - 1$ to the j^{th} neuron at layer l, b is the bias and f is the activation function. Pooling layer is discarded to ensure that the extracted number of features is not reduced. In our 1D CNN model, there are four hidden layers, where the first layer is the convolution layer that computes an element-wise dot product between the kernel and the output of the previous layer's neurons to generate the convolutional neuron. The second layer is the dropout layer with a dropout ratio of 0.5 that simulates a sparse activation from a given layer and prevents network layers to co-adapt to rectify mistakes from the previous layer during training. The third layer is a flatten layer of size (1×640) that transforms the entire feature maps (20×64) into a one-dimensional array for inputting it to the fully-connected layer for classification. The fourth layer is the fully-connected layer where all the inputs from the previous layer are linked to every activation unit of the softmax layer to predict a discrete probability of each output class (different risk-grades). The error/loss of each layer is calculated by the back-propagation algorithm to improve prediction. We obtained a finer structure of the network with repeated experiments over a longer duration.

3.3 Support vector machine

SVM is one of the most popular supervised machine learning methods used for binary or multi-class classification. In this method, data is considered to be linearly separable and boundary values range between –1 and 1, where a hyperplane/decision plane is established. We applied the SVM classifier to the convolution network for classification.

CNN performs the feature mapping, whereas the fully connected layer is replaced with the SVM classifier to classify the output into three risk grades. CNN model generates the intermediate output which is fed to the SVM classifier for multi-grade classification.

3.4 User interface design

An user interface is designed using Python and integrated with the proposed model LDACNN. This user interface facilitates a T2DM patient to input his/her clinical profile such as laboratory features, complications, past diseases, and demographics and based on the clinical data the proposed model stratify the risk of CVD in three different grades: grade-1 (low-risk), grade-2 (moderate-risk), and grade-3 (high-risk).

4 Result and analysis

All computations were performed on Intel ® core i5-8250U @3.4 GHz PC using Keras framework with Tensorflow libraries in Python 3.7. In this multi-centre study, the two cohorts of patients correspond to datasets S_1 and S_2. S_1 with 8,441 records is split in the ratio 8:2 to obtain a train set (6,752 records), and a test set (1,689 records). S_2 with 5,152 records is split to obtain a train set (4,121 records), and a test set (1,031 records). The proposed model is externally validated in a retrospective cohort of 4,719 patients with similar predictor variables as used in S_1. In our experiments, we compared the proposed model with the two baseline models with respect to confusion matrix, classification accuracy, sensitivity, specificity, and positive predictivity (PP). To project the CVD risk, multi-variate analysis of independent predictor variables is performed using ordinal regression as given in Section 5. Following are the model parameters for all the experiments:

1 the number of filters is 64 in the CONV layer

2 filter size of 3 × 3

3 dropout rate p is 0.5

4 RELU activation function at each convolution layer

5 cross-entropy loss for softmax activation function at output layer

6 decisionfunctionshape is ovo for SVM.

Training is performed using Adam optimisation instead of stochastic gradient descent to facilitate iterative update of the network weights. In LDACNN and 1D CNN models, classification was considered accurate when CV risks are graded with the highest probability in 100 epochs. The confusion matrix of different models for the train datasets S_1 and S_2 is given in Tables 2 and 3 respectively. The misclassification rate of LDACNN model for S_1 was much lower with 2.19% and 0.42% (Table 2) than CNNSVM model. CNNSVM performed poorly with 100% misclassification rate. For dataset S_2, LDACNN misclassified only 3.39% and 0.04% patients while 1D CNN showed a high misclassification rate of 52.54% (Table 3).

Table 2 Confusion matrix of dataset S_1

	1	*2*	*3*
		LDACNN	
1	358	8	0
	97.8%	2.19%	
2	12	2855	0
	0.42%	99.5%	
3	0	0	3519
			100%
		CNNSVM	
1	0	366	0
		100%	
2	0	2867	0
		100%	
3	0	0	3519
			100%

Table 3 Confusion matrix of dataset S_2

	1	*2*	*3*
		LDACNN	
1	57	2	0
	96.6%	3.39%	
2	1	2321	0
	0.04%	99%	
3	0	0	1740
			100%
		1D CNN	
1	28	31	0
	47.5%	52.54%	
2	8	2314	0
	0.34%	99.6%	
3	0	2	1738
		0.11%	99.8%

Table 4 illustrates the prediction performance of the three models on train and test dataset S_1. The proposed LDACNN outperformed the two baseline models 1D CNN and CNNSVM in terms of accuracy, PP, sensitivity, and specificity. Mores so, for dataset S_2, LDACNN outperformed the other two models in terms of accuracy (train: 99%, test: 99%), PP (train: 99%, test: 95%), sensitivity (train: 98%, test: 96%) and specificity (train: 99%, test: 99%) for both train and test dataset S_2. The specificity of LDACNN and 1D CNN was slightly similar, the latter showed a decrease in accuracy (train: 98%, test: 97%), PP (train: 92%, test: 93%), and sensitivity (train: 82%, test: 81%). Although CNNSVM achieved an accuracy of 98%, PP and sensitivity were as low as 65% and 66% for both train and test datasets. All the three prediction models were finally validated on an external dataset S_3 with 4,719 T2DM patient records for baseline CVD and non-CVD. An accuracy of 99% was achieved in case of all the three models; The proposed model LDACNN reproduced well when compared with the baseline models, with PP (94%), sensitivity (90%), and specificity (99%).

Table 4 Metrics comparison of prediction models for dataset S_1

Methods	*Train dataset*				*Test dataset*			
	Accuracy	*PP*	*Sensitivity*	*Specificity*	*Accuracy*	*PP*	*Sensitivity*	*Specificity*
LDACNN	0.99	0.98	0.99	0.99	0.99	0.98	0.98	0.99
1D CNN	0.98	0.94	0.96	0.99	0.98	0.91	0.93	0.99
CNNSVM	0.94	0.62	0.66	0.97	0.95	0.63	0.66	0.97

5 Discussion

Cardiovascular events are a common and severe complication in T2DM and several studies have attempted to facilitate its prediction. Framingham risk score, QRISK3, etc. are essentially based on longitudinal data of mostly non-Indian communities. On the other hand, we are compelled to quote such studies in the literature (Chen et al., 2015; Wu et al., 2019; Ashraf et al., 2019) where demographics, vital signs, and laboratory features of single-visit patients were considered for predicting multiple disease risks using classical machine learning models. However, all these studies have been mostly based on non-Indian residents that predict the disease-onset as a binary classification without assigning grade or score to the disease. Leveraging the useful information about patient data such as ischemic foot ulcer and disease history is beyond the scope of such models. Heart disease prediction using CNN (Shankar et al., 2020) is a single-visit cohort study where specific risk factors of heart disease and laboratory results are not mentioned during dataset preparation. To the best of our knowledge, our study is the first to develop a risk stratification tool using deep learning algorithms that combined time-insensitive features such as demographics, disease duration, laboratory tests, and disease history. This study cannot be compared with Framingham risk score, QRISK3, and SCORE as it stratifies CVD risk considering single-visit Indian type 2 diabetes patients.

Experiments showed that LDACNN achieved optimal performance when compared with other baseline models, indicating the potential of coupling deep learning model with LDA. However, due to non-Gaussian data distribution, overlapping between the output categories is a major concern which is avoided by applying 1D CNN for improved prediction. The proposed method is flexible enough to accommodate new clinical features such as heart rate variability (HRV), and ejection factor for comprehensive prediction outcome. The robustness of the proposed model was justified by the use of an external validation dataset that confirmed superior performance.

5.1 Independent predictor variables analysis

The core variables of CVD risk factors in T2DM have been studied and statistically analysed in datasets S_1 and S_2. The multi-variate analysis of the predictor variables was performed using ordinal logistic regression as the response categories, that is, CV risk grades are ordered. Despite close similarities of evaluable risk factors between the two datasets, there are few observed differences which are depicted in Tables 5 and 6. CV risk is not projected for past stroke and CKD (Table 5) as these are not documented in S_1. Likewise, there was no CV risk projection for tobacco smoking (Table 6) since it was not recorded in S_2 for reasons unknown. In Tables 5 and 6, duration of diabetes, age, BMI > 30, FBS, HbA_1C, ischemic foot ulcer and past CVD are found highly significant for future prediction of CVD risk ($P < 0.05$) while gender, PPBS and LDL-c (Table 5) and gender, family history, SBP, HDL-c and triglycerides (Table 6) shows no statistical difference ($P > 0.05$). Smoking ($\geq$10 cigarettes/day) is highly significant ($P < 0.001$) contributing a CVD risk with (OR: 1.640, 95% CI: 1.473, 1.83) (Table 5). Stroke and CKD (Table 6) shows statistical difference ($P < 0.001$) with (OR: 1.628, 95% CI: 1.312, 2.019) and (OR: 1.986, 95% CI: 1.625, 2.425) respectively. The risk factors observed in both tables has almost identical prevalence for prediction.

Table 5 Ordinal regression analysis results of influencing factors of CVD risk in type 2 diabetes for dataset S_1

Factors	β	*SE*	*Wald* χ^2	*P*	*OR*	*95% CI*
Duration of diabetes	0.046	0.003	236.975	<0.001	1.048	1.042, 1.05
Age	0.005	0.001	11.262	<0.001	1.005	1.002, 1.01
Gender	0.006	0.061	0.011	0.914	1.01	0.892, 1.14
BMI	0.028	0.003	79.798	<0.001	1.029	1.023, 1.03
Family history of DM	0.654	0.053	149.67	<0.001	1.923	1.732, 2.14
Smoking	0.494	0.054	81.703	<0.001	1.640	1.473, 1.83
SBP	0.0054	0.0009	36	<0.001	1.01	1.003, 1.01
FBS	0.001	0.0002	64.208	<0.001	1.002	1.001, 1.00
PPBS	0.00022	0.00018	1.488	0.214	1.000	1.000, 1.00
HbA_1C	0.059	0.006	76.912	<0.001	1.062	1.048, 1.08
HDL-c	0.003	0.001	6.927	0.008	1.004	1.001, 1.01
LDL-c	0.001	0.001	1.560	0.212	1.00	0.999, 1.00
Triglycerides	0.0006	0.0002	7.022	0.008	1.001	1.000, 1.00
Ischemic foot ulcer	0.947	0.028	65.97	<0.0001	2.58	2.05, 3.24
Past CVD	0.721	0.026	53.98	<0.0001	2.06	1.70, 2.49

5.2 Limitations

The unavailability of longitudinal data restrains us to develop a time series model in our patients. In this research, we considered data from adults aged 25 years or older. Therefore, our results are not applicable to children or adolescents. More so, the result outcomes may not apply to type 1 diabetes patients. The overall results of the study would have impacted as we have analysed only prevalence studies; incidence studies would have generated different results due to different consideration. Method of data collection (national health repository versus clinical records) and the time period over which the data was collected may vary the outcome of the study. Furthermore, moderate to severe CKD stages, peripheral gangrene for risk stratification are not examined in our method.

Table 6 Ordinal regression analysis results of influencing factors of CVD risk in type 2 diabetes for dataset S_2

Factors	β	*SE*	*Wald* χ^2	*P*	*OR*	*95% CI*
Duration of diabetes	0.036	0.004	79.21	<0.001	1.036	1.028, 1.045
Age	0.014	0.0029	24.871	<0.001	1.015	1.009, 1.020
Gender	0.147	0.082	3.179	0.075	1.159	0.986, 1.364
BMI	0.018	0.005	10.588	0.001	1.019	1.007, 1.030
Family history of DM	0.060	0.083	0.513	0.474	1.062	0.901, 1.252
SBP	0.002	0.001	3.345	0.067	1.003	1.000, 1.006
FBS	0.001	0.00041	6.796	0.009	1.001	1.000, 1.002
PPBS	–0.001	0.0003	15.046	<0.001	0.999	0.998, 0.999
HbA_1C	0.035	0.012	8.439	0.004	1.036	1.012, 1.062
HDL-c	–0.00322	0.00391	0.677	0.409	0.997	0.989, 1.005
LDL-c	–0.011	0.002	15.00	<0.001	0.989	0.983, 0.995
Triglycerides	–0.00047	0.00062	0.574	0.444	1.000	0.998, 1.001
Past stroke	0.487	0.109	19.651	<0.001	1.628	1.312, 2.019
Past CKD	0.685	0.102	45.171	<0.001	1.986	1.625, 2.425
Ischemic foot ulcer	0.530	0.035	30.281	<0.001	1.701	1.411, 2.051
Past CVD	0.475	0.070	45.927	<0.001	1.672	1.405, 1.877

6 Conclusions

A CVD risk stratification tool coupled with a user-interface is proposed for T2DM patients on a single OPD-visit from two different hospitals of Kolkata. This method categorises patients in different risk groups to support clinicians for close clinical attention and management. The experimental comparison demonstrated that feature transformation combined with deep learning algorithm achieved optimal performance when compared with other contemporary classification methods. Future studies of follow-up data with deep learning models will be of added benefit.

References

Artigao-Rodenas, L.M., Carbayo-Herencia, J.A., Divison-Garrote, J.A., Gil-Guillen, V.F., Masso-Orozco, J., Simarro-Rueda, M., Molina-Escribano, F., Sanchis, C., Carrion-Valero, L., de Coca, E.L. et al. (2013) 'Framingham risk score for prediction of cardiovascular diseases: a population-based study from southern Europe', *PLoS One*, Vol. 8, No. 9, p.e73529.

Ashraf, M., Rizvi, M. and Sharma, H. (2019) 'Improved heart disease prediction using deep neural network', *Asian Journal of Computer Science and Technology*, Vol. 8, No. 2, pp.49–54.

Benali, R., Dib, N. and Reguig, F.B. (2019) 'Product unit neural network trained by an evolutionary algorithm for diabetes disease diagnosis', *International Journal of Medical Engineering and Informatics*, Vol. 11, No. 3, pp.286–298.

Bertoluci, M.C. and Rocha, V.Z. (2017) 'Cardiovascular risk assessment in patients with diabetes', *Diabetology & Metabolic Syndrome*, Vol. 9, No. 1, pp.1–13.

Booth, G.L., Kapral, M.K., Fung, K. and Tu, J.V. (2006) 'Relation between age and cardiovascular disease in men and women with diabetes compared with non-diabetic people: a population-based retrospective cohort study', *The Lancet*, Vol. 368, No. 9529, pp.29–36.

Carracher, A.M., Marathe, P.H. and Close, K.L. (2018) 'International Diabetes Federation 2017', *Journal of Diabetes*, Vol. 10, No. 5, pp.353–356.

Chen, Q., Li, H., Tang, B., Wang, X., Liu, X., Liu, Z., Liu, S., Wang, W., Deng, Q., Zhu, S. et al. (2015) 'An automatic system to identify heart disease risk factors in clinical texts over time', *Journal of Biomedical Informatics*, December, Vol. 58, No. Supplement, pp.S158–S163.

Conroy, R.M., Pyörälä, K., Fitzgerald, A.E., Sans, S., Menotti, A., De Backer, G., De Bacquer, D., Ducimetiere, P., Jousilahti, P., Keil, U. et al. (2003) 'Estimation of ten-year risk of fatal cardiovascular disease in Europe: the score project', *European Heart Journal*, Vol. 24, No. 11, pp.987–1003.

Cosentino, F., Grant, P.J., Aboyans, V., Bailey, C.J., Ceriello, A., Delgado, V., Federici, M., Filippatos, G., Grobbee, D.E., Hansen, T.B. et al. (2020) '2019 ESC guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD: the task force for diabetes, pre-diabetes, and cardiovascular diseases of the European Society of Cardiology (ESC) and the European Association for the Study of Diabetes (EASD)', *European Heart Journal*, Vol. 41, No. 2, pp.255–323.

Einarson, T.R., Acs, A., Ludwig, C. and Panton, U.H. (2018) 'Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007–2017', *Cardiovascular Diabetology*, Vol. 17, No. 1, pp.1–19.

Giri, D., Acharya, U.R., Martis, R.J., Sree, S.V., Lim, T-C., Vi, T.A. and Suri, J.S. (20130 'Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform', *Knowledge-Based Systems*, Vol. 37, pp.274–282, ISSN: 0950-7051.

Grundy, S.M., Pasternak, R., Greenland, P., Smith Jr., S. and Fuster, V. (1999) 'Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations: a statement for healthcare professionals from the American Heart Association and the American College of Cardiology', *Circulation*, Vol. 100, No. 13, pp.1481–1492.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., Minhas, R., Sheikh, A. and Brindle, P. (2008) 'Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2', *BMJ*, Vol. 336, No. 7659, pp.1475–1482.

Hippisley-Cox, J., Coupland, C. and Brindle, P. (2017) 'Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study', *BMJ*, Vol. 357, No. j2099.

Hossain, M.E., Khan, A. and Uddin, S. (2019) 'Understanding the comorbidity of multiple chronic diseases using a network approach', in *Proceedings of the Australasian Computer Science Week Multiconference*, pp.1–7.

Hossain, M.E., Uddin, S., Khan, A. and Moni, M.A. (2020) 'A framework to understand the progression of cardiovascular disease for type 2 diabetes mellitus patients using a network approach', *International Journal of Environmental Research and Public Health*, Vol. 17, No. 2, pp.596–614.

Kaasenbrood, L., Boekholdt, S.M., Van Der Graaf, Y., Ray, K.K., Peters, R.J., Kastelein, J.J., Amarenco, P., LaRosa, J.C., Cramer, M.J., Westerink, J. et al. (2016) 'Distribution of estimated 10-year risk of recurrent vascular events and residual risk in a secondary prevention population', *Circulation*, Vol. 134, No. 19, pp.1419–1429.

Khalil, M., Ayad, H. and Adib, A. (2021) 'Mr-brain image classification system based on SWT-LBP and ensemble of SVMs', *International Journal of Medical Engineering and Informatics*, Vol. 13, No. 2, pp.129–142.

Khan, A., Uddin, S. and Srinivasan, U. (2018) 'Comorbidity network for chronic disease: a novel approach to understand type 2 diabetes progression', *International Journal of Medical Informatics*, Vol. 115, pp.1–9, ISSN: 0957-4174.

Khanday, A.M.U.D., Rabani, S.T., Khan, Q.R., Rouf, N. and Din, M.M.U. (2020) 'Machine learning based approaches for detecting COVID-19 using clinical text data', *International Journal of Information Technology*, Vol. 12, No. 3, pp.731–739.

Knopfholz, J., Disserol, C.C.D., Pierin, A.J., Schirr, F.L., Streisky, L., Takito, L.L., Ledesma, P.M., Faria-Neto, J.R., Olandoski, M., Da Cunha, C.L.P. et al. (2014) 'Validation of the friedewald formula in patients with metabolic syndrome', *Cholesterol*, Vol. 2014, Article ID 261878, pp.261–878.

Krishnan, R., Razavian, N., Choi, Y., Nigam, S., Blecker, S., Schmidt, A. and Sontag, D. (2013) 'Early detection of diabetes from health claims', in *Machine Learning in Healthcare Workshop, NIPS*.

Lee, J., Morishima, T., Kunisawa, S., Sasaki, N., Otsubo, T., Ikai, H. and Imanaka, Y. (2013) 'Derivation and validation of in-hospital mortality prediction models in ischaemic stroke patients using administrative data', *Cerebrovascular Diseases*, Vol. 35, No. 1, pp.73–80.

Liu, J., Song, S., Sun, G. and Fu, Y. (2019) 'Classification of ECG arrhythmia using CNN, SVM and LDA', in *International Conference on Artificial Intelligence and Security*, Springer, pp.191–201.

McKown, A.C., Wang, L., Wanderer, J.P., Ehrenfeld, J., Rice, T.W., Bernard, G.R. and Semler, M.W. (2017) 'Predicting major adverse kidney events among critically ill adults using the electronic health record', *Journal of Medical Systems*, Vol. 41, No. 10, pp.1–7.

Mukherjee, S., Adhikari, A. and Roy, M. (2020) 'Malignant melanoma detection using multi layer preceptron with visually imperceptible features and PCA components from med-node dataset', *International Journal of Medical Engineering and Informatics*, Vol. 12, No. 2, pp.151–168.

Pattabiraman, S., Vyas, R. and Srinivasan, S. (2020) 'Dietary macro-nutrients intake and risk of obesity and type 2 diabetes: to compute a model to predict probability of developing hypertrophic obesity and type 2 diabetes based on the macro-nutrient intake levels', *International Journal of Medical Engineering and Informatics*, Vol. 12, No. 5, pp.457–474.

Ricciardi, C., Valente, A.S., Edmund, K., Cantoni, V., Green, R., Fiorillo, A., Picone, I., Santini, S. and Cesarelli, M. (2020) 'Linear discriminant analysis and principal component analysis to predict coronary artery disease', *Health Informatics Journal*, Vol. 26, No. 3, pp.2181–2192.

Shankar, V., Kumar, V., Devagade, U., Karanth, V. and Rohitaksha, K. (2020) 'Heart disease prediction using CNN algorithm', *SN Computer Science*, Vol. 1, No. 170, pp.1–8.

Tama, B.A. and Rhee, K-H. (2018) 'In-depth analysis of neural network ensembles for early detection method of diabetes disease', *International Journal of Medical Engineering and Informatics*, Vol. 10, No. 4, pp.327–341.

Wu, J-H., Li, J., Wang, J., Zhang, L., Wang, H-D., Wang, G-L., Li, X-l. and Yuan, J-X. (2019) 'Risk prediction of type 2 diabetes in steel workers based on convolutional neural network', *Neural Computing and Applications*, Vol. 32, No. 13, pp.9683–9698.

Young, B.A., Lin, E., Von Korff, M., Simon, G., Ciechanowski, P., Ludman, E.J., Everson-Stewart, S., Kinder, L., Oliver, M., Boyko, E.J. et al. (2008) 'Diabetes complications severity index and risk of mortality, hospitalization, and healthcare utilization', *The American Journal of Managed Care*, Vol. 14, No. 1, pp.15–23.

Zhang, J., Gong, J. and Barnes, L. (2017) 'HCNN: heterogeneous convolutional neural networks for comorbid risk prediction with electronic health records', in *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, IEEE, pp.214–221.