ITC 1/52
Information Technology and Control
Vol. 52 / No. 1 / 2023
pp. 215-227
DOI 10.5755/j01.itc.52.1.32008

Deep Learning Based Cardiovascular Disease Risk Factor Prediction Among Type 2 Diabetes Mellitus Patients

Received 2022/08/04

Accepted after revision 2022/11/08

https://doi.org/10.5755/j01.itc.52.1.32008

HOW TO CITE: Selvarathi, C., Varadhaganapathy, S. (2023). Deep Learning Based Cardiovascular Disease Risk Factor Prediction Among Type 2 Diabetes Mellitus Patients. *Information Technology and Control*, 52(1), 215-227. https://doi.org/10.5755/j01.itc.52.1.32008

Deep Learning Based Cardiovascular Disease Risk Factor Prediction Among Type 2 Diabetes Mellitus Patients

C. Selvarathi
Department of Computer Science and Engineering, M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India-639113

S. Varadhaganapathy
Department of Information Technology, Kongu Engineering College, Erode, Tamil Nadu, India-638060

Corresponding author: cselvarathires@outlook.com

Type 2 Diabetes Mellitus (T2DM) is a common chronic disease that is caused due to insulin discharge disorder. Due to the complication of T2DM, the outcomes of this disease lead to severe illness, death and cardiovascular disease (CVD). Given a larger number of diabetes patients, it is necessary to find the patients with a high risk of CVD complications. For this, the traditional methods are not sufficient and it is important to develop a deep learning-based efficient quantitative model to predict the risk of CVD among diabetes patients. The major objective of this research is to assess the efficient artificial intelligence approach toward the proposal of a personalized deep learning model that can able to predict the risk of fatal and non-fatal CVD among T2DM patients. First, the unbalanced dataset is preprocessed to make the dataset balanced for processing. Second, the features are reduced and important features are selected using Rank based Feature Importance (RFI) model which will improve the prediction accuracy. Third, the proposed Cascaded Convolution Graph LSTM (CCGLSTM) has been used as a classifier to predict the risk of CVD. Novelty of the work resides on ranking based feature analysis is cascaded with CGLSTM. The proposed model is implemented and experimented with various evaluation metrics using the data from 560 patients of five-year follow-up with T2DM. These evaluated results are compared with the state-of-the-art methods and the proposed model is proven to be superior to other approaches in terms of AUC (0.989), Accuracy (98.8%), recall (96.7%), precision (96.8%), specificity (97.4%) and F1-Score (97.5%).

KEYWORDS: Cardiovascular disease (CVD), Type2 Diabetes Mellitus (T2DM), Deep learning, cascaded, Long short term memory (LSTM), Convolution, Evaluation metrics.

1. Introduction

Diabetes is the most common disease that affects 90% of people worldwide and it is the major risk factor for Cardiovascular Disease (CVD) and renal dysfunction. Type 1 diabetes mellitus (Type 1 DM) is a kind of diabetes disease that affects younger's and grownups. The organ pancreas is located in the area of the midriff that stops insulin creation in the body. To control the sugar level, insulin is widely used by the patients. Type 2 DM is a non-secondary illness that affects grownups. The symptom of T2DM is family history, overweight, heftiness, undesirable eating regimen, and smoking. Pre-diabetes is defined as the before stage of T2DM when having a glucose level of more than sort 2 levels. Gestational DM is a diabetes illness that affected ladies and pregnant women. In the long run, DM may cause various illness that affects the heart, nervous systems, kidneys, retina, and other internal organs.

DM patients are suffered 2 to 4 times more from myocardial infarction, coronary heart disease, angina pectoris, etc., than those who do not have DM [24, 37]. The management of T2DM with the calculation of CVD risk may guide to take necessary treatment initiation. The treatment of CVD risk factors will reduce CVD occurrences and its burden on the economy. The clinical guidelines on the prevention of CVD risk can recommend the care given to evaluate the CVD risk factor of the patients that warrant the treatment [28]. It is necessary to build computational models to predict the CVD risk factors among diabetes patients for better diagnosis and treatment. The T2DM international guidelines management estimates the CVD risk for suitable treatment [38].

In general, there are various methodologies related to statistics and the machine learning field includes logistic regression, artificial neural network (ANN), decision trees, support vector machines, and Bayesian networks applied to predict clinical outcomes [5, 14, 30]. Due to the simplicity and good prediction capability, ANN models are widely accepted models for risk prediction. It can capture relationships among the data that are applied for various medical diagnoses. Different machine learning approaches were applied to determine the complication of CVD risk in diabetes patients [46]. Nowak et al. [33] proposed a CVD risk detection system using Gradient Boosting and LASSO cox regression with multi-protein arrays. This model could able to identify the T2DM patients with risk of CVD. Baum et al. [7] studied the impact of weight loss to reduce the CVD risk of T2DM using a causal forest approach and suggested the model has improved accuracy. However, the current methods are based on machine learning-based approaches to predict the risk factors. This motivates to choose deep learning-based models for research current research area and the efficiency of the deep learning approach can increase the prediction accuracy for the CVD risk among diabetes. With this motivation, this paper concentrate on developing a deep learning-based for CVD risk prediction. The major contribution of this work is as follows:

- Feature Selection: In this proposed system, Rank based feature importance (RFI) method is used as a feature selection method.
- Classification: the proposed cascaded convolution graph LSTM (CCGLSTM) based neural network with ranked features has been trained to predict the CVD risk among diabetes patients using the selected risk factor variables.
- Evaluation: the proposed model is evaluated using a confusion matrix and the results are compared with the existing approaches to prove the proposed model's efficiency.

The rest of the paper is organized as follows: works of literature related to the CVD risk from diabetes patients are reviewed in Section 2. The dataset used for this study has been described in Section 3. The proposed materials and methods using deep learning models are introduced in Section 4. Experimented results are discussed through illustrations in Section 5. The proposed model outcomes and its disadvantage with future extensions are concluded in Section 6.

2. Review of Literature

This section discusses the related work of the CVD risk prediction system for diabetes patients. Chu et al. [13] developed deep neural network (DNN) based CVD risk prediction systems among T2DM patients. DNN has been used to train and test datasets with the best performance. The receiver operating characteristics curve (ROC) was used for evaluation. Among the participants, 272 patients were diagnosed with CVD risk with a ROC value of 0.91. This model secured an accuracy of 87.5%, sensitivity of 88%, and specificity of 87.2%. The top risk factors of T2DM patients that influence CVD are Body mass index, depression, anxiety, systolic blood pressure, and total cholesterol. Longo et al. [26] predict the CVD complications called major adverse CVD from the administrative claims of 214676 patients with diabetes. They used hospitalization and pharmacy claims with patient information for their analysis. The obtained AUROC is 0.812.

Dinh et al. [16] used supervised ML approaches to predict the patients with a CVD risk factor. For their study, they used NHANES (national health and nutrition examination survey) dataset. They used all the features to predict the risk of CVD, diabetes, and pre-diabetes diseases. Abdalrada et al. [1] predict the co-occurrences of diabetes and CVD using ML models. They used DiScRi (Diabetes complications screening research initiative) dataset for their study. Initially, the risk factors of DM and CVD are determined using logistic regression and Evimp functions. These models are applied in the multivariate adaptive regression spline model. The redundancy is reduced through a correlation matrix. Next, a classification and regression algorithm was developed. The obtained accuracy was 94.09% with a specificity value of 95.8%.

Kibria and Matin [9] developed fused ML approaches for the prediction of binary and multi-class CVD. The ML approaches such as ANN, SVM, decision tree, logistic regression, Random forest, and Adaboost were applied to predict the disease. The class imbalance is handled using the random over sampler method. Kibria et al. [8, 10] proposed a decision-level fusion approach for the classification of heart disease. They fused two ML approaches to produce the best result. Instead of decision level fusion, they used weighted score fusion to improve the accuracy.

Birjais et al. [11] used K nearset neighbour (KNN) based imputation approach to handle the missing values. The classifiers such as gradient boosting, naïve Bayes, and logistic regression were used to predict the diabetes influence on CVD. Based on the evaluation using accuracy, sensitivity and specificity, gradient boosting performs better than other approaches for prediction. Aggarwal et al. [2] developed fuzzy inference with machine learning (ML) approaches for risk prediction of Covid 19 in diabetes patients. The most influential eight parameters are taken as input and 15 ML approaches were used based on a rule base. Among them, the CatBoost classifier gives better results with 76% of accuracy. Pal et al. [27] developed a CVD risk prediction system using two ML approaches such as multi-layer perceptron (MLP) and k nearest neighbor (KNN) using a public dataset. The experimental result secured 82.47% of accuracy and an AUC of 86.41%.Abdalrada et al. [40] developed cardiac autonomic neuropathy prediction using ML approaches from diabetes patients. This model secured a ROC value of 0.962, 87% of accuracy, and 87.12% of sensitivity.

Monitoring the mental ability of population during covid 19 with intelligent algorithm is implemented using social network [42, 43]. The fog computing is used to store and manage the resources. The opinion mining, mental or psychiatric issues are addressed using machine learning with high accuracy [41, 18]. Information security in big data is recently addressed using blockchain technologies [32, 39, 23, 35]. Various diseases diagnosis using hybrid machine and metaheuristic techniques are successfully implemented with high accuracy [19, 6, 44].

3. Data Set Description

For the development of the proposed model, this paper used the data from medical records of 560 Type 2 diabetes patients that were collected for five years follow-up at HippoKraion general hospital with the period from 1996 to 2007 [46]. This dataset consists of 41 patients out of 560 Type 2 diabetes patients who developed fatal or non-fatal CVD during five years of follow-up. Among the 41 CVD patients, four patients have a stroke, and the rest of the persons experienced Coronary heart disease (CHD). The presence of fatal or non-fatal CVD (positive instance) is coded as the binary value 1 and non-CVD (negative instance) is coded as 0. Table 1 summarizes the risk factors of CVD of T2DM patients along with their medical analyzed data.

Based on various studies, these identified risk factors are influences CVD in T2DM patients. With

Table 1

Fatal or non-fatal CVD risk factors in T2DM patients

Risk factors	Average ± Std. Dev	Variable type
Age	58.5±10.7 (years)	Continuous
Diabetes duration	7.6±7 (years)	
BMI (Body mass index)	29.4±5.5	
Glycosylated Hemoglobin	7.4±1.8 (%)	
PP (Pulse Pressure)	56.7±15.8 (mmHg)	
FG (Fasting Glucose)	165±56 (mg/dL)	
TC (total cholesterol)	226.6±50 (mg/dL)	
Triglycerides	167±110.8 (mg/dL)	
HDL cholesterol	48±16.4	
Medical analysis	**No. of Patients (%)**	**Variable type**
Smoking habit		Categorical
Non-Smokers	286 (51.6%)	
Current smokers	146 (26%)	
Previous smokers	125 (22.3%)	
Sex		
Male	263 (46.9%)	
Female	297(53%)	
Hyper tension	260 (46.4%)	
Lipid-lowering therapy		
No	469 (83.7%)	
Statins	74 (7.8%)	
Fibrates	17 (3%)	
Aspirin		
No	509 (90.8%)	
100 mg	44 (7.8%)	
325 mg	7 (3%)	
Insulin		
No	494 (88%)	
Yes	66 (11%)	
Parental history of DM		
No	304 (54%)	
Yes	256 (45.7%)	

each decade, age is an important factor that increases the CVD risk [17]. Diabetes duration and BMI also increase the risk of CVD. Over two to three months, HbA1c is the independent and important risk factor for CVD [12]. The relation between PP and CHD is irrelevant and patients with more than 45 to 55 mmHg have a risk of CHD [31]. FG levels and abnormal TC including high LDL and low HDL occur in patients who have premature CHD [29, 4]. It is proven that an active smoking person has a risk of CVD [34] and also men are having a maximum risk of CVD than females and females with diabetes are influenced by heart disease as twice as males [36]. Hypertension is the highest relevant risk of CVD [22]. Lipid-lowering therapy and aspirin are the protective factors that influence CVD [20]. Various research trials show that insulin does not affect the CVD risk and parental diabetes is the lowest risk of CVD in T2DM [25].

3.1. Training and Testing Dataset Ratio

The entire diabetes dataset is divided into two parts training dataset and a testing dataset with a ratio of 80:20, respectively. Based on the cross-validation, the training set is further divided into training and validation datasets in the ratio of 80:20, respectively.

4. Proposed Methodology

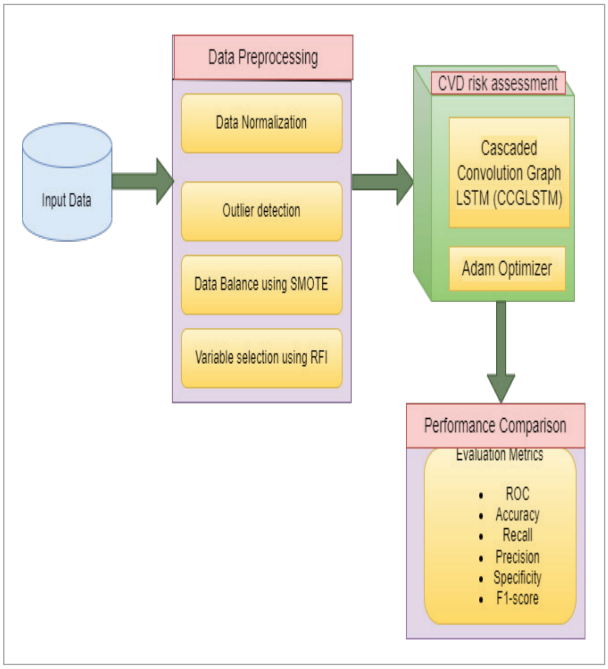
The proposed CVD risk prediction system is shown in Figure 1. It consists of three stages of processing such as (i) Preprocessing using outlier treatment, normalization, and dataset balance: in this stage, the input raw data are preprocessed to make the data balanced and efficient for further processing (ii) Variable selection using Rank based Feature Importance (RFI) model: this stage is used to select the most important and relevant variables based on its rank to improve the prediction accuracy and (iii) prediction using proposed Cascaded Convolution Graph Long short term memory (CCGLSTM) neural network which is optimized by Adam Optimizer: this section predicts the risk of CVD from T2DM patients data and classify it belongs to fatal CVD or non-fatal CVD.

4.1. Preprocessing

The efficient prediction systems have optimal data preparation and data preprocessing approaches. The

Figure 1

Overview of proposed CVD risk assessment from T2DM patients



major issues encountered in data preparation are outlier detection and missing values. The missing values in the data will reduce the reliability of the system and it needs to be replaced with appropriate values from statistical methods or ignored. In this work, the missing values are replaced by using the normalization approach. The second issue is outlier treatment. Outliers are the values that lie outside compared to the other data points and it is extremely abnormal to process. It causes the proposed system to errors and produces overestimated or underestimated results. The detected outliers are treated using a weight adjustment approach. Next, the imbalanced data are converted into balanced data using SMOTE model. Once the preprocessing is over, the raw dataset becomes balanced and efficient data for further processing.

4.1.1. Normalization

The raw data is inconsistent to process which will reduce the classifier performance. These raw data are transformed into an understandable format using the min-max normalization method denoted in Equation (1). Now, each data points have a similar range of values that lies between 0 and 1 the here minimum value of column 0 (D_{min}) and the maximum value of 1 (D_{max}) concerning the other column in the dataset.

$$D_{normalization} = \frac{D - D_{\min}}{D_{\max} - D_{\min}}. \tag{1}$$

4.1.2. Outlier Treatment

The outliers in the dataset are handled using the weight adjustment approach. Let us consider 'w' is the weight assigned to the data and the outlier data weight is considered as 'ω'. The weight computation and outlier treatment are denoted in Equations (2)-(3), respectively.

$$w^{\omega} = w * \omega \tag{2}$$

$$w^{\omega} = \begin{cases} 1\, for\ outliers \\ w\left(\frac{n}{n-k}\right) = w\left(1 + \frac{k}{n-k}\right) for\ non-outliers \end{cases} \tag{3}$$

where n is the total number of data and k is the number of outliers.

4.1.3.Dataset Balancing Using SMOTE

The collected dataset is an imbalanced dataset which causes the overfitting issue. That is the dataset is work well in training data and impact the testing data performance. To overcome this issue, SMOTE method has been used. It selects the data points subset from the minority class and compute k nearest neighbor instance from needed number of sampling. This data is added to the real dataset D. for each data 'x' in the dataset X, k nearest neighbor is computed as $X_{knn}(x)$. The new point 'p' is selected by searching 'p' in each segment from x to x_j which have the maximum distance from class c_i computed using Equation (4)

$$p = argmax_{p \in x, x_j} \frac{1}{k} \sum_{x \in c_i} \left\| p - x \right\|. \tag{4}$$

After preprocessing, the dataset is now complete and balanced for further processing to assess the risk of CVD among T2DM patients.

4.1.4. Variable Selection Using RFI

The features that are irrelevant and less important will degrade the classifier's performance. Those features are identified and the features that are important to train the model are selected using rank based feature importance method. This relevant feature set increases the classification accuracy with reduced training time. This approach consists of a group of decision trees to find the significant feature subset among the feature set. The steps for this variable/feature selection are as follows:

Algorithm 1: Variable selection using RFI

Input: balanced data set X

Output: Selected most relevant features subset

Step 1: Initially all the features in the dataset are selected

$f_x = f_1, f_2, f_3, \dots f_{x-1}$

Step 2: The model is defined to find the rank of features

Step 3: declare the number of decision trees needs to built and number of features as considered as random sample

Step 4: entropy is computed using Equation (5)

$$Entropy\left(subsetS\right) = \sum_{i=1}^{n} p_i \log_2 \, p_i, \tag{5}$$

where p is the number of instances, n is the number of classes (0 or 1)

Step 5: Information Gain(IG) from entropy is computed as in Equation (6)

$$IG\left(S, A\right) = Entropy\left(S\right) - \sum_{v \in values\left(A\right)} \frac{\left|S_v\right|}{\left|S\right|} . Entropy\left(S\right) \tag{6}$$

Step 6: the total score is calculated by aggregating each feature score with other feature using Equation (7):

$$totalscore = \sum_{i=1}^{Nimpacts} \sum_{j=1}^{N} \left(IG(A_{i)}, IG\left(A_j\right)\right), \tag{7}$$

where N is total number of features, A is attribute.

Step 7: the highest value features are ranked in ascending order and considered as relevant variables. For the considered dataset, the relevant feature with its score, rank are shown in Table 2.

Table 2

Variables score and rank using RFI

Risk factors	Score	Rank
Age	0.562	2
Diabetes duration	0.316	8
BMI (Body mass index)	0.524	3
Glycosylated Hemoglobin (HbA1C)	0.483	4
PP (Pulse Pressure)	0.401	6
FG (Fasting Glucose)	0.235	11
TC (total cholesterol)	0.442	5
Triglycerides	0.191	13
HDL cholesterol	0.211	12
Smoking habit		
Non-Smokers	0.082	24
Current smokers	0.378	7
Previous smokers	0.111	20
Sex		
Male	0.291	9
Female	0.268	10
Hyper tension	0.631	1
Lipid lowering therapy		
No	0.067	25
Statins	0.132	17
Fibrates	0.145	16
Aspirin		
No	0.097	22
100 mg	0.124	19
325 mg	0.127	18
Insulin		
No	0.091	23
Yes	0.183	14
Parental history of DM		
No	0.102	21
Yes	0.167	15

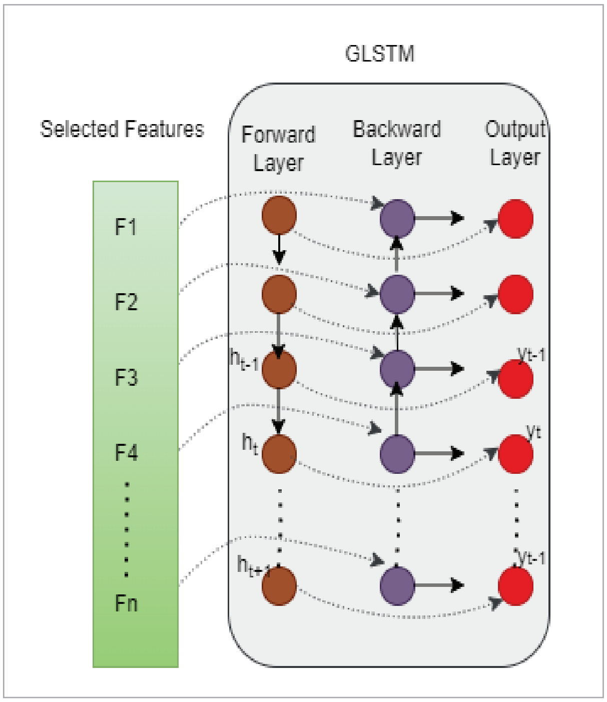
From Table 2, one can understand that the features with increased score is ranked in ascending order and the risk factors that influence the CVD at maximum possibilities are considered as significant variables. The variables with first seven ranking such as Age, BMI, Hb1Ac, Pulse pressure, cholesterol, current smokers, and hyper tension are chosen as significant features and used for classification.

4.2. Classification Using Proposed Cascaded Convolution Graph LSTM (CCGLSTM)

The selected variables of CVD risk among diabetes patients are fed as input to the classifier for the prediction of positive or negative response of CVD influence using the proposed Cascaded Convolution GLSTM. The standard LSTM consists of four gates includes forget gate (f), input gate (i), control gate (c) and output gate (o) with memory cell [48]. The architecture of Graph LSTM (GLSTM) is shown in Figure 2. In traditional LSTM, there is no correlation between previous and current memory cell which will reduce the prediction system performance while the output gate is closed. To overcome this issue, convolution based GLSTM has been proposed.

Figure 2

Structure of GLSTM



Compares to traditional LSTM, graph LSTM adds tree node at each single LSTM in both forward and backward directions. In forward direction, previous node history are captured and in backward direction, the node responses are captured. Input for this model is previous cell state called h_{t-1}, input feature x_t and bias b with extra convolution link. The output of this model c_t to represent present memory content and s_t represents the current cell state. Figure 3 illustrates the structure of CGLSTM.

The input node xt consists of both submission context and comment text with timing and hierarchical layer. Standard LSTM [47] has been used to represent the node transition as vector. The additional parameter ct-1 has been as to the link of GLSTM which makes CGLSTM. If t has no child node, no previous node, and no next node then the child node pointer, forward and backward pointer are set as null. Each variable selected from Section 4.2 are represented in the input layer. The selected seven features represent seven input neuron for the neural network. With this input, the data is processed to find the risk of CVD influence in T2DM using the following CGLSTM. In forward CGLSTM, the gate such as input gate i_t, temporal forget gate f_t, hierarchical forget gate h_t, control gate c_t and output gate $o_t$are updated using Equations (13-17).

$$i_t = \sigma\left(w_i x_t + U_i h_{t-1} + V_i c_{t-1} b_i\right) \quad (13)$$

$$f_t = \sigma\left(w_f x_t + U_f h_{t-1} + V_f c_{t-1} + b_f\right) \quad (14)$$

$$c_t = f_t \times c_{t-1} + i_t \times \sigma_h (w_c x_t + U_c h_{t-1} + b_c \quad (15)$$

$$h_t = o_t \times \sigma_h c_t \quad (16)$$

$$o_t = \sigma_g\left(w_o x_t + U_o h_{t-1} + V_o c_{t-1} + b_o\right), \quad (17)$$

where σ is sigmoid function. In backward CGLSTM, the gates are updated using the Equations (18-22)

$$i_t = \sigma\left(w_i x_t + U_i h_{t+1} + V_i c_{t+1} + b_i\right) \quad (18)$$

$$f_t = \sigma\left(w_f x_t + U_f h_{t+1} + V_f c_{t+1} + b_f\right) \quad (19)$$

$$c_t = f_t \times c_{t+1} + i_t \times \sigma_h (w_c x_t + U_c h_{t+1} + b_c) \quad (20)$$

$$h_t = o_t \times \sigma_h c_t \quad (21)$$

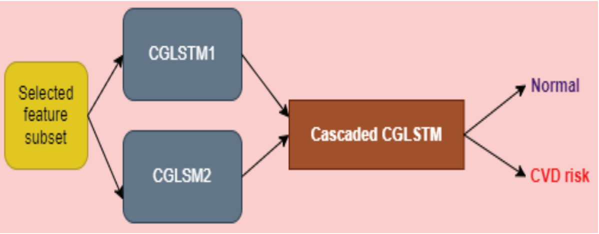
$$o_t = \sigma_g\left(w_o x_t + U_o h_{t+1} + V_o c_{t+1} + b_o\right) \quad (22)$$

The weights in the proposed CGLSTM model have been further optimized with an Adam optimizer for better performance. The convolution GLSTM is cascaded to make the network deep which predicts the

CVD influences in diabetes patients with improved accuracy. The structure of CCGLSTM is shown Figure 3.

Figure 3

Proposed CCGLSTM



The CCGLSTM produce two decision of the risk of CVD from each CGLSTM, respectively. The final result is chosen from voting scheme of estimated risk of CVD. For every testing instance, the maximum as well as minimum decision was selected based on the decision from both classifiers were lower or greater than a certain threshold of 0.2. In some cases, this decisions are chosen using DWC (Dynamic Weighting based on Certainties) method [45].

5. Experimented Results and Discussions

This section experiments the proposed RFI feature selection based CCGLSTM model for the prediction of CVD risk among diabetes patients [21, 3] using the considered dataset. The methods are implemented in Scikit package in Tensorflow library of python. This evaluation could find the answers for the following questions:

- Q1: How the risk factor variables are important to predict the CVD risk among the diabetes patients or it is necessary to have all the features for prediction?
- Q2: How the proposed model is accurate to predict the CVD risk compared to the conventional approaches?

5.1. Evaluation Criteria

The performance of the prediction system is measured using discrimination and calibration evaluation. It is the measure with the ability to separate the patients having disease risk from those who do not have by providing the improved score value of the former one. To calculate this, the reliable and popular metric is receiver operating characteristics curve (ROC). It is 100% denotes perfect ability of discrimination. And AUC of 50% indicates worst performance. With addition to AUC, the metrics are accuracy, sensitivity (recall), specificity, precision and F1-score are computed from the confusion metrics shown in Table 3.

Table 3

Confusion matrix

	Actual positive	Actual negative
Predicted positive	True positive (T^P)	False positive (F^P)
Predicted negative	False negative (F^N)	True negative (T^N)

The specificity is the ratio between actual negatives predicted as negative. It is also known as true negative rate as shown in Equation (23)

$$Sp = \frac{T^N}{T^N + F^P} \tag{23}$$

$$recall = \frac{T^P}{T^P + F^N} \tag{24}$$

$$Precision = \frac{T^P}{T^P + F^P} \tag{25}$$

$$Accuracy = \frac{T^P + T^N}{T^P + T^N + F^P + F^N} \tag{26}$$

F1 score is the mean between precision and recall as in Equation (27)

$$F1 - Score = \frac{2 * T^P}{2 * T^P + F^P + F^N} \tag{27}$$

Calibration measures how the predicted results are close to the actual probabilities. The impact of clinical model is assessed by the calibration measure called Net benefit criteria [15] which is denoted in Equation (28)

$$Net\ Benefit = \frac{T^P}{N} - \frac{F^P}{N} \cdot \left(\frac{P}{1-P}\right), \tag{28}$$

where N is the number of patients and P is the threshold probability. The threshold probabilities is varied for Net Benefit computation based on decision curve. Using this decision curve analysis, the P value is identified. The parameter settings of the NN are shown in Table 4.

Table 4

Hyper Parameter setting of classifiers

Parameter	CLSTM	CCGLSTM
Learning rate	0.02	0.01
Batch size	35	35
Hidden layers	50	50
Number of epoch	100	100

5.2. Evaluation in Terms of Feature Selection

This section answers the first question where the RFI model has been used to select the relevant and most important risk factor variables from the whole variables. These selected features are fed as input to the classifier. The features that degrade the prediction model performance are removed based on ranking scheme. Based on RFI, the ranking of risk factor variables are shown in Figure 5.

From the illustration of Figure 4, the variables are ranked with the features having highest scored. Among the 25 features, the first seven rank features such as Age, BMI, Hb1Ac, Pulse pressure, cholesterol, current smokers, and hyper tension were selected for further processing. The metrics of proposed classifier with all the features and reduced features is shown in Table 5. Compare to the classifier performance with all the features, the reduced feature selection will enhance the classifier performance efficiently. With the selected risk factor variable sets, the proposed model secured the accuracy of 98.81%.

Figure 4

Feature ranking using RFI

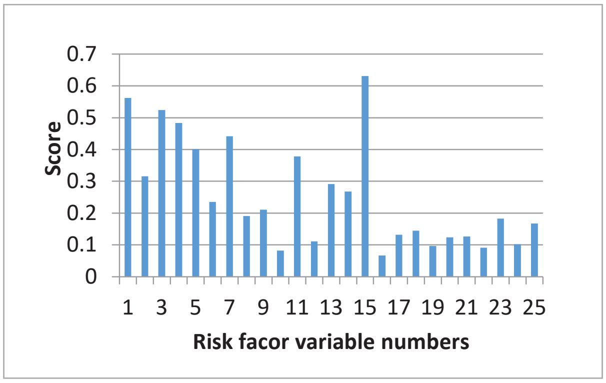


Table 5

Performance evaluation of the proposed prediction system with respect to feature selection

Evaluation Metrics	Input features	
	Original risk factor variables	Selected variables using RFI
No of selected features	25	7
Accuracy	92.5	98.81
Precision	91.2	96.72
Recall	92.9	96.81
Specificity	93.1	97.45
F1-Score	93.4	97.53
Training time (s)	94.9	3.31
Testing time (s)	182.32	7.22

5.3. Evaluation in Terms of Classification System

The proposed CCGLSTM performance in terms of AUC is compared with the other conventional approaches such as Traditional LSTM (TLSTM), Convolution LSTM (CLSTM), Convolution GLSTM (CGLSTM), Ensemble Hybrid wavelet neural network (HWNN) [46] with Self organizing map (SOM) for analysis. The confusion matrix for the proposed model evaluation is shown in Figure 5. Based on this confusion matrix outcome, other metrics are computed. The results for ROC computation are illustrated in Figure 6. Compare to the other approaches, the proposed RFI-CCGLSTM secured improved ROC value of 0.989. Various the other approaches secured, 0.71, 0.74, 0.84 and 0.69, respectively.

The comparison of other metrics such as Accuracy, Precision, Recall, Specificity, F1 score are shown in Table 6. The proposed CVD risk prediction system secured improved accuracy of 98.8% than other approaches such as TLSTM, CLSTM, CGLSTM and

Figure 5

Confusion matrix for proposed CVD risk Prediction system

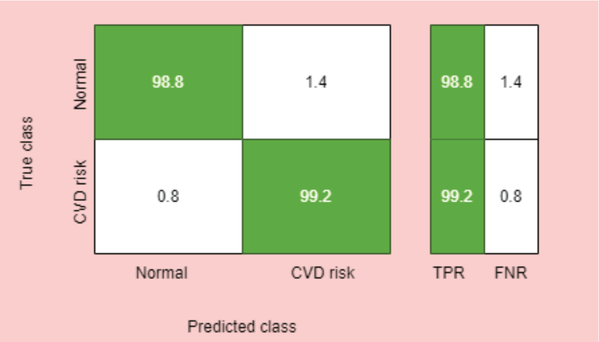


Figure 6

Classification system performance in terms of AUC

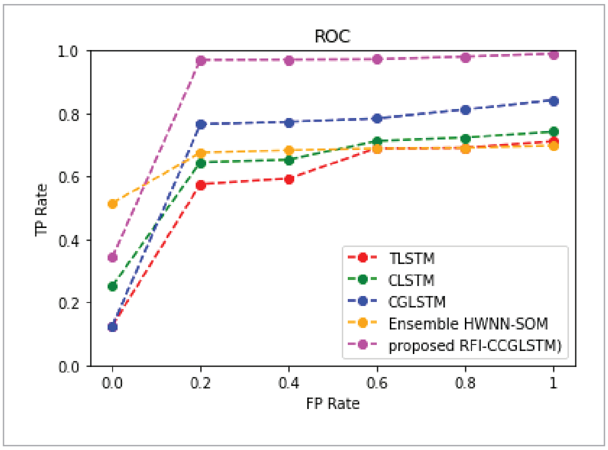


Table 6

Comparative analysis of proposed vs existing heart disease prediction systems

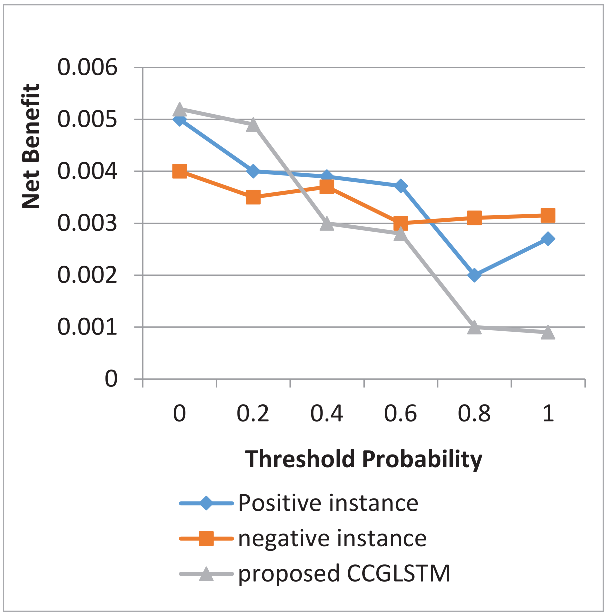
Classifiers	Precision	Recall	Specificity	F1-Score	Accuracy
TLSTM	0.81	0.86	0.88	0.90	0.92
CLSTM	0.85	0.88	0.90	0.91	0.93
CGLSTM	0.87	0.89	0.91	0.92	0.95
Ensemble HWNN-SOM	0.91	0.92	0.96	0.93	0.95
Proposed RFI-CCGLSTM	0.97	0.97	0.97	0.98	0.99

Ensemble HWNN with SOM which secured the accuracy of 92%, 93%, 94.9% and 94.7%, respectively. In terms of the other metrics also, the proposed model is superior to other approaches.

The calibration metric called Net Benefit across various threshold evaluations is shown in Figure 7. Net Benefit is computed across various threshold probabilities based on the risk detected by the proposed model (green color) and with the assumption of all patients are positive (blue) and negative (red) for CVD.

Figure 7

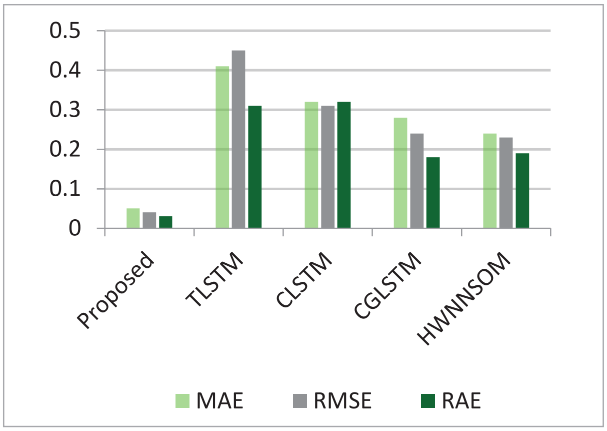
Evaluation of Net benefit metric of proposed model



The performance comparison in terms of error calculations such as Mean absolute error (MAE), Root Mean Square Error (RMSE) and Relative absolute error (RAE) are computed and compared with conventional approaches. The evaluated results are illustrated in Fig 8. Compared to other approaches, the proposed model secured minimum MAE, RMSE and RAE of 0.05, 0.04 and 0.03% of error. Various the other approaches such as TLSTM secured 0.41, 0.45 and 0.31, respectively. CLSTM secured 0.32, 0.31 and 0.32, respectively. CGLSTM secured0.28, 0.24 and 0.18 and HWNNSOM secured 0.24, 0.23 and 0.19% of error, respectively. Hence, the proposed CVD risk prediction model efficiently predicts the risk of CVD among the diabetes patients with improved accuracy, ROC and reduced error rate.

Figure 8

Error evaluation of the CVD risk prediction Classifiers



6. Conclusion

CVD is highly prevalent to the patients with Diabetes complication particularly T2DM. The artificial intelligence approaches provides the automated prediction of risk of CVD among the T2DM patients. In this paper, the risk of CVD complication among the diabetes patients is predicted using deep learning model. Among the 25 risk factors of diabetes patients, seven features are ranked as highest using RFI and selected for further processing. Using the selected risk factors variables, the proposed cascaded CGLSTM is trained for the prediction of patients having risk of CVD. By experimenting the proposed model using the constructed dataset and the evaluation is performed based on the metrics. By comparing the efficiency of the proposed model with other conventional systems, the proposed RFI-CCGLSTM secured improved ROC value of 0.989 and the metrics such as accuracy, precision, recall, specificity and F1score as 98.8%, 96.7%, 96.8%, 97.4% and 97.5%, respectively. Hence, the proposed risk of CVD prediction among the T2DM patients is predicted efficiently and perform superior to other approaches on the considered dataset. In future, the larger datasets of the patients with various ethnicity is considered for the proposed model applicability to the other cohorts of patients. With the addition to the traditional risk factors, the psychological factors are also important factors that influence the CVD among the diabetes patients. Those psychological factors are also considered for our further research.

References

1. Abdalrada, A. S., Abawajy, J., Al-Quraishi, T., et al. Machine Learning Models for Prediction of Co-occurrence of Diabetes and Cardiovascular Diseases: A Retrospective Cohort Study. Journal of Diabetes and Metabolic Disorders, 2022, 21, 251-261. https://doi.org/10.1007/s40200-021-00968-z
2. Aggarwal, A., Chakradar, M., Singh, M., Manojkumar, B., Thompson, S., Gupta, S. K., Alsamhi, S. H., AL-Dois, H. COVID-19 Risk Prediction for Diabetic Patients Using Fuzzy Inference System and Machine Learning Approaches. Hindawi Journal of Healthcare Engineering, 2022, Article ID 4096950. https://doi.org/10.1155/2022/4096950
3. Alade, O. M., Sowunmi, O. Y., Misra, S., Maskeliūnas, R., Damasevicius, R. A Neural Network Based Expert System for the Diagnosis of Diabetes Mellitus. Proceeding of Advances in Intelligent Systems and Computing, 2018, 14-22. https://doi.org/10.1007/978-3-319-74980-8_2
4. American Heart Association, Cardiovascular disease & diabetes, 2017.
5. Ayer, T., Chhatwal, J., Alagoz, O., Kahn, C. E., Woods, R., Burnside, E. Informatics in Radiology: Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation. Radio Graphics, 2010, 30, 13-22. https://doi.org/10.1148/rg.301095057
6. Bacanin, N., Zivkovic, M., Al-Turjman, F., Venkatachalam, K., Trojovský, P., Strumberger, I. Bezdan, T. Hybridized Sine Cosine Algorithm with Convolutional Neural Networks Dropout Regularization Application. Scientific Reports, 2022, 12(1), 1-20. https://doi.org/10.1038/s41598-022-09744-2
7. Baum, A., Scarpa, J., Bruzelius, E., Tamler, R., Basu, S., Faghmous, J. Targeting Weight Loss Interventions to Reduce Cardiovascular Complications of Type 2 Diabetes: A Machine Learning-based Post-hoc Analysis of Heterogeneous Treatment Effects in the Look AHEAD Trial. Lancet Diabetes Endocrinology, 2017, 5, 808-815. https://doi.org/10.1016/S2213-8587(17)30176-6
8. Binte, H., Abdul, K. M. An Efficient Machine Learning-based Decision-level Fusion Model to Predict Cardiovascular Disease. International Conference on Intelligent Computing & Optimization, Springer, 2020, 1097-1110. https://doi.org/10.1007/978-3-030-68154-8_92

9. Binte, H., Abdul, K. M. The Severity Prediction of the Binary and Multi-class Cardiovascular Disease - A Machine Learning-based Fusion Approach, arXiv:2203.04921v1, 2022.

10. Binte, H., Abdul, K., Sanzida, M. I. Comparative Analysis of Two Artificial Intelligence Based Decision Level Fusion Models for Heart Disease Prediction. International Semantic Intelligence Conference, CEUR Workshop Proceedings, 2020, 2786, 314-322.

11. Birjais, R., Kumar, A., Ritu Chauhan, M., Kaur, H. Prediction and Diagnosis of Future Diabetes Risk: A Machine Learning Approach. SN Applied Sciences, 2019, 1(9), 1112. https://doi.org/10.1007/s42452-019-1117-9

12. Cavero-Redondo, I., Peleteiro, B., Alvarez-Bueno, C., Rodri F., Aıguez Artalejo, Martiınez-Vizcaiıno, V. Glycosylated Haemoglobin as a Predictor of Cardiovascular Events and Mortality: A Protocol for a Systematic Review and Meta-analysis. BMJ Open, 2016, 6, 1-5. https://doi.org/10.1136/bmjopen-2016-012229

13. Chu, H., Chen, L., Yang, X., Qiu, X., Qiao, Z., Song, X., Zhao, E., Zhou, J., Zhang, W., Mehmood, A., Pan, H., Yang, Y. Roles of Anxiety and Depression in Predicting Cardiovascular Disease Among Patients With Type 2 Diabetes Mellitus: A Machine Learning Approach. Frontiers in Psychology, 2021, 12, 645418. https://doi.org/10.3389/fpsyg.2021.645418

14. Dalakleidi, K. V., Zarkogianni, K., Karamanos, V. G., Thanopoulou, A. C., Nikita, K. S. A Hybrid Genetic Algorithm for the Selection of the Critical Features for Risk Prediction of Cardiovascular Complications in Type 2 Diabetes Patients. IEEE 13th International Conference on Bioinformatics Bioengineering, 2013, 1-4. https://doi.org/10.1109/BIBE.2013.6701620

15. Dieren S. Prediction Models for the Risk of Cardiovascular Disease in Patients with Type 2 Diabetes: A Systematic Review. Heart, 2012, 98, 360-369. https://doi.org/10.1136/heartjnl-2011-300734

16. Dinh, A., Miertschin, S., Young, A., et al. A Data-driven Approach to Predicting Diabetes and Cardiovascular Disease with Machine Learning. BMC Medical Informatics and Decision Making, 2019, 19, 211. https://doi.org/10.1186/s12911-019-0918-5

17. Finegold, J. A., Asaria, P., Francis, D. P. Mortality from Ischaemic Heart Disease by Country, Region, and Age: Statistics from World Health Organization and United Nations. International Journal of Cardiology, 2012, 168, 934-945. https://doi.org/10.1016/j.ijcard.2012.10.046

18. Gautam, R., Sharma, M. Prevalence and Diagnosis of Neurological Disorders Using Different Deep Learning Techniques: A Meta-analysis. Journal of Medical Systems, 2020, 44(2), 1-24. https://doi.org/10.1007/s10916-019-1519-7

19. Hassan, M. R., Islam, M. F., Uddin, M. Z., Ghoshal, G., Hassan, M. M., Huda, S., Fortino, G. Prostate Cancer Classification from Ultrasound and MRI Images Using Deep Learning Based Explainable Artificial Intelligence. Future Generation Computer Systems, 2022, 127, 462-472. https://doi.org/10.1016/j.future.2021.09.030

20. Ittaman, S. V., VanWormer, J. J., Rezkalla, S. H. The Role of Aspirin in the Prevention of Cardiovascular Disease. Clinical Medicine Research, 2014, 12, 147-154. https://doi.org/10.3121/cmr.2013.1197

21. Jothi, Prakash, V., Karthikeyan, N. K. Dual-layer Deep Ensemble Techniques for Classifying Heart Disease. Information Technology and Control, 2022, 51(1), 158-179. https://doi.org/10.5755/j01.itc.51.1.30083

22. Kaplan, N. M. Cardiovascular Risks of Hypertension, 2016. [Online]. Available: http://www.uptodate.com/contents/cardiovascular-risksof-hypertension

23. Kaur, P., Sharma, M., Mittal, M. Big Data and Machine Learning Based Secure Healthcare Framework. Procedia Computer Science, 2018, 132, 1049-1059. https://doi.org/10.1016/j.procs.2018.05.020

24. Larsson, S. C., Wallin, A., Hakansson, N., Stackelberg, O., Back, M., Wolk, A. Type 1 and Type 2 Diabetes Mellitus and Incidence of Seven Cardiovascular Diseases. International Journal of Cardiology, 2018, 262, 66-70. https://doi.org/10.1016/j.ijcard.2018.03.099

25. Law, J. R. Association of Parental History of Diabetes with Cardiovascular Disease Risk Factors in Children with Type 2 Diabetes. Journal of Diabetes Complications, 2015, 29, 534-539. https://doi.org/10.1016/j.jdiacomp.2015.02.001

26. Longato, E., Fadini, G. P., Sparacino, G., Avogaro, A., Tramontan, L., Di Camillo, B. A. Deep Learning Approach to Predict Diabetes' Cardiovascular Complications From Administrative Claims. Journal of Biomedical and Health Informatics, 2021, 25(9), 3608-3617. https://doi.org/10.1109/JBHI.2021.3065756

27. Madhumita, P., Smita, P., Ganapati, P., Kuldeep, D., Ranjan, K., M. Risk Prediction of Cardiovascular Disease Using Machine Learning Classifiers. Open Medicine, 17(1), 1100-1113. https://doi.org/10.1515/med-2022-0508

28. Matheny, M. Systematic Review of Cardiovascular Disease Risk Assessment Tools. Agency Healthcare Res. Qual., Rockville, MD, USA, May 2011.

29. Mongraw-Chaffina, M. A Prospective Study of Low Fasting Glucose with Cardiovascular Disease Events and All-Cause Mortality: The Women's Health Ini-

tiative. Metabolism, 2017, 70, 116-124. https://doi.org/10.1016/j.metabol.2017.02.010

30. Mougiakakou, S., et al. SMARTDIAB: A Communication and Information Technology Approach for the Intelligent Monitoring, Management and Follow-up of Type 1 Diabetes Patients. IEEE Transactions on Information Technology in Biomedicine, 2010, 14(3), 622-633. https://doi.org/10.1109/TITB.2009.2039711

31. Nargesi, A. Nonlinear Relation Between Pulse Pressure and Coronary Heart Disease in Patients with Type 2 Diabetes or Hypertension. Journal of Hypertension, 2016, 34, 974-980. https://doi.org/10.1097/HJH.0000000000000866

32. Nguyen, H., Kieu, L. M., Wen, T., Cai, C. Deep Learning Methods in Transportation Domain: A Review. IET Intelligent Transport Systems, 2018, 12(9), 998-1004. https://doi.org/10.1049/iet-its.2018.0064

33. Nowak, C., Carlsson, A. C., Ostgren, C. J., Nystrom, F. H., Alam, M., Feldreich, T., et al. Multiplex Proteomics for Prediction of Major Cardiovascular Events in Type 2 Diabetes. Diabetologia, 2018, 61, 1748-1757. https://doi.org/10.1007/s00125-018-4641-z

34. Pan A., Wang, Y., Talaei, M., Hu, F. B. Relation of Smoking with Total Mortality and Cardiovascular Events Among Patients with Diabetes: A Meta-Analysis and Systematic Review. Circulation, 2015, 136, 795- 804.

35. Ramanan, M., Singh, L., Kumar, A. S., Suresh, A., Sampathkumar, A., Jain, V., Bacanin, N. Secure Blockchain Enabled Cyber-Physical Health Systems Using Ensemble Convolution Neural Network Classification. Computers and Electrical Engineering, 2022, 101, 108058. https://doi.org/10.1016/j.compeleceng.2022.108058

36. Regensteiner, J. G., Sex differences in the cardiovascular consequences of diabetes mellitus, Circulation, 2015, 132, 2424-2447. https://doi.org/10.1161/CIR.0000000000000343

37. Roobini, M. S., Lakshmi, M. Autonomous Prediction of Type 2 Diabetes with High Impact of Glucose Level. Computers and Electrical Engineering, 2022, 101, 108082. https://doi.org/10.1016/j.compeleceng.2022.108082

38. Ryden, L. Diabetes Pre-diabetes and cardiovascular Diseases Developed with the EASD. European Heart Journal, 34, 3035-3087. https://doi.org/10.1093/eurheartj/eht108

39. Saleem, M. H., Potgieter, J., Arif, K. M. Automation in Agriculture by Machine and Deep Learning Techniques: A Review of Recent Developments. Precision Agriculture, 2021, 22(6), 2053-2091. https://doi.org/10.1007/s11119-021-09806-x

40. Shaker, A., Jemal, A., Tahsien, A., Sheikh, A.-Q., Shariful, M. Prediction of Cardiac Autonomic Neuropathy Using a Machine Learning Model in Patients with Diabetes. Therapeutic Advances in Endocrinology and Metabolism. Therapeutic Advances in Endocrinology and Metabolism, 2022, 13, 1-10. https://doi.org/10.1177/20420188221086693

41. Sharma, M., Romero, N. Future Prospective of Soft Computing Techniques in Psychiatric Disorder Diagnosis. EAI Endorsed Transactions on Pervasive Health and Technology, 2018, 4(15), e1-e1. https://doi.org/10.4108/eai.30-7-2018.159798

42. Sharma, M., Sharma, S., Singh, G., Remote Monitoring of Physical and Mental State of 2019-nCoV Victims Using Social Internet of Things, Fog and Soft Computing Techniques. Computation Methods Programs Biomed, 2020, 105609-105609. https://doi.org/10.1016/j.cmpb.2020.105609

43. Sharma, M., Singh, G., Singh, R. Design of Ga and Ontology Based NLP Frameworks for online Opinion Mining. Recent Patents on Engineering, 2019, 13(2), 159-165. https://doi.org/10.2174/1872212112666180115162726

44. Tuba, E., Strumberger, I., Tuba, I., Bacanin, N., Tuba, M. Acute Lymphoblastic Leukemia Detection by Tuned Convolutional Neural Network. In 2022 32nd IEEE International Conference Radioelektronika (RADIOELEKTRONIKA), 2022, 1-4. https://doi.org/10.1109/RADIOELEKTRONIKA54537.2022.9764909

45. Valdovinos, R., Sanchez, J., Barandela, R. Dynamic and Static Weighting in Classifier Fusion. Pattern Recognition and Image Analysis, 2005, 3523, 59-66. https://doi.org/10.1007/11492542_8

46. Zarkogianni, K., Athanasiou, M., Thanopoulou, A. C. Comparison of Machine Learning Approaches Toward Assessing the Risk of Developing Cardiovascular Disease as a Long-term Diabetes Complication. Journal of Biomedical and Health Informatics, 2018, 22, 1637-1647. https://doi.org/10.1109/JBHI.2017.2765639

47. Zhou, J., Xiang, J., Huang, S. Classification and Prediction of Typhoon Levels by Satellite Cloud Pictures Through GC-LSTM Deep Learning Model. Sensors 2020, 20, 5132. https://doi.org/10.3390/s20185132

48. Zhu, G. Redundancy and Attention in Convolutional LSTM for Gesture Recognition. IEEE Transaction on Neural Networks Learning Systems, Jun. 2019.