2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA) | 978-1-7281-9437-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/CITISIA50690.2020.9371843

Predicting Early Phase of Type 2 Diabetic by Deep Learning

Prabir Pathak[1] and Amr Elchouemi[2]
[1]Study Group, Sydney, Australia
[2]American Public University System, USA
prabirpathak68@gmail.com; Amr.elchouemi@mycampus.apus.edu

Abstract— **Deep Neural Network with prediction is the one of the main deep learning technologies which has been used by many researches for early prediction of Type 2 Diabetics (T2D). For the prediction of the T2D, the taxonomy with the components are proposed with Data, Prediction processing and Display (DPD). Those components are evaluated for the better performance of the system and are validated with the different parameters for the early diagnosis of the T2D. The system being proposed has the higher accuracy for the prediction of the T2D and early detection of the diabetics in different age group in comparison to research paper reviewed and with current findings. It also helps to diagnose the diabetics in the patients. The critical analysis of the literature review of the latest published research paper available on the T2D and on deep learning has better accuracy for the prediction of T2D. On basis of the analysis, an effective system for T2D based on Deep Neural Network (DNN) has been developed in the system that can predict the diabetics in the early stage.**

Keywords- Deep Learning, Prediction, Type 2 Diabetics (T2D), Deep Neural Network

I. Introduction

Predicting the disease has become one of the focus areas of the medical sector for the detection of the T2D in the early phase. The technique is used to detect and predict the number of different diseases including Type 1 Diabetic, Diabetic Retinopathy, Cancers etc. The purpose of predicting is to process different kind of dataset of the patients. The process is accomplished using large sets of data with multiple standard set, such as diagnoses, medications, and laboratory tests and pre-diagnosed diabetics patients for the process of the detection of the T2D. The data are processed for the result. With this system, the pre-processed data, patient's diagnosis, and prediction algorithms are the necessary factors to create a working deep learning system. The *tools and data sets* are zipped into prediction, delivering the end users for detection of the T2D. For the review on this application, readers can refer to the review of the authors [1] .

In the latest system, we can find wide and deep network from deep learning have been proposed for the prediction of the T2D. In the generic domain the technology has been used to predict the diabetics and, in the subdomain, to find the accuracy of the prediction of T2D. The main limitation of the technology is that it is not able to predict some of the important risk factors [1] . Along with this, it lacks the auto feature selection method to process the large set of data [2]. Therefore, the purposed system in this paper will focus on Deep Neural Network for the early prediction of the T2D. This technique can be found on diabetic retinopathy which uses prediction method for the study of diabetes mellitus [3]. The taxonomy of Data, Prediction algorithms, Display (DPD) will be the components of the research paper for the review paper. The proper evaluation and the prediction algorithms for the early prediction will help the researches and the health sector and they can evaluate the components section for the improvements or any updates.

The purpose of this research paper is to search for the latest and most accurate research papers on deep learning for the prediction of the T2D. The available literature review is from deep learning focusing on different technique of deep model such as neural network, conventional network and so on. The papers also reviewed and focused on predicting different types of diabetes and the model which provide the maximum accuracy rate. The main objective of the research paper is to highlight the taxonomy of the paper which focuses on the deep learning model for the prediction of the T2D based on the datasets and on the feature selection and extraction of the data before processing the data for the prediction and evaluation before delivered to the end users. The purpose of defining such taxonomy is to solve the issue of the prediction of the T2D by using deep neural network which can processed and evaluated on the large or electronic datasets for the early detection of the T2D.

The research papers which focused on the deep learning of the prediction of the diseases specially for the diabetics has been focused in this paper. The taxonomy was classified by using 10 state-of-the-art publications which describe the deep neural network for the prediction method. The paper was selected from the active group and the authors form the database of 60 papers in the field of deep learning model. The system is verified by using accuracy to find the prediction level.

The rest of the paper will be followed by literature review section which will describe all the collected papers and their method along with their components being used. Then it will be followed by the state of art solution with its components. The 4th section will be composed of proposed model with the diagram and algorithms.

II. Literature Review

A. *Field of Research*

The gap on prediction of T2D is conducted under deep learning but has less accuracy data and most of the paper are found focusing on machine learning for the prediction of diabetic. The use of deep learning has been only been used on few papers for prediction of different diseases but only few for T2D.

B. *Review of previous paper*

In this paper, a total of 25 journal papers were reviewed. Most of the paper focused on the prediction of the diabetics and other disease for the early prediction.

978-1-7281-9437-0/20/$31.00 ©2020 IEEE

1. Deep Neural Network

[1] has enhanced the prediction technique by combining the linear model strength with different features. The deep neural network has helped to improve the prediction of onset of Type 2 Diabetics. The Accuracy rate of 84.13% denote the higher level then the individual rate. The modeling averaging ensemble is bettter and robust model and better in providing accuracy than any other single model. The technique has better solution on prediction with an accuracy of 84.13%. The solution also shows the importance of deep neural network and the model proposed by the research paper. The other testing tool - SMOTE has an improved sensitivity but has low improvement on other metrics [4]. Here the prediction process are tested for better accuracy with the electronic database and tested under Deep Neural Network and the purposed solution for the T2D prediction will be better at predicting the diabetics with higher accuracy and efficiency.

[5] carries a significant potential to shift and put the irregular data collected by extracting the meaningful which helps to boost efficency and effectiveness. This method helps to better use of EMRs for prediction of the diseases. They also provide a promising tool for any identification of the individual for any potential risk to the disease and a mean to help the health authority to adopt the efficient measure for prevention of diabetics. The project provides the potential risk identification of the individual which are prone to high risk. It also provides a guideline for the health department for the preventive measure of the diabetics.

[6] delivers the algorithm which addresses data pre-processing along with classification process. As the insensitivity of the imbalanced data, different performance indicators are considered to overcome error. It provides better performance than other algorithms on different classifier indicator for performance. It has shown great potential for predicting diabetics. This research paper shows better outcome for diabetic's prediction because of its algorithms for performance indicators such as accuracy, precision an AUC. The process of the imbalanced datasets has some error for processing the data, but the performance indicators will be able to manage the accuracy by ignoring the error.

[7] has enhanced the accuracy of prediction by working on the three model of prediction. The authors have also focused on risk factors which lead to the increase of a high-risk potential on patients. The paper works on graph theory and analysis of social network with comorbidity prevalence and transition pattern match to assign a score for reflecting the effectiveness of the method to predict the disease. As a result, the prediction system for any disease has increased and the capacity of the research paper has helped the health system to cope with the patient flow ratio. This solution dealing with the three components for prediction will certainly help to enhance the work in further research paper. This research provides a less expensive method of knowing the high-risk cohorts which in future can have diabetic.

[8] has enhanced the prediction and estimation of the diabetics among the different group of people. The authors have worked to develop a new model for detecting the undiagnosed model from using the data provided from a hospital. The solution was achieved by examining the data of 8000 patients with non-diabetics and 3845 diabetics patients. The combination of the five model has the AUC of 0.97. These algorithms have a prediction with random forest with highest accuracy (ACC = 0.8084). As a result, the prediction of the early patients using AUC model has an improved condition in health sector. The research paper provides the significant outcome of using machine learning components of combining five model and delivers the accuracy rate for comparison. The feature of using five models has also been focused on different research paper which gives certain degree of accuracy.

[9] enhanced the deep learning models to upgrade the prediction of diabetics in patient. The authors have used the Multimodal Deep Learning (MMDL) for overcoming lack of different model. The accuracy level with the MMDL while calculating mean rank was 1.15 and Super learner model has 4.32 which indicated the MMDL model is better performer as it has the lower rank. As a result, the prediction and detection of early phase of diabetics is possible with the correct scale of database provided. The MMDL provides the better accuracy on the patient health data and works on different deep model to provide nearest possible probability if any, in the patients.

[4] has enhaced the novel approach by supporting multiple predictive models. It can handle the data that has been missed at the prediction time and deliver the prediction. The approach holds the missing data by passing an alternative model which is equivalent too. As a result, this model does not require any missing model. This research provides more accuracy than the other few models by sorting the data. Its model does help to find missing data and calculate the prediction level from the data provided.

[2] has enhanced the prediction level and the accuracy of the prediction on the Koreans. The clinical model included the 10 selected risk factors such as SNPs, PTPRG, RIC1, TNED2, ADAM12 and CGNL1. The p-value of NRI for the SNP are significantly important for the point-based score. As a result, the clinical model has been able to predict the incidence among an assigned group of Korean and has a significant prediction. The solution provides reliable factor to be considered for predicting the outcome. The model suggested adding an SNP point which improves the testing power. The research has only covered a certain age group and has left patients from age group 25 – 40.

2. Patient2Vec

[10] enhances the prediction accuracy and interpretability. The Patient2Vec has a categorized mechanism which allows the clinical events to interpret with the weight. The framework model Patient2Vec has better performance than the baseline methods. The two terms sensitivity and F2 score are better performance for proposed framework. Along with this feature, learned feature can interpret both the Individual as well as population levels mobilized clinical insights. As a result, learning framework on recurrent neural networks and attention mechanism helps improve the prediction. This solution under the proposed framework provides the possible prediction of the disease and helps the health sector to improve the understanding of the disease correlations.

[11] enhances the techinical intuitions about the EHR and the approaches for the deep learning process. The reasearch worked on the tehnical side of the different efforts which have been applied in the clinical knowledge and its discovery using Electronic Health Record (EHR) of vast data set. EHR gives many information of the patients and the data processing delivers more accuracte components for prediciton [9]. As a result, deep learning has been able to handle the electronic health record by providing the prediction data or the patients management for the clinical process. This solution provides good knowledge for the deep learning on electronic health record and on different components of the deep learning which has the clinical knowledge using EHR with the deep data sets.

[12] enhances the knowledge of dataset being used for prediction of the disease. They focus on using the data set on an informative way rather than in poor quality context. The paper emphasizes on machine learning model for the prediction of the outcome. It also emphasizes on negative aspects of prediction and its wrong impact. This research paper has provided with better solution for the machine learning model. The use of machine learning can be challenging for some sector. The solution provides the deep understanding in dataset and its effective use, while minimizing the error for prediction.

[13] enhances the estimation of the diabetic prediction without initial medical diagnosis. The paper helps to predict the patients with undiagnosed diabetics to early medical care. The research paper is based on the non-invasive variable for analyzing the dataset. It provides the early prevention model for the diabetics' patients for the disease which can be remain undiagnosed. The area under curve (AUC) 80.11 has better performance than the other previously used model used for the research.

C. Best Solution

Among all the research paper and the methods purposed by the authors [1] describe the components of the Prediction algorithms for the taxonomy of the DPD on different stages (Data source - Feature extraction and selection, Prediction and Display). They used wide and deep neural model for the proceed data for assessing the accuracy and level of prediction of T2D. However, there is limitation on the paper with an auto feature selection method.

Our interpretation for the [1] DPD model in the prediction of the T2D is displayed in the Fig. 2. The process step starts with the use of raw datasets acquired from different sources. The raw data are obtained in an unclassified format and is further processed for feature extraction based on various factors such as BMI, age, gender and so on. The data is further analyzed under feature selection and further feature extraction with the data testing and validation from Nguyen research paper. The view of the predicting the T2D from the Deep Neural Network by analyzing and training the dataset will deliver the result which the user will find and interpret the data and the results in the requirement of the user. For the better understanding of the model proposed by the author example has been drawn at Fig. no 1. The Bind P. Nguyen, Hung N. Pham and Hop Tran analyzed the data in the testing and validation method that helps to extract the feature and further process for the ensemble model for prediction. The motive being held for the development of the technique which will be used in the research paper.

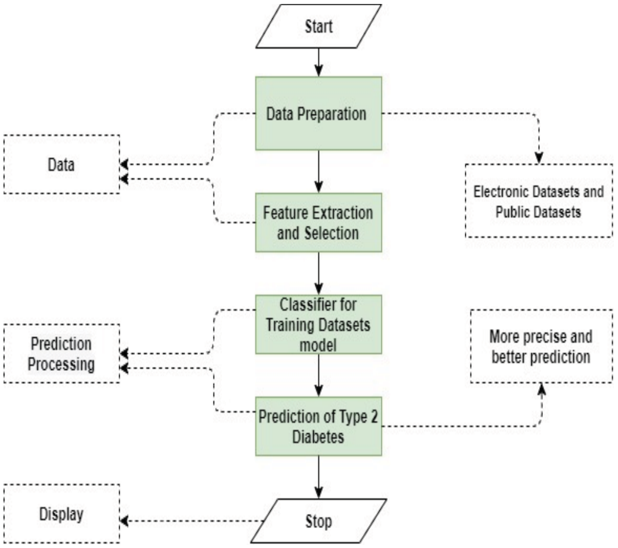
Start
Data Preparation
Data
Electronic Datasets and Public Datasets
Feature Extraction and Selection
Classifier for Training Datasets model
Prediction Processing
More precise and better prediction
Prediction of Type 2 Diabetes
Display
Stop

Fig. 1 The figure illustrates the factors for the prediction algorithm. The top to bottom shows the data been analyzed and process.

However, the prediction model of the deep model is based on the electronic health record which processes the prediction based on analyzing the dataset. The purpose of our work is to deliver a taxonomy to classify the technique and prediction accuracy applied in the domain area. The purposed taxonomy is based on three factors: 1) Data, which is collected from different sources and the feature are extracted, 2) Prediction Processing which gives the best possible technique to analyze the data and 3) Display, which display the results. In the following section, the taxonomy DPD will be discussed followed by their respective classes and their associated sub classes.

III. SYSTEM COMPONENT

The research paper has been developed based on the knowledge acquired from all the literature review process. The research paper has focused on the review of the Type 2 diabetics from Deep Learning - Deep Neural Network (DNN), for the early stage of prediction. Along with the literature review, the experts of the domain have helped for framing out the useful system for Type 2 Diabetic. The taxonomy of the system will be Data, Prediction Algorithm and Display (DPD) which has been developed through the possible review of the journal accessible. The taxonomy will be able to deliver a system that can be effective for the prediction, evaluation and then to provide indication of Type 2 diabetic onset.

The review section of the research for the Type 2 Diabetic and the technique search gave 408 journals. From the finalization, 30 papers were found matching inclusive criteria of the research: the frameworks which discuss about the

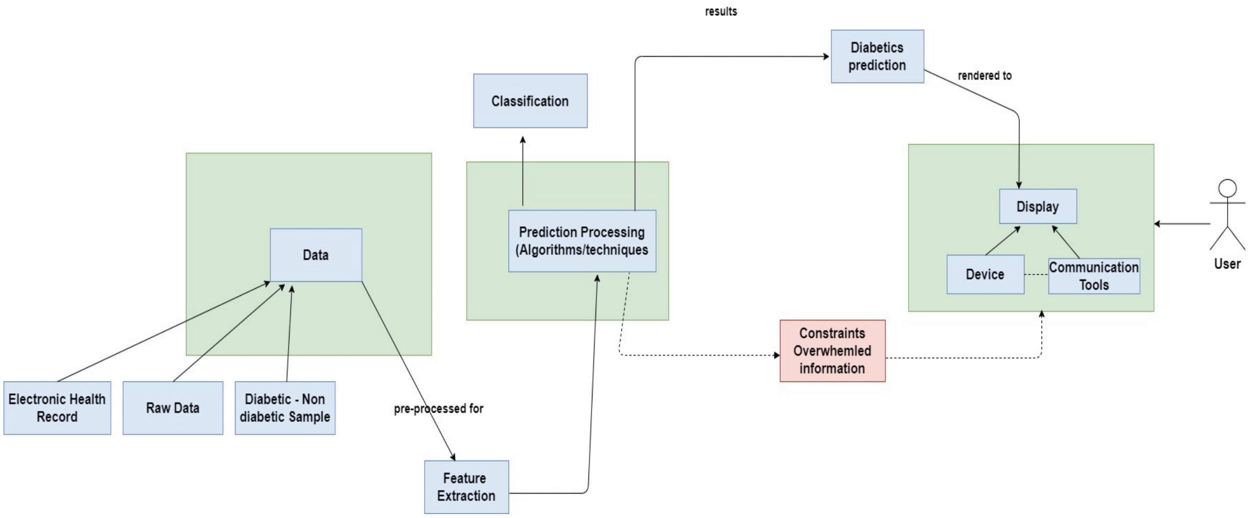

Fig. 2 The above figure illustrate the three factors of the taxonomy Data, Prediction Processing and Data - 'DPD' with their class and subclass. The line (dash line) relationship among them.

prediction of Diabetics (Type 2 Diabetic) using Deep Learning were included. The paper which worked on deep learning and prediction of the diabetics basically on the early stage of stage was given greater importance for the review. The most possible published journal was considered along with the level of publications (Q1 and Q2). Among the total number of the journals, 378 paper was rejected as it was not relevant to the inclusive criteria. 150 papers have not met the inclusive criteria, 40 paper were found focusing on various methodology than deep learning. Some of them, 28 paper were found written in various language and 35 were to level of Q2 and Q3. Also, papers discussed other technique in addition to deep learning on diabetics were excluded. After considering all those journals and its relevance on different section, 30 journals were selected for the research.

As per my understanding and knowledge and the review of the literature, the model which is proposed for the prediction of the type 2 diabetics in the early stage will help the health sector professionals to detect any chances of Diabetic in patient based on medical history. On this basis, three major points are considered. 1. Type of Data which will be used in the system 2. How the Prediction of Type 2 Diabetics can be carried out on the actual given context 3. The Model which will carry all the process of the prediction result displayed and the way it can be interacted with. Hence, the Prediction Processing by deep learning has been classified under three factors: Data, Prediction Processing and View.

VI. System Classification

The preliminary search result was 408 in total, out of which 25 journals are finalized and fitted the inclusive criteria for the research paper. For better analysis of the data and for getting closer look on the Prediction Processing research journals from 2018 to 2020 has been considered for gathering some deep knowledge. Only the journal which has levelled to Q1 and Q2 has been brought into consideration. High level journal will provide the research paper the quality in the work for both the researcher as well as users. Along with this, most of the journal were searched for the peer reviewed. Most of the search paper were mainly search for which has Prediction Processing with the deep learning and been focused on Type 2 Diabetics. For example: journal with Deep learning Prediction Processing for Diabetics prediction, Type 2 Diabetics prediction, early stage of diabetic detection and so on. With this, journal which focused on the Electronic Health Record and on feature extraction and classification was also included for research purpose.

After all the research paper were searched, the selection criteria among those papers were implemented, 30 journals were finalized. All the selected paper is illustrated in Table 2, along with used columns and each component are allocated into their subcomponent and sub columns are also illustrated, respectively. The table below present classification followed by evaluation table.

TABLE I. DPD CLASSIFICATION OF SYSTEM COMPONENTS

Reference	**Type of Diabetic**	**Data**		**Prediction Processing**	**Display**	
		Raw	**Feature Processing**		**Data**	**Interaction tools**
[3]	DM	Clinical datasets	N/A	Artificial neural networks Support vector machine ANN Naive Bayes Random Forest Deep learning	M	Computer-Aided Diagnosis (CAD)
[14]	T2D	Pima Indians Diabetes - 8 attributes	Oral glucose tolerance test	Deep neural networks	Softmax layer	Autoencoders
[7]	Chronic disease prediction - T2D	Administrative Health Data (De-identified patients)	BF	Parameter estimation model Logistic regression Binary tree classification		Graph cluster match score
[2]	T2D	Korean cohort data (disease history, parental disease history, and lifestyle habits)	Genetic & Clinical factor	Cox proportional hazards	N/A	SNP point-based score
[4]	T2D	Electronic health record (EHR)	Self-declared Report, earlier diagnosed type 1 diabetes	MMTOP algorithm (Multiple models for Missing values at Time of Prediction)	N/S	Coxph
[1]	T2D	De - identified electronic health records of 9948 patients in total and 1904 diagnosed Diabetics	Feature extraction and selection	Deep learning neural network architecture – Naive Bayes, Decision Tree Random Forest Support Vector Machine Logistic Regression	N/S	QDiabetes - Cox proportional hazards models
[15]	T2D	Pima Indian Diabetes (PID) data set	k-fold cross-validation	DNN	M	WEKA
[9]	Diabetics	MIMIC -II Dataset	FSE	Deep Learning models Feedforward Neural Network Super Learner algorithm Recurrent Neural Network	N/S	Area under the ROC curve (AUROC) and Area under Precision-Recall Curve (AUPRC) is applied for prediction
[13]	DM	KNHANES Dataset - Diagnosed DM	Non invasive factors	Deep neural network (DNN) Model Screening model Non-invasive variables	N/S	Bivariate

VII. VALIDATION AND EVALUATION

We have evaluated both qualitative and quantitative model for the model being analyzed. The evaluation criteria are set to be including the research paper and the domain experts based on criteria prediction model, accuracy, dataset, completeness, and prior knowledge of T2D. The most important part of the evaluation is under the analysis of Numerical data and its evaluation. The researched paper which was reviewed has the numerical analysis for the accuracy of the prediction of the data. Further, we have analyzed the research paper from 2018 to 2020 and some of the qualitative method from the author suggests the importance the deep learning with the electronic data base. [15] suggests the model for the prediction at the best level. In review, many constraints on the prediction technique are outlined and came to a step near to prediction model.

TABLE II. EVALUATION OF THE PREDICTION ALGORITHMS

Author	Applied Technique/Algorithm	Datasets				Study Criteria		Measurement Criteria	Validation/ Evaluation method or dataset	Results	Mathematical formula
		Number	Name	Samples	Types	Input	Output				
[3]	Artificial neural networks Support vector machine ANN Deep learning	1	Pima Datasets	20000	Clinical data	Clinical datasets	Better prediction	Sensitivity Specificity Precision Accuracy F measure	Result analysis – Numerical	The result suggests using tool AI and DL as auxiliary tools to aid during the medical diagnosis	N/S
[14]	Deep neural networks	1	Pima Indian diabetes dataset	768 records and 8 attributes	Diabetic (Positive class) and non-diabetic (Negative Class)	Pima Indians Diabetes - 8 attributes (total of 768 patients record)	Prediction of diabetics has an accuracy of 86.26 % enhances the model	Precision Recall Specificity F1 - score	Numerical analysis and comparison	Deep learning model for T2D has outperform the prediction on various classification model.	Accuracy $=\frac{TP+TN}{TN+FN+TN+FP}$ Precision $=\frac{TP}{TP+FP}$ Recall $=\frac{TP}{TP+FN}$ $F1-\text{Score}=2\cdot\frac{\text{Precision}\cdot\text{Recall}}{Precision+Recall}$ Where True positive (TP), true negative (TN), false positive (FP) and false negative (FN)
[7]	Parameter estimation model Logistic regression Binary tree classification	1	Administrative health data from private healthcare funds	749,0 0 0	longitudinal dataset of 6 years of span coverage	Administrative Health Data (Total of 749,0 0 0 de-identified patients)	Increase of prediction accuracy	Risk Factors	Experiment and Numerical analysis	The accuracy rate between 82% to 87% was predicted using social and network analysis.	$rf_{cluster}=$ $=\frac{N_{match(N\ test)}}{count\ of\ total\ edges\ in\ N\ C}$ Where, ntest is match t other nodes
[2]	Cox proportional hazards	1	Korean cohort data	10030	Clinical and Genetic factors	Korean cohort data (disease history, parental disease history, and lifestyle habits)	The p value of survival NRI of 0.001 shows that combining genetic factors and clinical factors improves the Prediction	Integrated (clinical factors + genetic factors) model	Numerical analysis and evaluation	To predict the occurrence of T2DM Clinical and integrated models can be effective.	ProbT2DM $=1-$ S0 (t) $\cdot$ exp $(\beta_1 x_1+\beta_2 x_2+..\beta_k x_k)$ Where, $x1,\ldots,xk$ are risk factors, $\beta1,\ldots,\beta k$ are the corresponding coefficients of the risk factors
[4]	MMTOP algorithm	1	Action to Control Cardiovascular Risk in Diabetes	10251	Clinical diagnosis of T2DM	Electronic health record (EHR) (total of 10251 patients having T2D)	Better prediction	Comparison model	Comparison and evaluation	Novel strategy which makes prediction even at missing value.	N/S
[1]	Deep learning neural network architecture	1	Practice Fusion	9948	Patients with diabetics	De - identified	Improved accuracy for the prediction from	AUC Sensitivity Specificity	Result analysis and Comparison	The accuracy level for the proposed method has 84.28 % which is	Sensitivity $=\frac{TP}{TP+FP}$ Specificity $=\frac{TN}{TN+FP}$

			Raw Data	1904	and non-diabetic	electronic health records	the ensemble model	Accuracy SMOTE		better score than using SMOTE.	Accuracy = $\frac{TP+TN}{TN+FN+TN+FP}$ Where True positive (TP), true negative (TN), false positive (FP) and false negative (FN)
[15]	DNN	1	Pima Indian diabetes dataset	Total of 768 samples and 8 attributes	UCI machine learning repository database	Pima Indian Diabetes (PID) data set	Prediction Accuracy of 98.35 %	Accuracy Sensitivity Specificity F1 Score	Numerical analysis	A promising system for prediction of diabetics having accuracy of 98.35%, F1 score of 98, and MCC of 97 for five-fold cross-validation.	Sensitivity = $\frac{TP}{TP+FP}$ Specificity = $\frac{TN}{TN+FP}$ Accuracy = $\frac{TP+TN}{TN+FN+TN+FP}$ $F1 = \frac{2TP}{(2TP+FP+FN)}$ Matthews Correlation Coefficient (MCC) $= \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN -}}$ Where True positive (TP), true negative (TN), false positive (FP) and false negative (FN)
[9]	Deep Learning models Feedforward Neural Network Super Learner algorithm RNN	2	MIMIC III Dataset (Intensive Care patients)	53423	Adult patients (above 15 years old) Neonates sample hospital admitted patients	MIMIC-II Dataset	Better performance of Deep Mode than Super Learner models when trained with lot of data.	Area under the ROC curve (AUROC) Area under Precision-Recall Curve (AUPRC)	Experiment and compare and discussion of the results	Deep models have better performance than all other models when raw clinical data are of large numbers.	$X(l+1) = f(l)\ (X(l)\ ) = s(l)\ (W(l)X(l) + b(l))$ Where, where $W(l)$ and $b(l)$ are respectively the weight matrix and bias vector of layer l, and $s(l)$ is a nonlinear activation function.
[13]	Deep neural network (DNN) Model Screening model Non-invasive variables	1	Korean National Health and Nutrition Examination Survey	11456	Excluding diagnosed data and missing data	KNHANES Dataset - Diagnosed DM (total of 11,456)	Area under Curve (AUC) is higher in Deep Learning Model (DLM) compare to s	Logistic regression model	Numerical Analysis	AUC is 80.11 which has better performance than any other model and the proposed DLM can helped the early medication at early stage.	Accuracy = $\frac{TP+TN}{TN+FN+TN+FP}$ Precision = $\frac{TP}{TP+FP}$ Recall = $\frac{TP}{TP+FN}$ Where True positive (TP), true negative (TN), false positive (FP) and false negative (FN)

We have evaluated Deep learning model – Deep Neural Network (DNN) in taxonomy and the dataset for the input. Diabetes and non-diabetic patients, data of different age group and some workers, dividing the cohorts of dataset. We reviewed the article based on electronic dataset with the deep learning, prediction of T2D using deep learning and some different model using paper. All the papers have pre-processing and feature extraction and prediction algorithms. The feature extraction was used for validating and evaluating the paper. Feature extraction plays important role and has different algorithms for the prediction. Moreover, mathematical analysis is concluded for the numerical analysis of the data.

The dataset being used by the research paper are unclassified and de-identified record of the patients and lack the test on the complex system. The data handling technique is crucial for the analyses of the prediction system for the validating and evaluating the research paper.

Second most important evaluation of the framework is the prediction algorithms and its analysis. The algorithms that has most accuracy and the best prediction needs to be compared and detent and compare to other model.

Most of the paper uses the numerical analysis for the prediction of the T2D. Dataset was tested on the different model as of accuracy, sensitivity, and area under the curve. Those papers were evaluated based on raw dataset.

VIII. SYSTEM VERIFICATION

We reviewed the taxonomy based on the prediction technique used on the paper. We have used DNN as the prediction algorithms. [14] suggests the use of DNN and analyzed the data from numerical result based on accuracy and different model including AUC. Further, the research paper from the 2018 to 2020 also suggested to review the electronic dataset and comparing it with different model.[7] paper suggest the CNN model can be effective for the diabetic prediction to the certain group of the people.

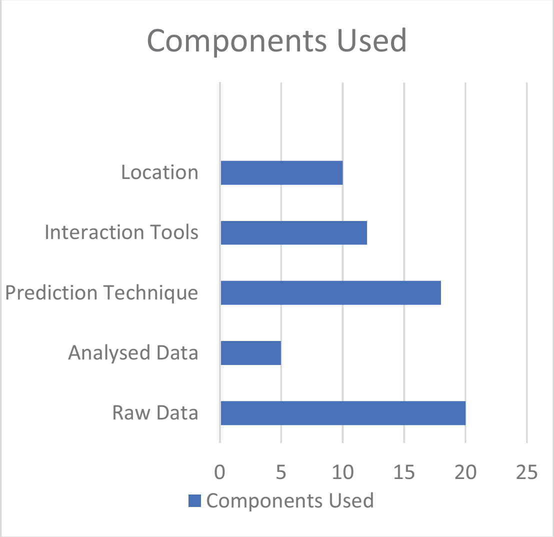

Fig. 3 Graphical representation of the components used in the paper

Based on evaluation and the data analyzed along with the knowledge of the prior data, the taxonomy chosen can be corrected. The overlaps were less in the system and research paper has less occurrence of such corpus. The use of the machine learning in many papers shows lack of research on the deep learning on the T2D and the prediction.

IX. DISCUSSION

While working on literature review, many research papers was collected, and information was processed. After studying we collected all the required information and a taxonomy was drawn which is unique in all the paper reviewed. The required examples and the important contents were included to support the taxonomy DPD (Data, Prediction processing and Display) was developed and discussed. We have also highlighted the components of the taxonomy.

A. *Data*

Almost all the paper has the contents of using health record dataset especially Electronic Health Record (EHR), laboratory report, patients' diabetics history and clinical data. The prior knowledge data can be found only at some of the paper reviewed. The data where it will be displayed was discussed only at some paper. The discussion was based on the basis prediction technique and most was for the accuracy level and the dataset and the content of it. The virtualization of the system works on the accuracy of the prediction technique which is important part of the research paper. The knowledge and experience will not be helpful for the process of the visualization. The dataset cannot be simply proceeded by the system directly as the processing will not be effective enough.

The improper coverage of the patient's dataset and their long history is not covered. Along with this prior knowledge and the derived dataset lacks the contents which can be covered in this taxonomy. Prediction technology will cover the main dataset and analyzed the data and process the steps. The system will be able to proceed the patient's diabetics history, clinical reports, and laboratory data to help the prediction technique.

One of the papers,[1] has covered most of the required dataset and analyzed the labelled data such as patient medical history, laboratory report and clinical report. This research used for the electronic dataset and the dataset of different format for the prediction of the T2D. Many author like[7] and [2] has included laboratory results and clinical dataset.

Some of the research paper focuses only on one set of datasets having one age group will lacks the coverage of the of more feature extraction. The use of limited dataset and the limited knowledge from the derived data lacks the support in the taxonomy of their research paper. Moreover, we opted that the maximum data coverage and feature extraction gives the high accurate rate with the prediction process.

B. Prediction Processing

In some of the paper prediction process has been discussed on the prior dataset processing. The process of the prediction is followed by the feature extraction from the dataset which carries various data related to the patients. [1] has suggested the uses of Deep Neural Model which has emsemble classifier for the dataset analyses for the accuracy or the prediction of T2D. Further [15] uses Recurrent Neural Network for the prediction of the diabetics which training the datasets. Further [14] has more detail method has been analyzed for the better accuracy of the results. Various technique has been discussed and been applied for the prediction process of the T2D. The recurrent neural network discussed on different paper has different results upon ae group [2]. Research [2] and [1] uses deep learning models which uses Deep Conventional Neural Network which uses autoencoder for the extracting and processing the data and for the final result of the prediction and diagnosing the T2D. However, [13] has focused on the deep conventional network in the prediction and data analyses and prediction using electronic dataset. However, more methods have been discussed on the prediction of the T2D under deep learning.

In the paper we have described the taxonomy based on Data type, Prediction algorithm and Display (DPD). The paper mainly focused on reviewing the paper based on this technology and the T2D and has analyzed the latest research paper to provide the solution for the best prediction method for the early prediction of the T2D.

The findings and the literature review suggest that the Deep Neural Network and the combination of the Electronic Dataset or Electronic Health Record (EHR) will be able to predict the T2D with better accuracy rate [11] and [15].

However, the limitation of the technique will be only able to predict the T2D not any other diabetics and the accuracy will not be hundred percentage. The paper is on reviewed based, so it has not been tested on practical dataset nor has predicted any T2D. The future papers are suggested to carry more deep analysis of the Dataset with the most input of previously diagnosis dataset as well as dataset of different age group. The system in future can be suggested to develop to predict different types of diabetics under one algorithm.

The need of deep learning in the prediction of the T2D is important for the prevention of the patient from falling ill. The lack of the dataset being included in the prediction model and the prior knowledge has been covered relevantly in the research area of the research papers. To develop the prediction technique the coverage of vast dataset needs to be included to improve the quality of the paper and predicted result.

C. *Display*

Almost all the of the selected 25 papers view of the prediction of the T2D or the prediction process has focused on the accuracy of the result. [15] and [4] focuses on the display of the result on the accuracy rate while displaying in an interactive tool.

The data view for the prediction processing has uses various tool as per the research requirement.[7] uses cTake, SPSS and graph cluster for the interaction of the data for the prediction of the T2D.

X. RECOMMENDATION

The purposed system covers all the for an effective model for the prediction of the T2D. However, some limitations, issues and improvements are discussed within this section for the improvements of the accuracy for the prediction and early detection. The feature extraction and classifying the data plays an important role in the continuous accurate detection of the T2D. The deep learning model is the most important factors for the technology for the prediction. The Deep Neural Network has the significance to provide better accuracy rate for the prediction of the early stage of the T2D. The increase of dependency to detect the disease on the technique can lead to automation of the system which will lead to the better performance. Also, using the auto filter for the data processing will enhance the work for the prediction and delivering effective result. The implementation issue need to be considered and the complexity of the large dataset need to be considered for the effective functioning of the algorithm.

XI. CONCLUSION

In the paper we have described the taxonomy based on Data type, Prediction algorithm and Display (DPD). The paper mainly focused on reviewing the paper based on this technology and the T2D and has analyzed the latest research paper to provide the solution for the best prediction method for the early prediction of the T2D.

The findings and the literature review suggest that the Deep Neural Network and the combination of the Electronic Dataset or Electronic Health Record (EHR) will be able to predict the T2D with better accuracy rate.

However, the limitation of the technique will be only able to predict the T2D not any other diabetics and the accuracy will not be hundred percentage. The paper is on reviewed based, so it has not been tested on practical dataset nor has predicted any T2D. The future papers are suggested to carry more deep analysis of the Dataset with the most input of previously diagnosis dataset as well as dataset of different age group. The system in future can be suggested to develop to predict different types of diabetics under one algorithm.

REFERENCES

[1] B. P. Nguyen *et al.*, "Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records," *Comput Methods Programs Biomed,* vol. 182, p. 105055, Dec 2019, doi: 10.1016/j.cmpb.2019.105055.

[2] S. H. Kim, E. S. Lee, J. Yoo, and Y. Kim, "Predicting risk of type 2 diabetes mellitus in Korean adults aged 40-69 by integrating clinical and genetic factors," *Prim Care Diabetes,* vol. 13, no. 1, pp. 3-10, Feb 2019, doi: 10.1016/j.pcd.2018.07.004.

[3] G. Oumaima, E. Lotfi, E. Fatiha, and B. Mohammed, "Deep Learning Approach as New Tool for Type 2 Diabetes Detection," presented at the Proceedings of the 2nd International Conference on Networking, Information Systems & Security - NISS19, 2019.

[4] S. Ma, P. J. Schreiner, E. R. Seaquist, M. Ugurbil, R. Zmora, and L. S. Chow, "Multiple predictively equivalent risk models for handling missing data at time of prediction: With an application in severe hypoglycemia risk prediction for type 2 diabetes," *J Biomed Inform,* vol. 103, p. 103379, Mar 2020, doi: 10.1016/j.jbi.2020.103379.

[5] S. Perveen, M. Shahbaz, T. Saba, K. Keshavjee, A. Rehman, and A. Guergachi, "Handling Irregularly Sampled Longitudinal Data and Prognostic Modeling of Diabetes Using Machine Learning Technique," *IEEE Access,* vol. 8, pp. 21875-21885, 2020, doi: 10.1109/access.2020.2968608.

[6] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng, and D. N. Davis, "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values," *IEEE Access,* vol. 7, pp. 102232-102238, 2019, doi: 10.1109/access.2019.2929866.

[7] A. Khan, S. Uddin, and U. Srinivasan, "Chronic disease prediction using administrative data and graph theory: The case of type 2 diabetes," *Expert Systems with Applications,* vol. 136, pp. 230-241, 2019, doi: 10.1016/j.eswa.2019.05.048.

[8] X. L. Xiong, R. X. Zhang, Y. Bi, W. H. Zhou, Y. Yu, and D. L. Zhu, "Machine Learning Models in Type 2 Diabetes Risk Prediction: Results from a Cross-sectional Retrospective Study in Chinese Adults," *Curr Med Sci,* vol. 39, no. 4, pp. 582-588, Aug 2019, doi: 10.1007/s11596-019-2077-4.

[9] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmarking deep learning models on large healthcare datasets," *J Biomed Inform,* vol. 83, pp. 112-134, Jul 2018, doi: 10.1016/j.jbi.2018.04.007.

[10] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, and L. E. Barnes, "Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record," *IEEE Access,* vol. 6, pp. 65333-65346, 2018, doi: 10.1109/access.2018.2875677.

[11] G. Harerimana, J. W. Kim, H. Yoo, and B. Jang, "Deep Learning for Electronic Health Records Analytics," *IEEE Access,* vol. 7, pp. 101245-101259, 2019, doi: 10.1109/access.2019.2928363.

[12] P. Doupe, J. Faghmous, and S. Basu, "Machine Learning for Health Services Researchers," *Value Health,* vol. 22, no. 7, pp. 808-815, Jul 2019, doi: 10.1016/j.jval.2019.02.012.

[13] K. S. Ryu, S. W. Lee, E. Batbaatar, J. W. Lee, K. S. Choi, and H. S. Cha, "A Deep Learning Model for Estimation of Patients with Undiagnosed Diabetes," *Applied Sciences,* vol. 10, no. 1, 2020, doi: 10.3390/app10010421.

[14] K. Kannadasan, D. R. Edla, and V. Kuppili, "Type 2 diabetes data classification using stacked autoencoders in deep neural networks," *Clinical Epidemiology and Global Health,* vol. 7, no. 4, pp. 530-535, 2019, doi: 10.1016/j.cegh.2018.12.004.

[15] S. Islam Ayon and M. Milon Islam, "Diabetes Prediction: A Deep Learning Approach," *International Journal of Information Engineering and Electronic Business,* vol. 11, no. 2, pp. 21-27, 2019, doi: 10.5815/ijieeb.2019.02.03.

Appendix

Abbreviation

N/S – Not Specified	N/A- Not Available	T2D – Type 2 Diabetic	SVM - Support Vector Machines	DNN-Deep Neural Network
MIMIC – Multi parameter Intelligent Monitoring in Intensive Care	SMOTE-Synthetic Minority Oversampling Technique	ANN-Artificial Neural Networks	LSTM – Long Short-Term Memory	D – Derived Data
M-Monitor	FSE – Feature selection and extraction	ED – Early Diagnosis Report	BF- Basic feature	DM – Diabetics Mellitus