Chronic Disease Prediction by Analyzing Clinical Data

Pandey Himanshu Kumar Sachchidanand
Computer Science and Engineering
ITM UNIVERSE, VADODARA
Gujarat Technological University,
Gujarat, India

Patel Jaxitkumar D.
Computer Science and Engineering
ITM UNIVERSE, VADODARA
Gujarat Technological University,
Gujarat, India

Pallavi Hire
Computer Science and Engineering
ITM UNIVERSE, VADODARA
Gujarat Technological University,
Gujarat, India

ABSTRACT
This paper reviews various applications of machine learning and deep learning models and concepts in the diagnosis of chronic diseases. Patients suffering from these diseases need lifelong treatment. At the moment, Predictive models are frequently applied in the diagnosis and forecasting of these chronic diseases. In this study, the most common chronic diseases are been reviewed and analysed. This paper mostly focused on chronic diseases like Diabetes, Heart Disease and Skin Diseases. The outcomes of this journal suggest the diagnosis of chronic diseases, but there is no standard method to determine the best approach in real-time medical/clinical practice since these methods have their own advantages and disadvantages. Among the most commonly used methods, this paper considered Support Vector Machines (SVM), logistic regression (LR), clustering and convolutional neural network. These models are highly applicable in the classification, and diagnosis of chronic diseases and are expected to become more important in medical practiceshortly.

General Terms
Machine Learning and Deep Learning Model for Chronic Diseases prediction.

Keywords
Medical/Clinical Data Analysing, Image Recognising, Convolutional Neural Network (CNN), Disease prediction models, chronic diseases (Diabetes, Heart Diseases, Skin Diseases), Accuracy of models.

1. INTRODUCTION
Artificial intelligence is the technology that uses computer-based knowledge to represent intelligent behaviour with minimal human involvement, whereas machine learning is considered to be a subset of (AI) artificial intelligence [1]. Medical Science is developing very quickly with the help of Artificial intelligence and its applications. In medical science, artificial intelligence (AI) refers to the utilization of automated diagnosis and the treatment of patients who require care for the disease [1]. An increase in the use of artificial intelligence in medicines will play an important role to automate it, reducing the time and efforts of medicinal experts', opening up time to be used in performing different obligations, ones which can't be automated due to lack of time. Machine learning and deep learning is used for

determining complex models and extracting medical knowledge, exposing ideas to the users and specialists [1]. In the present time period, medical science is improving with help of computers. Many types of research in the field of medical sciences are been carried out with the help of artificial intelligence. Which saves time for the researches and at the same time gives an accurate report to the researcher. Machine learning provides a platform to detect many chronic diseases whether it is present or not. To date no such application is present which can predict chronic diseases, only research papers and models are been carried out for the prediction of the disease. The application will give the precautions and the remedies that the user should take if they suffer from the disease. The benefit of the application is that if diseases can be predicted, then early treatment can be given to the patients. With machine learning concepts, it is possible to improve medical data quality, reduce variations in patient rates, and save treatment costs [1]. Early detection and effective treatments are the only options to reduce the death rates, caused by chronic diseases. Models with the highest accuracies will gain large importance in medical diagnosis. Chronic diseases are low-progress in nature, that's why it's important to make an early prediction and provided effective treatment and medication to a patient. Therefore, one must propose a decision model which can help to diagnose these chronic diseases and predict future outcomes of patients. This journal, have proposed a decision model which is a real-time mobile healthcare system for chronic disease prediction based on the current medical/clinical data

2. LITERATURE REVIEW
The healthcare sector continues to propose new concepts and many optimal solutions in the chronic diseases prediction model. A research paper presented by Dr Ravi S. Behra, PhD. Ritesh Jain. of IEEE 14th International conference on bioinformatics and bioengineering, on predictive modelling for wellness and chronic condition [6].
Thomas Bodenheimer, MD, Edward H. Wagner, MD, MPH and Kevin Grumbach, MD.A study suggests that out of 27 cases, 18 have confirmed that the use of proposed chronic disease models has reduced the treatment cost and lead to a decrease in the use of health services [7].

Research carried out by Allen M. Glasgow, Jane Turek, EvBeliveau in "Readmissions of children with diabetes mellitus to a children's hospital" they've, mentioned about the readmission of diabetic children increased due to missed insulin doses [8].

Another point of view presented by, Beata track, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios and John N. Clore in "Impact of HbA1c measurement on Readmission Rates in hospitals: Analysis 70,000 Medical Records of patients" where they used the multivariate regression method. they concluded that thegreatest attention to HbA1c determination may improve outcomes and also reduces the treatment cost. The use of a proposed model to develop interventions in hospitals has provided better patient outcomes, it has decreased length of stay, and lessened rates of patients' re-admission in a 30-day [9].

In one study by Jill Koproski, CDE, ZoraydaPretto, MD and Leonid Poretsky, MD, they conducted a random study on the effects of intervention in hospitalized diabetic and they found that this resulted in decreased length of stay in hospital as well as overall improvement in glycaemic levels [10].

Research paper presented by Yang Guo, GuohuaBai, Yan Hu School of computing Blekinge Institute of Technology Karlskrona, Sweden, the revelation of information from clinical records is critical keeping in mind the end goal to make a successful therapeutic determination [11].

3. METHODOLOGIES

The used classification models arebased on the following algorithms: Support Vector Machines (SVM), logistic regression (LR), clustering and Deep convolutional network, Convolutional Neural Network (CNN) [2][3][5].

3.1 MachineLearning Model Development

Methodology for Two of our final models is based on machine learning algorithms [1]. The final model involves taking the logistic regression (LR), clustering and support vector machine (SVM) as our model to find our prediction results [3][2].

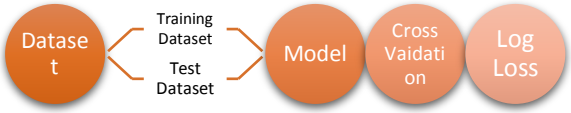

Fig.1: Model Architecture

In Fig.1, model architecture is shown that how the research reached to end of this model or the procedures we followed to get our results.

The reason behind the use of logistic regression and support vector machine is that they both are binary classification model. In logistic regression (LR), data is classified into two different classes 0 and 1 [3]. In logistic regression (LR), the output is always a discrete value which is (1) or (0) (yes or no) [3].

Support vector machine (SVM) is a model which is used for both classification and regression problems [2]. In comparison with logistic regression (LR), the support vector machine (SVM) separates the classes which reduce the risk of error on the data [3][2].

Clustering is an unsupervised learning method, it identifies and group similar data, it is only used for the larger datasets. Usually, it is used to classify data into structures that make it easier to understood and manipulate.

3.1.1 Pre-processing

Pre-processing was performed on the dataset; it is only used for data cleaning purposes. Pre-processing also includes normalization, feature extraction etc. It identifies and removes duplicate rows from the datasets, also the outliers (i.e., values that are outside the specified range of - 3δ and +3δ). Data cleaning was already done, all the missing values and missing parameters were either removed or replaced so pre-processing was not required.

3.1.1.1 Feature Reduction

It reduces the number of features i.e., the number of variables in the dataset present were also been reduced. A single value decomposition method was used to construct enriched features in the data. This simplifies features from 13 to 4 features (heart disease) and from 8 to 3 (Diabetes).

3.1.1.2 Data Transformation

Normalization was performed on the valued features that max heart rate, age, etc of both training and test sets. It changes the values in a dataset to a common scale. The following features were normalized age, old peak, insulin, BMI, etc. to range between 0 and

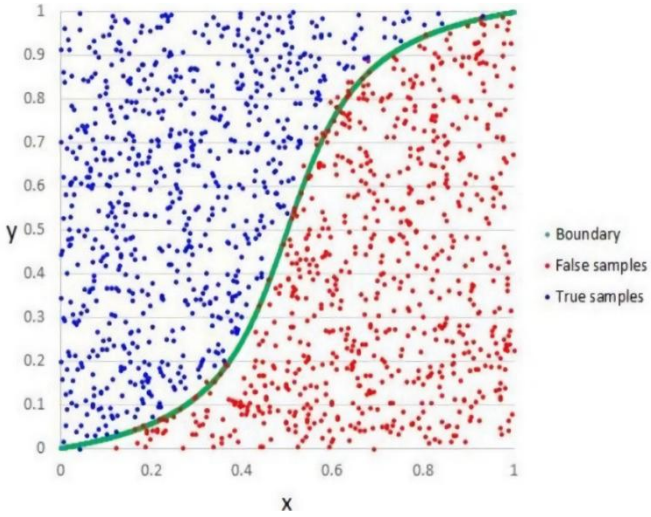

Fig.2: Linear Regression (LR)

In Fig.2,Logistic regression (LR) divides the data into two different classes is 1 and 0 (True or False) samples separated by a boundary and the same goes for our problem [3]. While support vector machine (SVM) identifies the best margin i.e., the distance between the line and the support vectors [2]. this separates the classes and also reduces the risk of error on data. Support vector machine (SVM) is based on geometrical properties, while Logistics regression (LR) is based on statical approaches of the data [2][3].

3.2 DEEPLEARNING MODEL DEVELOPMENT

The dataset used is consists of 7000+ RGB images of the infected skin areas, categorized into 20 different skin diseases. A complete dataset was employed to train the system model. In the deep learning model, the dataset is spilt into a training set and testing set [1] (which is the same as in the machine learning model).

3.2.1. PRE-PROCESSING

Pre-processing was performed on the dataset; it is only used for data cleaning purposes. At whatever point the information is assembled from the sources it is collected in an informal way which isn't achievable for the training or testing purpose. For accomplishing better accuracy from the applied model in deep learning, the organization of the information must be in proper form [1].

3.2.1.1 DEVELOPMENT

For training model in deep learning, passing an algorithm with the training data [1]. CNN finds patterns in the training data, that the input parameters correspond to the target [5].The output of the training process is a machine learning model which can be used for prediction. In deep learning the estimation of the objective capacity that has been prepared to utilize preparing information, sums up to new information [1][5].

3.2.1.2 USE OF CONVOLUTIONAL NEURAL NETWORK (CNN)

A convolutional neural network (CNN) is a deep learning algorithm used for analysing visual image that we can accept an image as an input and attaches some significance to each point in images which makes it easy to differentiate [5].

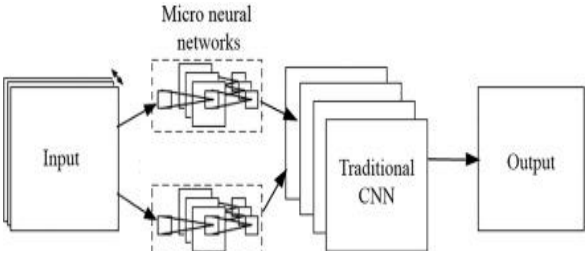

Fig. 3: Convolutional Neural Network

The convolutional neural network (CNN) model uses a feed-forward network for image recognition [5]. Convolutional neural network (CNN) uses Multilayer perceptron (MLP), for convolutional processes in which each neuron in a layer is associated with another neuron in the other layers. Convolutional neural network (CNN) techniques are well known for their better performance and accuracy in image recognition. A small amount of pre-processing is needed for convolutional neural networks (CNN). Various filters are used in a convolutional layer; each filter makes a two-dimensional activation map by moving across the input data. A convolutional neural network (CNN) is highly used for image recognition.

The convolutional neural network (CNN) model consists of the input layer, which is composed of artificial neurons and the output layer also having multiple hidden layers [5]. Hidden layers of convolutional neural network (CNN)are convolutional layer, pooling layer, fully connected layer, receptive field and weights [5]. Feature extraction and classification are two components of convolutional neural networks (CNN) [5].

Feature extraction is performed by convolution and pooling layer. fully-connected layers then act as a classifier on top of these extracted features and assign a probability to provide the final output.

A better outcome achieved to predict and prevent skin diseases using the concept of deep learning (CNN) [5].

4. PROPOSED METHOD

In this study,three complex models have been implemented together in one application, which patients can use for their benefits. This will help them to predict or know the status of their diseases. This will save millions of lives and also cut down the treatment cost for these diseases.
First, the user has to pass on some medical parameters, these data will be then passed to the models, which will give the output(predict) of the disease. This applies to all models in the application.

The uniqueness of application will be, only for diseases like Diabetes and heart diseases; users have to pass the medical parameters which will be in text form. But for skin disease prediction, they just have to upload the infected skin area, this will give them an accurate prediction and also some home treatment(first-aid) and precautions one should take for those specific diseases.

This application not only saves time for the patients but also the treatment cost, the complexity of the diseases and most important will save millions of lives.

This study is unique in itself because no one at present has neither implemented nor used three complex models at the same time and that too in one application.

5. SYSTEM ARCHITECTURE

Datasets from different sources are been collected (i.e., for diabetes and heart diseases we are taking. CSV data collection while for skin disease we're collecting image files and using that as a dataset for training and testing purpose of the model). Dataset for diabetes has 8 parameters like age, glucose level, insulin etc. while in the heart diseases dataset total number of parameters present were 13. Unique is our skin diseases dataset, it is a collection of skin infected areas (particular diseases). The total number of datasets used here are approx. 5000+.
Steps for pre-processing are as follows: (i) Very the first step is, cloning of dataset for using it in our model. (ii) Submission format these are because for heart diseases the dataset is divided into two columns with patient_id and disease present. (iii) Then, cleaning data is the process which is the most important, it’s a check that whether there's any missing value or false values present in the dataset, ensuring it before using it.

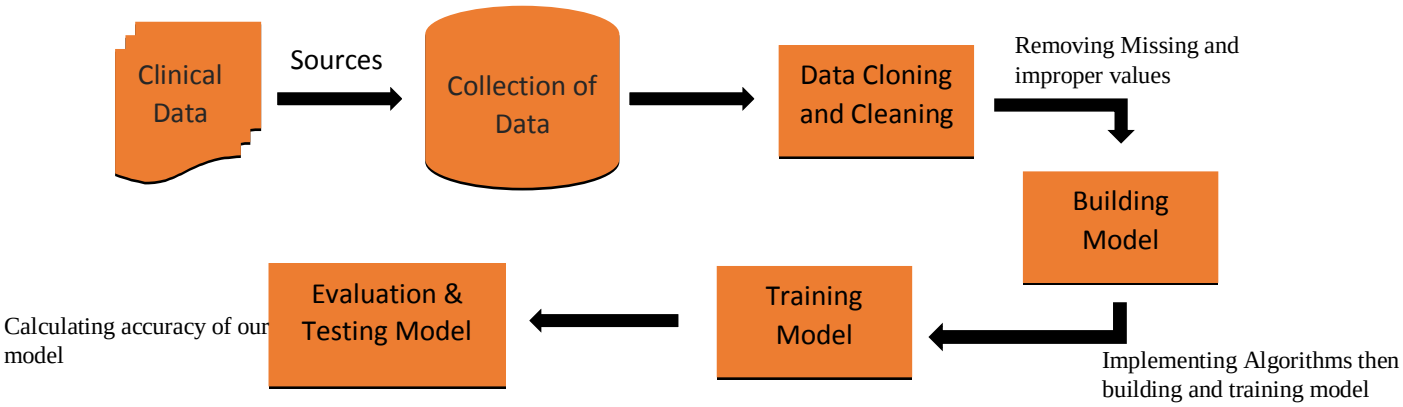

Fig.5: System Architecture

5.1 Build & Train Model

After completing the whole task like deciding feature columns, cleaning of the datasets and cloning, now the use ofthe datasets come, here they are, the model is been build and then trained. These include, features with their outcomes and the final output will be the summary, outcomes and the prediction result.

5.2 Evaluation& Testing Model

The evaluation and testing model is an important part of the whole process. Here, this evaluates the model and then it can get to know about the accuracy of the proposed model.

6. RESULTS

The dataset we collected from the Kaggle site [12].
The training dataset consists of 13 essential features (heart disease) and 8 essential features (Diabetes) [12]. The metric used for the evaluation id log loss as shown in equations.:

$$J(\emptyset) = C\left[\sum_{i=1}^{m} y^i Cost_1\left(\emptyset^T\left(x^i\right)\right) + \left(1-y^i\right)Cost_0\left(\emptyset^T\left(x^{(i)}\right)\right)\right] + \frac{1}{2}\sum_{j=1}^{n}\emptyset_j^2$$

m = number of samples, n = number of features

$$Log\ Loss = \frac{1}{n}\sum_{i=1}^{n}[y_i \log(y_i\hat{}) + (1-y_i)\log(1-y_i\hat{})]$$

where $\boldsymbol{y_i}$ *represents* ***y=1***

By log, loss implies a good accuracy for the
model. The method involved was passing the data through the model and repeating this method for different models

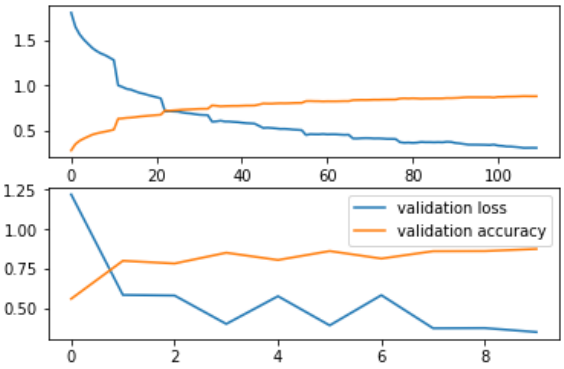

Fig.6: Validation Loss and Accuracy.

In Fig.6, the main goal was to reduce log loss which implies a proper accuracy to model. The method involved was passing the data (medical parameters) through the model, noting the log loss and repeat this method for different models.

```
evaluator = BinaryClassificationEvaluator
evaluator.evaluate(model.transform(test))
```

```
0.838156931080637
```

Fig.7: Accuracy of a model (Diabetes).

	precision	recall	f1-score	support
0.0	0.80	0.74	0.77	27
1.0	0.79	0.84	0.81	31

Fig.8: Accuracy of a model (Heart diseases)

class	Cellulitis Impetigo and other Bacterial Infections
score	76.2

Fig.9: Accuracy of a model (Skin Disease)

In Fig. [7][8][9], the accuracy score obtained from both models is shown. Diabetes model accuracy was approx. 83% and for the Heart Disease model that was approx. 79%. While for Skin diseases we get 76.2% accuracy.

7. CONCLUSION

Improving the Health care sector and helping many peoples who are suffering from these chronic diseases. Based on this journal, an application is been developed that can predict these diseases easily and fast. Many people around the world are been suffering from most chronic diseases which increases the health the complication and also has the high treatment cost, in the application; only selected/major or can say the most common chronic diseases for thepredictions like heart, diabetes and skin diseases which can be diagnosed as early as possible and will reduce the risk of life and also lowers the treatment cost for these diseases.

In the future, Artificial Intelligence (AI) concepts like Machine Learning (ML) and Deep learning may play a critical role in the interpretation of chronic diseases [1].

However, many researchers are progressively been attracted towards proposed predictive model techniques in the advancement of healthcare. As involvement and advancement in healthcare is been done and is expanding access to new electronic clinical data opens up a new door to decision support and productivity improvement.

These models are designed to emphasize the responsibility of patient care quality and cut down on treatment and save millions of lives.

8. REFERENCES

[1] Artificial Intelligence (AI) “https://ai.google/”

[2] Support Vector Machine (SVM) “https://scikit-learn.org/stable/modules/svm.html”

[3] Linear Regression (LR) “https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html”

[4] Clustering Machine Learning “https://developers.google.com/machine-learning/clustering”

[5] Convolutional Neural Network (CNN) “https://www.tensorflow.org/tutorials/images/cnn”

[6] Dr Ravi S. Behra, PhD. Ritesh Jain. “IEEE 14th International conference on bioinformatics and bioengineering, on predictive modelling for wellness and chronic condition”.

[7] A Study by Thomas Bodenheimer, MD, Edward H. Wagner, MD, MPH and Kevin Grumbach, MD.A on Aug.19,2010 JAMA.

[8] Allen M. Glasgow, Jill Weissberg-Benchell, W. Douglas Tynan, Sandra F. Epstein, Chris Driscoll, Jane Turek, EvBeliveau. "Readmissions of children with diabetes mellitus to a children's hospital"

[9] Beata track, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios and John N. Clore. "Impact of HbA1c measurement"

[10] Jill Koproski, CDE, ZoraydaPretto, MD and Leonid Poretsky. “The effects of intervention n hospitalized diabetic”.

[11] Yang Guo, GuohuaBai, Yan Hu School of Computing Blekinge Institute of Technology Karlskrona, Sweden.

[12] Kaggle for datasets“https://www.kaggle.com”