IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 1, 2021, accepted July 4, 2021, date of publication July 9, 2021, date of current version July 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3096139

An Optimised Multivariable Regression Model for Predictive Analysis of Diabetic Disease Progression

V. K. DALIYA[1], T. K. RAMESH[1], (Member, IEEE), AND SEOK-BUM KO[2], (Senior Member, IEEE)

[1]Department of Electronics and Communication Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru 560035, India
[2]Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, SK S7N 5B5, Canada

Corresponding author: V. K. Daliya (daliyavk1@gmail.com)

ABSTRACT With the advent of smart systems and smart IoT network all over the world leading to enormous amount of data generation; the right analysis and decision making based on the relevant data plays a crucial role. Various industries such as transportation, retail, healthcare etc. rely on analysis using this huge volumes of data for intelligent decision making. In smart healthcare system, accurate analysis of patients' data and prediction of diseases and medicine is important. To a great extent, fatalities can be avoided by timely recommendation of healthcare measures and immediate alert on emergency conditions. The use of machine learning algorithm for precise predictive analysis of data can be very promising in the field of healthcare. In this paper, optimised Multivariable Linear regression method is used to predict the diabetic disease progression of 442 patients based on various parameters such as age, gender, Body Mass Index and 6 different blood serum measurements. Here optimisation is performed using feature reduction and logarithmic transformation. The predicted output is found to be closely associated with actual output data with a Root Mean Square Error of 1.5 units; which indicates higher accuracy in comparison with the non-optimised model with the error of 54 units. There has also been a comparison with the results obtained from other state of the art regression methods, which proves that the proposed model exhibits maximum accuracy. This method can be used to provide promising medical advice to the patients on how to reduce the diabetic disease progression over a year by controlling various health parameters.

INDEX TERMS Data analysis, diabetic management, linear regression, machine learning, multivariable regression.

I. INTRODUCTION

The advancement in technology has taken the world towards devices which are connected to each other and with the internet. Because of the enormous number of connected devices in use, large volume of data is getting generated. It poses data analysis to be a challenging task. At the same time, the modern day technology of analysing the data and predicting various outcomes can benefit the mankind and impact the life of people positively across industries. Prediction of certain diseases from symptoms, analysis of customer behaviour to estimate how a certain product will generate revenue in market, predicting the future traffic conditions based on various parameters in a particular area of a city etc. can contribute towards improving the quality of life [1]. Hence, the concept of data analysis plays a major role in today's connected world. One of the important steps of data analysis is data mining. It is about application of specific algorithm for extracting patterns from the data. The combination of data mining and computer science has given rise to data science, which is one of the most prominent technologies of the present world. The steps involved in data analysis or the layers of data science can be depicted with the help of Fig. 1 [2].

Referring to Fig. 1, in the first stage known as data collection stage; the data is obtained from sensors, instruments, online sources or it is generated by the user. The data can be primary data obtained through direct observation or it can be secondary dataset, tested or used by others for various analysis. If the collected data is lacking in some manner, data augmentation can be done by adding extra relevant data

The associate editor coordinating the review of this manuscript and approving it for publication was Yin Zhang.

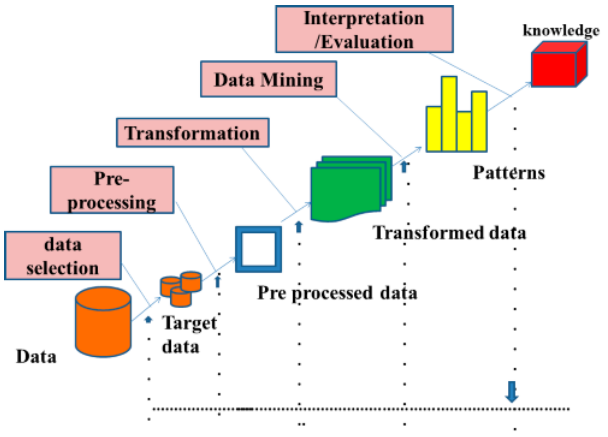


FIGURE 1. Steps of data analysis.

parameters [3]. If the type of data is not available, it can also be generated. The second stage is about data preprocessing; which involves cleaning of data, anomaly detection and formatting as per the requirements. In this stage missing data is identified, irrelevant data are removed and outliers are detected which can interfere with the efficiency of the model. Various statistical parameters are also checked to identify the behaviour of the data. Normalised data is often preferred as it gives good model performance. As per stage 3, we may transform the data using any of the transformation techniques such as identity transform, square root transform, wavelet transform, logarithmic transform etc., if the dataset does not fit properly after using the machine learning algorithm [4]. After proper transformation, data can be labeled. If the dataset is large, data aggregation, can be carried out. It is the process of selecting a part of the dataset, that may predict proper behavior of the entire dataset [5]. In fourth stage called data mining, learning and optimisation is performed using machine learning methods and outcome is predicted. This is followed by interpretation and evaluation of the results in the fifth stage. Finally, the obtained knowledge is presented in stage 6. The data analysis and prediction if done meticulously can be very useful in telemedicine and remote monitoring of patients [6]. The use of machine learning in predictive analysis, data clustering and reinforcement learning is well known. Machine learning methods outperform the traditional techniques with respect to its efficiency in knowledge discovery and prediction [7].

There are different types of learning techniques such as supervised and unsupervised learning in machine learning. In supervised learning, machine is trained using a set of data, whereas in unsupervised learning, there is no training dataset given to the machine to learn from. There are methods based on continuous data as well as discrete classes too. If the predictive analysis is conducted on continuous data and prediction is delivered as a continuous value, the method is named as regression. If the prediction is used to classify the output in any of the discrete classes, the method is termed as classification [8]. The technique used in this paper is regression.

Machine learning techniques can be used individually or in combination of different models known as Ensemble method. The individual models which are used in common include K-Nearest Neighbours (KNN), Linear Regression(LR), Support Vector Machines(SVM), Classification and Regression Tree(CART), Artificial Neural Network(ANN) based models etc. Various methods such as bagging, boosting and stacking are used to combine the predictions of individual models to produce the best results in ensemble technique [9]. In bagging, similar type of individual models are considered, which learn independently from each other in parallel and the results are combined using averaging or other deterministic strategies. In boosting technique, the homogeneous individual learners learn sequentially, so that each model becomes better learners from the previous learners and predictions are combined at the end. Stacking ensemble considers learners which are different in characteristics. They learn in parallel and the predictions are combined using a meta-model.

Machine learning in healthcare sector generally involves prediction of any disease from the symptoms, classification of images based on certain disease parameters etc. Here, diabetic disease progression, one of the major health issues faced by people all over the globe is considered for analysis. Diabetes Mellitus is a chronic metabolic disorder that causes abnormal regulation of blood glucose, either due to lack of insulin production in the body (Type I) or due to resistance of the body to insulin (Type II). Insulin is a hormone that helps in regulating the blood glucose levels. The number of adults expected to suffer from this illness over this disease is expected to be 640 million by the end of 2040 [10]. The complexity in managing the disease and giving proper predictive insights into the control of the disease has given rise to the use of machine learning techniques.

A. RELATED WORK

Since Diabetes Mellitus is a disease prevalent all over the world, there has been a lot of work related to blood glucose prediction and analysis in the literature. A number of machine learning techniques have been used and the performance matrices have been discussed in the papers [10]. Among the various articles in journals and conferences considered in recent years, research work with different machine learning algorithms were examined. From these papers after eliminating algorithms which are similar and repetitive, some of the related articles were chosen for analysis. Among this, research work with single machine learning algorithm and articles with hybrid method as a combination of various algorithms; were identified. Based on the preprocessing techniques, validation approach and class of machine learning algorithm used, these data are categorized into rows and columns. The summary of the literature review is given in Tables 1 and 2. Table 1 narrates the summary of single machine learning algorithm used, while Table 2 lists the features of hybrid methods used. The various preprocessing

TABLE 1. Summary of individual methods used for diabetic prediction.

Input Type	Input Pre-processing	Machine Learning algorithm used	Validation approach	Prediction Horizon	Performance Metrics
Blood Glucose	Smoothing (LPF-order 11)	Recurrent Neural Network (RNN) [11]	Random Subsampling (10% test, 90%train)	15,30,45,60 minutes	RMSE= 0.14, 0.84, 1.32
Blood Glucose	Smoothing,noise and time lag reduction	Feed forward Neural Network (FNN) [12]	Random sub sampling (10% test ,90% train)	15,30,45,60 minutes	RMSE=0.15, 0.42, 0.8
BG,Insulin,Diet	Normalisation	Online and adaptive RNN [13]	Hold out (50% train and 50% test)	30 and 45 minutes	RMSE(mean)= -14 ± 4.1
17 medical exam items	Random forest, SBS algorithm	Support Vector Machines and Random Forest [14]	Random sub sampling (60% training, 40% testing)	30 minutes	RMSE= 19 ± 0.3
BG	NA	Support Vector Regression based on Differential Evolution [15]	Hold out -70% training,30% testing	30 and 60 minutes	RMSE= 10.78, 12.95
BG, Insulin, Diet, Heatflux, Skin temp, METs	NA	Gaussian Processes,Bayesian framework [16]	Hold out- 6 days training data, 3 days validation	25 minutes,1 hour,4 hours	NA
BG ,Diet, Insulin	NA	Genetic programming-Grammatical Evolution [17]	Hold out: 100% data taken at different time	NA	PAE = -14.12 with std deviation 2.11
Real world categorical medical data	NA	K-Nearest Neighbours [18]	10-fold cross validation	NA	MAE= 0.3

methods and other performance parameters of the same is shown in the tables.

Referring to the tables 1,2 and the corresponding papers cited, briefing of the literature is presented in this section. In article [11], paper [12] and [13], deep learning using Neural Network is applied as the machine learning algorithm. Neural networks work similar to the neuron in human body which takes the data in and passes it through much iteration and adjusts the weights. It will have input layer, output layer and hidden layers. In [11], random sub sampling is used as a measure of splitting the data. 90% of the data is used for training the model and 10% of the data is consumed for testing. Blood glucose measurement (BG) of nine patients taken from Continuous Glucose Machine (CGM) is used as the input parameter and the prediction is taken after a time period of 15, 30, 45 and 60 minutes. Low pass filter is used for preprocessing the data. Root Mean Squared Error (RMSE) is chosen as the performance metric and it has been observed that the error increases with prediction horizon. In [12] also, blood glucose of nine real patients from CGM is taken as the input parameter and RMSE increased with prediction horizon. Article [13] considered blood glucose, insulin and diet of a single patient as the input parameters. In this case, input data is normalized and hold out method is used for splitting the data into 50% test and 50% train set. Here adaptive Recurrent Neural Network (RNN) is used. In [14], medical exam data of 17 patients were analysed for blood glucose prediction with the help of Support Vector Machine (SVM) algorithm and Random Forest (RF) approach is chosen to optimize the output. Random sub sampling of data into 60% and 40% is employed for training and testing data respectively. Similarly paper [15], article [16] and [17] have applied Support Vector Regression (SVR) with Differential Evolution (DE) optimisation, Gaussian Process, Genetic Programming and Grammatical Evaluation algorithms respectively. A few patients were observed in the literature considered with input parameters either as blood glucose alone or BG with insulin and diet. In all the above cases, the RMSE value increase as prediction horizon increases from 15 minutes to 60 minutes. Article [18] analyses the performance of real world categorical data by using optimised KNN (opt-KNN)model. The value of K is optimised in accordance to least RMSE. It uses 10-fold cross validation method to obtain Mean Absolute Error(MAE) of 0.3 for $K = 3$.

Apart from these methods, there are hybrid methods used by researchers, where they have taken combination of different algorithms to produce best results. Reference [19] discusses a hybrid method, where Feed forward Neural Network and Linear Prediction algorithm is used in combination with Physiological method. Scaling of the input data was done and K-fold cross validation was performed to get equal chance for each fold of data to appear as training and testing data. In this method, average of all the validation is taken to assess the performance in general. RMSE of 9.4 ± 1.5 is obtained as per this task. This analysis was conducted on 35 patients' data with BG, insulin and diet as input parameters. In [20], Jump Neural Network with physiological model is chosen for analysis. In this case, the data is normalized, scaled and Bayesian smoothing was carried out for preprocessing to obtain best performing model. In another experiment [21], 10 patients' real data was considered with BG, change in blood glucose levels and physical activities as input parameters. A hybrid method consisting of many algorithms such as Feed forward Neural Network (FNN), Self Organizing Map (SOM), Neuro fuzzy network with Wavelet as activation function (WFNN) and Linear

TABLE 2. Summary of hybrid methods used for diabetic prediction.

Input Type	Input Pre-processing	Machine Learning algorithm used	Validation approach	Prediction Horizon	Performance Metrics
Blood Glucose, Insulin, Diet	Scaling, adding noise on CGM	Hybrid (Feed forward Neural Network + Linear Prediction algorithm) with Physiological model [19]	K fold cross validation	30 minutes	RMSE= 9.4 ± 1.5
Blood Glucose, Derivative of BG, Diet	Normalization, scaling, Bayesian Smoothing	Hybrid (Jump NN with physiological model) [20]	K-fold cross validation	30 minutes	RMSE= -16.6 ± 3.1
Blood Glucose, change in BG, physical activities	Normalisation, quantizing input space	Hybrid (FNN +SOM+WFNN+LRM) [21]	10 fold cross validation	30, 60 and 120 minutes	RMSE(mean)= $13.31 \pm 4.47, 22.66 \pm 6.86, 37.62 \pm 11.79$
BG	NA	Ensemble approach,Hybrid-fusion(AR, ELM, SVR-kernel function-Gaussian) [22]	Hold out (60% training, 40% testing and validation)	15,30, 45 minutes	RMSE (30 min.) $= 19 \pm 0.3$
BG, Insulin, Diet and Exercise	Fuzzy approximation of food and exercise	Hybrid (Compartment model&FNN, fuzzy logic and expert system) [23]	Hold out	75 minutes	MAD= -15.9%
BG, Insulin, Diet, Exercise, Sleep, Hypoglycemic symptoms	Pooled panel data, regression, clustering	Support Vector Machines, Decision tree, Random Forest [24]	NA	NA	RMSE= $(SVM - 68.76, DT - 41, RF - 39.73)$
BG, Diet, Insulin, Physical activity	NA	Hybrid (SVM, Linear kernel and Compartmental model) [25]	K fold cross validation	15,30,60 and 120 minutes	RMSE= $9.3, 15.6, 24.06, 31.24$
5 diabetic children's medical data	Sliding window	Extreme Gradient Boost(XGBoost) [26]	NA	30 and 60 minutes	RMSE $= 23.219, 35.8$
US Hospital data and PIMA Indian dataset	Feature selection with Genetic Algorithm	Stacking Ensemble [27]	5,10,15 fold cross validation	5 years	Acuuracy $= 99\%$
3 sets of online medical data	Normalization	Bagging, Boosting, Random subspace, DECORATE, Rotation Forest [28]	Repeated K-fold cross validation	NA	Area Under the Curve(AUC) $= 90\%$
Medical competition data from china	feature selection using Lasso Regression	Adaboost, Gradient Boost Decision Tree(GBDT), RF, Bagging [29]	10 fold cross validation	NA	RMSE $= .005, .007, .445, .241$

Regression Method (LRM) were used to perform predictive analysis. Normalization of input data and its quantization were executed as preprocessing methods. Among the statistical parameters considered, RMSE, Correlation Coefficient (CC), Mean Absolute Relative Difference (MARD) and Continuous Glucose Error Grid Analysis (CG-EGA) were chosen for performance assessment. In another approach [22], Support Vector Regression (SVR) was combined with Extreme Learning Machine (ELM) for an ensemble method. 10 real patients' CGM data was taken. Blood glucose was the only parameter under consideration as the input parameter. 60% of the data was used for training and the rest for testing and validation. RMSE and Clark Grid Analysis were considered for error analysis. Reference [23] used combination of Compartment model with Neural Network and [24] discusses hybrid of SVM, Decision Tree and RF; while in [25], SVM is combined with Linear kernel and Compartmental model. XGBoost algorithm is used for predictive analysis of 5 diabetic children's medical data in [26]. It predicts future Blood Glucose levels ahead of time, with prediction horizon of 30 and 60 minutes. Here, the RMSE value increases from 23 to 35.8, as the prediction horizon increases. In [27], a Stacking Ensemble is proposed in combination with Genetic Algorithm for Pima Indian Diabetic Dataset and US hospital data. The article claims to predict onset of diabetes in a time span of 5 years with 99% accuracy. Paper [28], analyses the performance of various tree based ensemble models such as bagging, boosting, random subspace, DECORATE and Rotation Forest, with different tree based base models. The ensemble has obtained 90% classification accuracy for bagging and boosting methods, with Logistic Model Tree (LMT) as the base model, taking AUC of Receiver Operating Characteristics (ROC) as the performance metric. An empirical study using various tree based ensemble models have been proposed in [29], which discusses performance of Adaboost, GBDT, RF and Bagging techniques, resulting in least RMSE of 0.005 with Adaboost technique compared to other ensembles.

From these analysis, it is clear that many machine learning algorithms and its combinations are used for predictive

analysis of blood glucose, which predict if a person is diabetic or not. Most of these papers considered RMSE as the major performance metric.

From the reviewed articles, it can be understood that the input parameters considered are generally blood glucose levels, diet and insulin. Only a few of them considered exercise and other activities of the person as inputs. Many of the researchers have extracted data from CGM machines and the number of data points utilized for experiment is less. In general, 10 to 20 patient's data were taken, while very few papers considered data points which are in 100's or in 1000's range. In the research works which considered prediction horizon varying from 10 minutes to 60 minutes after recording the input parameters, the error rate increased as the prediction horizon increased. A person's conditions such as infection or other diseases or medical conditions such as hypertension or blood serum measurements were taken into account by a few articles only. But it is very much known that many such factors also affect the diabetic disease manifestation in a person. The long term progression of the disease is not monitored in these papers.

B. MAJOR CONTRIBUTIONS

- In this paper, diabetic disease progression of 442 patients over one year of baseline, after taking the input parameters such as Age, Sex, BMI and six blood serum measurements has been predicted.
- Based on various statistical analysis, optimisation is performed on one of the supervised machine learning methods called Multivariable Linear Regression.
- By exploring the actual, predicted values, error parameters and other statistical performance measures, the model performance is found to be promising compared to the non-optimised model.
- The RMSE of the optimised model is 1.5 while that for the non-optimised model is 54. It shows the efficiency of the proposed optimised model.
- The dataset is also tested with other popular machine learning algorithms as well as its combinations. In comparison with these methods, the application of proposed model has resulted in least error.
- The papers in literature have considered predicting the immediate diabetic information, while the proposed method predicts the progression of the disease over a period of one year.
- The proposed prediction can be useful for the doctors and medical professionals in providing proper medical advice to the patients in controlling their disease.
- The optimised model can be insightful in predicting any other disease or health parameter and can function as an intelligent decision making element of telemedicine and remote monitoring of patients.

The rest of the paper is organised as follows. Section II discusses system architecture. Narration of theoretical aspect of the algorithm used is given in section III and section IV deals with proposed method in detail. Section V discusses the results and section VI concludes the paper followed by references.

II. SYSTEM ARCHITECTURE

Considering the shortcomings identified from the literature review, in this paper a dataset of 442 patients with input parameters AGE, SEX, BMI, Average Blood Pressure (ABP) and six blood serum measurements such as Low Density Lipoproteins (LDL), High Density Lipoprotein (HDL), Total Cholesterol (TC), Serum concentration of Lamotrigine (LTG), Thyroid stimulating hormone (TCH) and blood glucose level (Glu) are considered as input parameters and a Quantitative Measure of Diabetic Disease Progression (QMDDP), one year after the baseline is taken as the target variable. The parameters AGE, SEX, BMI and ABP represents the statistics of the patients' body while LDL, HDL and TC stands for the individual cholesterol levels and total cholesterol respectively. LTG indicates the serum concentration of diabetic neuropathy medicine consumed by the patients. TCH is a measure of thyroid stimulating hormone which can impact the diabetic condition, while the parameter Glu signifies the blood glucose level, which has a direct impact on the disease. QMDDP, which depicts the disease progression of the patients after one year, gives the quantitative values, that can be assessed to analyse the progression to be high risk or not. This dataset is taken from the following link available from scikitlearn; https://www4.stat.ncsu.edu/ boos/var.select/diabetes.html.

The diagrammatic representation of the system architecture of the proposed method is depicted in Fig. 2.

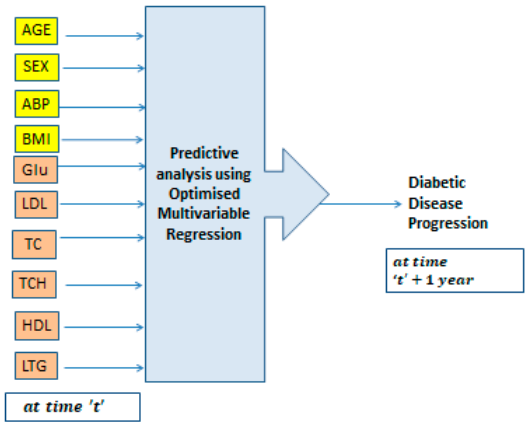


FIGURE 2. System architecture of the proposed method describing the ten input parameters, optimised method used and the predicted outcome.

Referring to Fig. 2, we can see that the system architecture consists of 10 input parameters which represent health parameters of the patients with blood serum measurements taken at time 't' and the predicted value of quantitative measurement of the disease progression as the output, at time 't'+ 1 year. This paper discusses the performance of the

system by measuring the error between actual outcome and the predicted outcome.

In the proposed method, prediction is carried out on how the diabetic disease progresses over one year, after taking the blood measurements and other individual parameters. So, unlike the reviewed articles, instead of predicting if a person is diabetic or not, based on blood glucose level, insulin and diet; here the prediction is carried out on the diabetic disease progression over a long period of time of 1 year. Here, optimisation process by means of feature reduction and logarithmic transformation is applied on Multivariable Linear Regression method. Hence, this method can give insight into the kind of precaution and healthy habits to be considered to reduce the progression of the disease over a long period of time. Even though exercise or diet is not directly considered in this model, it is assumed that the blood serum measurement and body parameters considered is a reflection of the person's daily lifestyle and genetics. Moreover, by assessing the predictive model, corrective action for the lifestyle can very well be advised to the concerned patients. Linear regression being a simple and straight forward technique, this analysis stands out in its own way.

The theoretical aspect of the main algorithm used is narrated in the following section.

III. MULTIVARIABLE REGRESSION

To perform the data analysis, a lot of statistical parameters are considered and optimisation techniques are used. However, the main data mining algorithm used is Multivariable Linear Regression.

Regression is a technique used to predict the value of one or more continuous target variables t, given the value of a D- dimensional vector Z of input variables. The available data is usually called features and the value to be estimated is known as target. Given a training set of N observations of any parameter where $n = 1, 2, \ldots N$, together with corresponding target values t_n, the aim is to predict the value of t for a new value of Z. The prediction should be done in such a way, so as to minimize the error function (loss function). This function is usually a squared loss function.

Consider a simple linear model consisting of input variables.

$$Y(Z, G) = G_0 + G_1 Z_1 + \ldots\ldots + G_D Z_D \tag{1}$$

where $Z = (Z_1, \ldots\ldots Z_D)^T$

Y represents the dependent variable or response variable and Z represents matrix of independent variables $Z_1, Z_2, Z_3..Z_D$ and $()^T$ stands for the transpose. G signifies parameter vector, where G_0 is the intercept term, and elements of the vector $G_1, G_2, G_3, G_4 \ldots .G_D$ are known as regression coefficients. Equation(1) can also be written as

$$Y(Z, G) = G_0 + \sum G_j \phi_j(Z) \tag{2}$$

where, $j = 1 \rightarrow M - 1$ and $\phi_j(Z)$ is known as the basis function and M represents the number of regression coefficients. If we define the function $\phi_0(Z) = 1$, the function can be expressed as

$$Y(Z, G) = \sum G_j \phi_j(Z) \tag{3}$$

where, $j = 1 \rightarrow M - 1$

That is

$$Y(Z, G) = G^T \phi(Z) \tag{4}$$

where $G = (G_0, G_1, G_2 \ldots G_{M-1})^T$ and $\phi(Z) = (\phi_1, \ldots \phi_{M-1})^T$

There will be differences between the predicted value and actual value. This difference can be positive or negative. To convert it into positive value, squares of each difference is counted and the sum of these squared values is estimated. This sum of squared error function is known as residuals. The sum of square error function should be minimized to obtain the maximum likelihood function; which makes the prediction reliable. The sum of square error function $E_D(G)$ can be represented as

$$E_D(G) = 1/2 \sum (t_n - G^T \phi(Z_n))^2 \tag{5}$$

where $n = 1, 2, 3 \ldots .N$, t_n represents target variable and $G^T \phi(Z_n)$ is the response variable or outcome. In practice, when we visualise a scatter plot using math function, a plot of output variable with respect to a series of input variables is obtained. Now, the focus is to find a line which fits best in the above scatter plot, so that the prediction of the response variable for any new values of the feature can be performed. This line is called regression line [30]. The equation of regression line is represented as:

$$h(Z_i) = b(0) + Z_i b(1) \tag{6}$$

Here, $h(Z_i)$ represents the predicted response measure for i^{th} observation. $b(0)$ and $b(1)$ are regression coefficients and signifies y-intercept and slope of the regression line respectively.

To create our model, we must "learn" or estimate the values of regression coefficients $b(0)$ and $b(1)$ and once we have estimated these coefficients, the model can be tested to predict responses. The step by step analysis performed on the data including the preprocessing methods is depicted in the following section. Taking multivariable regression as the base method, optimisation is performed on the data after checking the probability of improvement.

IV. PROPOSED METHOD

The Proposed method is divided into three major parts. Part A describes the preprocessing techniques used and part B discusses predictive analysis of non-optimised Linear Regression model. Part C is about log transformed model's performance while part D elaborates predictive performance of the Optimised Multivariable Regression model (reduced log. model).

A. PRE-PROCESSING

Preprocessing of data is the methods adopted to prepare the dataset for the next stage of predictive analysis, which involves application of machine learning algorithm [31]. The techniques utilized for the preprocessing stage of the proposed method are listed below.

1) CLEANING OF DATA AND FINDING DISTRIBUTION

Using various functions, the data is checked for any missing value and any characters which are not relevant to the context. It is also checked if there are any data points which are totally out of range. These types of points are known to be outliers and they tend to interfere with the efficiency of the model. So the distributions of data with various parameters are observed to see if they are normalised. Ideal distribution is normalised and outliers if any will be seen separately in the graph. Such outliers can be removed for better accuracy in performance. Some of the distribution plots obtained are shown in Fig. 3, Fig. 4 and Fig. 5.

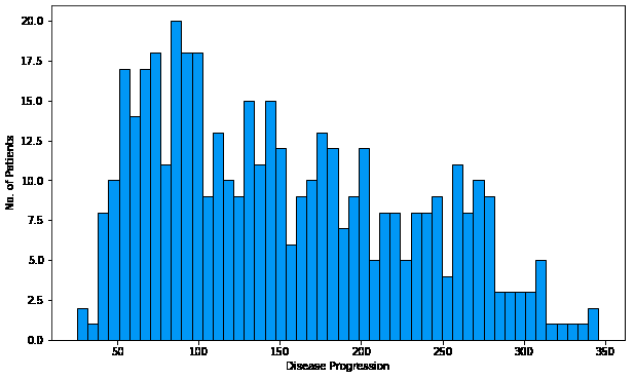


FIGURE 3. Histogram representation of number of patients versus diabetic disease progression after 1 year.

Fig. 3 represents disease progression versus number of patients corresponding to each value of the disease progression. It can be understood that the number of patients with very high value for the quantitative measurement of disease progression (300 and above) is less. Patients with value of disease progression above and below 140 are close to 50%. A plot of number of patients with Average Body Mass Index is as shown in Fig. 4.

Here BMI in the range of 20 to 30 are more in number and people with high BMI, greater than 40 are very less in number. The histogram plot of Average Blood Pressure versus Number of patients is depicted in Fig. 5.

From Fig. 5, it is clear that blood pressure more than 130 is away from the distribution. But the number of patients in this range is very less. So even though blood pressure values more than 130 can be taken as an outlier, it is not required to be removed as it will not affect the predictive analysis in a significant manner. The correlation between each parameter to one another is indicated in the Fig. 6, with the help of numerical values.

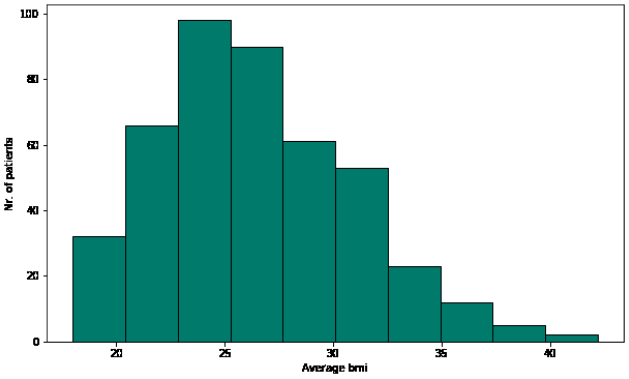


FIGURE 4. Histogram representation of number of patients versus average BMI of patients.

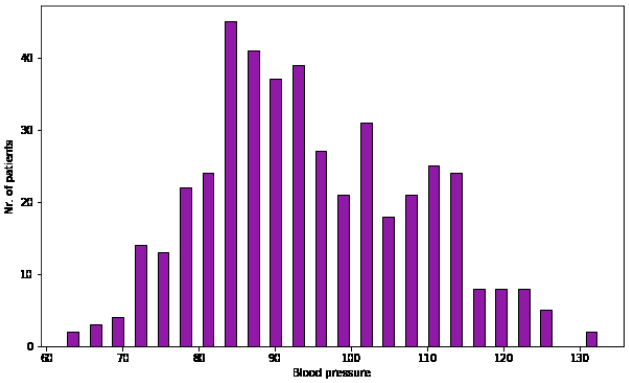


FIGURE 5. Histogram plot of number of patients versus average blood pressure of patients.

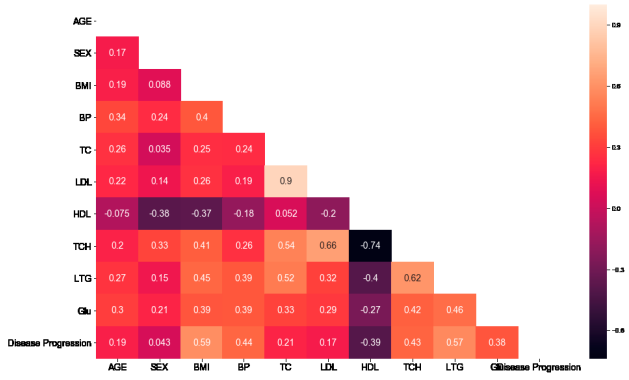


FIGURE 6. Heat map of correlation among all the features and target with colour coding.

Referring to fig. 6, it can be inferred that light colour indicates high correlation between parameters while dark colour signifies less correlation. If the correlation is very high, in the range of 0.9s, the issue of multicollinearity occurs; which indicates that if two variables are highly correlated, its appearance causes redundancy and affects the overall performance of the model. If the correlation of any factor is very less, then it indicates that the removal of that factor will not

affect the performance much. Here LDL and TC are highly correlated, which may cause multicollinearity and we can understand that LDL is less correlated with other parameters too. So elimination of LDL may not cause performance issue with the model than TC, which has better correlation with other factors. The correlation of variables whose values are not linear is not valid. Here variable SEX has only two values and hence its correlation with respect to any of the variables cannot be taken into consideration. A scatter plot of the correlation among some of the variables is shown in the Fig. 7.

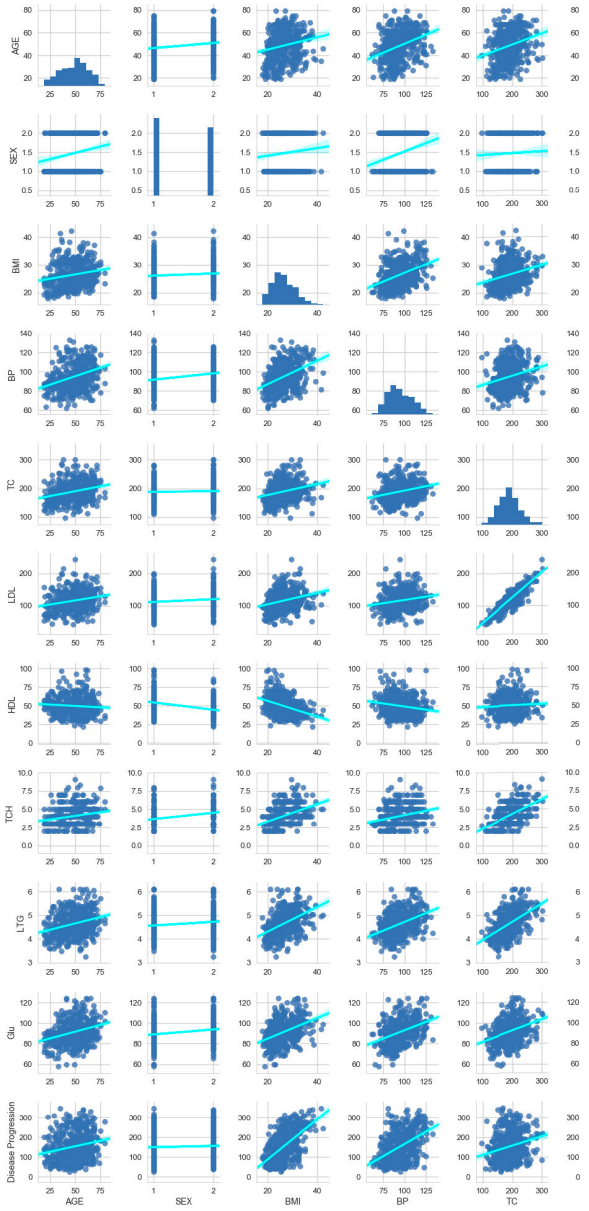


FIGURE 7. Scatter plot representation of correlation among some of the features.

From the scatter plot, relationship of each variable with one another can be inferred. It gives a pictorial representation of the correlation among different features, which was depicted in Fig. 6. Here the variables with only 2 values such as SEX is shown in discrete bars, whereas the variables having continuous values are indicated with the help of clustered data points. The high correlation between LDL and TC can be easily identified by analysing the highly linear plot between the two features.

Once the distribution of data is examined, values of certain statistical parameters of the dataset have been evaluated. Based on the observation, the process of feature selection is carried out, retaining the important features and eliminating the unimportant features, which can impact the performance of the model. The details of the method is presented next.

2) FEATURE SELECTION

To perform feature selection, the following factors are analysed.

P values: In statistics, P value indicates the probability of obtaining results as extreme as observed results of a statistical hypothesis test. A small value of P indicates robust evidence in favour of alternate hypothesis. Hence a P value more than 0.05 is not desirable in general. TC, LDL,HDL, TCH and Glu are parameters which show P value more than 0.05.

Variance Inflation Factor (VIF): The VIF parameter is used to test the issue with respect to multicollinearity. It indicates how much the variance of regression coefficient increases because of collinearity. VIF value more than 10 indicates severity with respect to multicollinearity and hence undesirable in general. Here TC, LDL, HDL values have values of VIF more than 10; While LTG is having a value of 10.48 (borderline).

Bayesian Information Criteria (BIC): It is a criterion for model selection among a set of finite models. When we choose parameters to increase the likelihood, it is also possible that over fitting may occur. In order to resolve this issue, BIC introduces a penalty term based on the number of features. According to statistics, as the BIC value gets lower, the model becomes better.

Considering the above parameters in mind, the model with all the parameters and reduced parameters after dropping features like LDL, TCH and TC are tested and the results are compared. Table 3 indicates the effect of dropping the parameters.

TABLE 3. Effect of feature reduction on BIC and R squared.

Dropped parameters	BIC	R squared
TC	407.582	0.475
TC, LDL and TCH	396.086	0.474

From Table 3, it is concluded that without much variation in R squared value (measure of variance), BIC value obtained after dropping parameters TC, LDL and TCH (which showed undesirable P value and VIF) is less compared to the value obtained after dropping the feature TC alone. Hence it

indicates that it is a better model when we drop the above 3 parameters. After these statistical tests, the data needs to be prepared for regression analysis, for which the splitting of data is performed with 80% as training data and 20% as testing data.

The regression plot is visualised to see the goodness of fit. Fig. 8 shows the plot between Disease Progression and one of the features, BMI. The data does not seem to follow a proper pattern and the slope of the regression line tends to vary in lower and upper regions, showing deviation of the predicted and actual values.

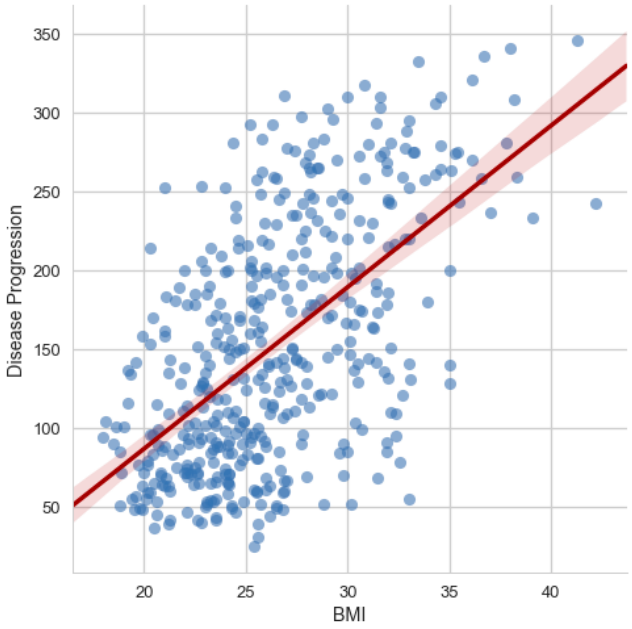


FIGURE 8. Regression plot of BMI versus disease progression.

B. PREDICTIVE ANALYSIS OF LINEAR REGRESSION MODEL

After conducting preprocessing and the regression analysis test, predictive performance of different models are experimented and compared. First and foremost, the linear regression model, without reducing the features is verified for its performance. The graph of actual versus predicted data of the model and the statistical parameters are as shown in Fig. 9.

In Fig. 9, the correlation is 0.69 and data points are not fitted to the line effectively. After dropping the features, the plot of regression model is as given in Fig. 10. We can observe the increase in correlation as 0.72, but still there is room for improvement with respect to the goodness of fit.

The residual plot of the regression model after feature reduction is as shown in Fig. 11.

Here, the residual plot indicates certain data points to be away from zero which is not desirable. A good residual plot should be clustered around zero and there should not be any pattern with respect to the plot. The probability density function (PDF) of the residual should be normally distributed for a good model. The obtained PDF is as shown in Fig. 12.

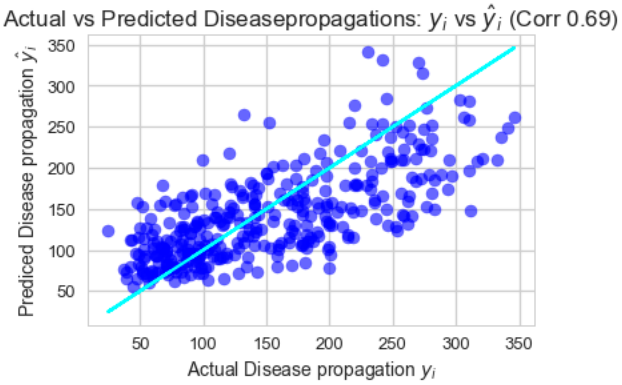


FIGURE 9. Scatter plot representation of actual versus Predicted disease progression of regression model.

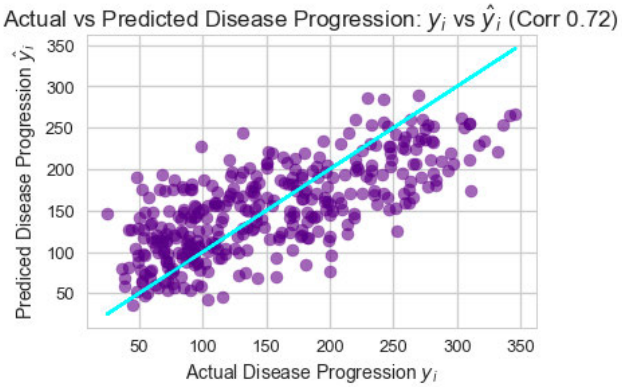


FIGURE 10. Scatter plot representation of actual versus predicted disease progression of regression model after feature reduction.

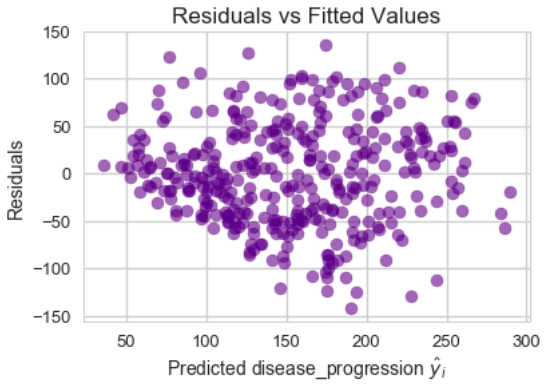


FIGURE 11. Residual plot of predicted values of the regression model.

The error analysis of the regression model is as shown in Table 4.

Referring to Table 4, RMSE, the root mean squared error tends to be 54.25 as per the regression model. As this error measurement does not indicate a very high precision, in order to reduce it and to reduce the skew; logarithmic transformation was tested on the model and the error analysis was conducted. The details of the process is given in the

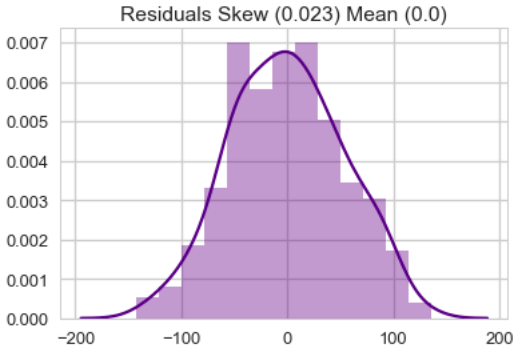


FIGURE 12. Probability density function plot of residuals of the regression model.

TABLE 4. Error analysis of the regression model.

MSE	R-squared	RMSE
2942.7370	0.51254	54 .247

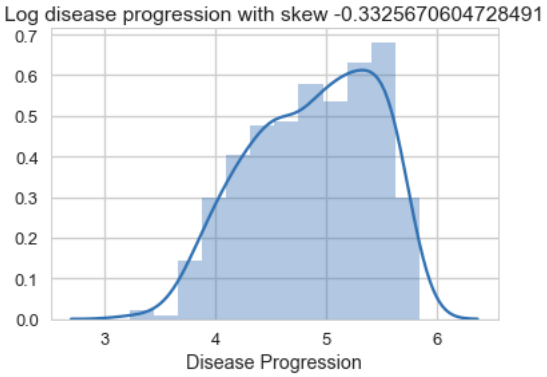


FIGURE 13. Probability density function of disease progression after logarithmic transformation.

following section. The probability density function (PDF) of the log. transformed disease progression is plotted in Fig. 13.

It can be found out that the reduced value of skew is −0.33, after applying logarithmic transformation, against 0.44 of the original model. The log. transformed plot of the disease progression with respect to one of the features, BMI is as shown in Fig. 14.

The plot of predicted versus actual value after performing log. transformation is represented in Fig. 15.

It can be understood from Fig. 15, that the goodness of fit is improved after transformation in comparison with the non-transformed regression model. The residual plot of the log transformed model is shown in Fig. 16.

It is observed that the data points are clustered around zero and it does not follow any pattern, which indicates the goodness of fit. In comparison to the regression model, the log transformed model's plot of residuals gives a better outcome. The PDF of log transformed residual plot in represented in Fig.17.

The plot in Fig. 17 shows less value for skew (−0.489) against 0.023 of the non-transformed regression model (B, Fig.12), and hence more efficient.

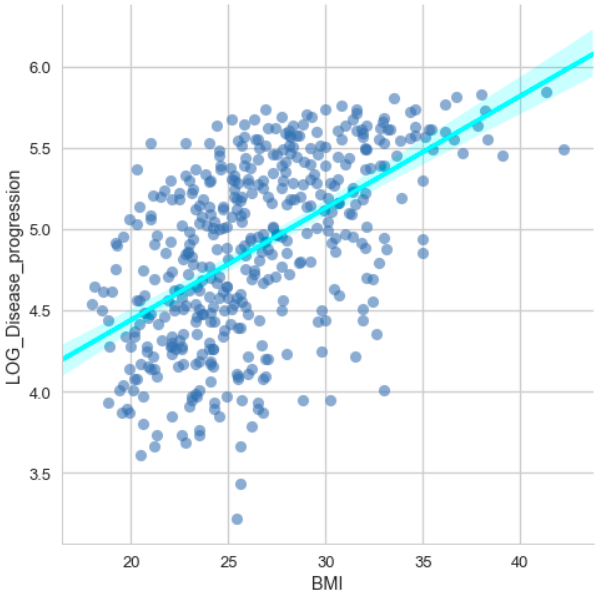


FIGURE 14. Scatter plot representation of log. transformed disease progression versus BMI.

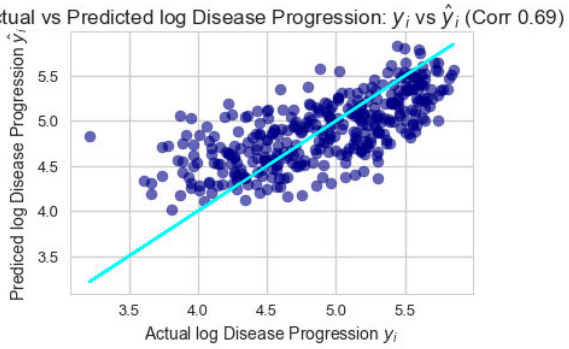


FIGURE 15. Scatter plot representation of predicted versus actual disease progression of log. transformed model.

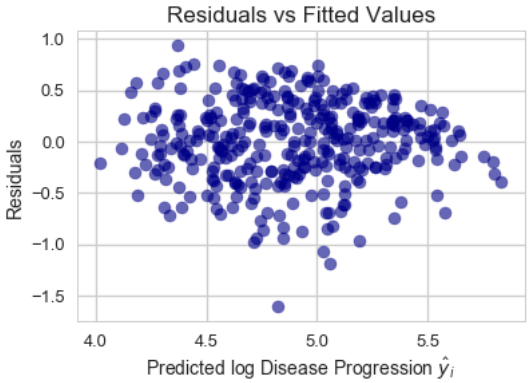


FIGURE 16. Residual plot of predicted disease progression.

Here, the model was tested for logarithmic transformation, without feature reduction. The difference in the performance by the optimised model, which performs both feature reduction and logarithmic transformation, can be explored from section C.

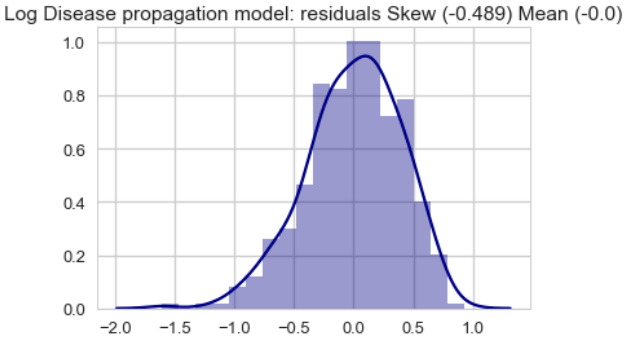


FIGURE 17. Probability density function of residuals of log transformed disease progression.

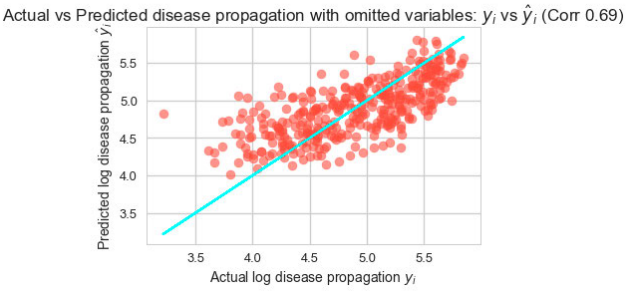


FIGURE 18. Scatter plot of predicted vs. actual values of the disease progression of the optimised model.

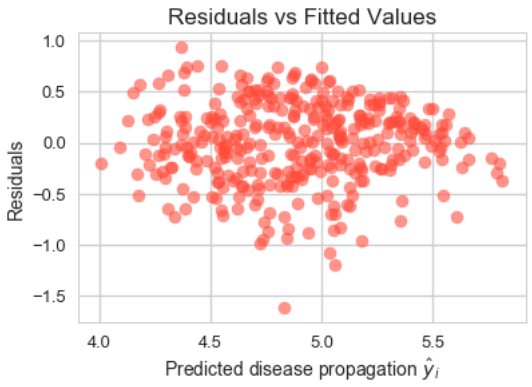


FIGURE 19. Residual plot representation of the optimised model.

C. PREDICTIVE ANALYSIS OF THE OPTIMISED MODEL

Here, after reducing or dropping the features, the log. transformation is performed. Fig. 18 shows the actual vs. predicted plot of the optimised model (reduced log model).

The scatter plot in Fig.18 shows improvement in fitting. The residual plot of the model is depicted in Fig. 19.

The residual plot is promising as most of the data points are centered on zero and there is no pattern. The results indicating the performance parameters are mentioned in section V.

V. RESULTS AND DISCUSSION

After conducting the analysis, the performance parameters of both the models are tabulated in Table 5.

TABLE 5. Performance analysis of non-optimised and optimised models.

Methods used	MSE	R- squared	RMSE
Non-optimised regression model	2942.73	0.51254	54.247
Optimised regression model(reduced log model)	0.1610	0.4750	1.5

From Table 5, it can be inferred that the logarithmic transformed model with feature reduction has given an RMSE value of 0.40 in log scale which indicates RMSE of 1.5 units in normal scale. So the predicted values and actual values differ by 1.5 units on an average in log transformed model with feature reduction. The regression model, without any transformation, produced RMSE value equal to 54.247 units which is not an accurate prediction compared to the predicted RMSE of optimised model. R-squared values of the non optimised regression model and optimised model are 0.51 and 0.47 respectively, which does not show much variation.

With reference to the literature used in Table 1 and 2, the dataset is also tested with existing popular and high performing individual models in literature such as KNN, SVR, Classification and Regression Tree(CART/Decision Tree) and RF. Promising Ensemble techniques such as Bagging Regressor, Adaboost, Xgboost and Stacking are also applied to compare and analyse the performance. Bagging, Adaboost and Xgboost are experimented on the data with Decision Tree as base model, while stacking was tested with KNN,SVR and CART as the base learners and Linear Regression as the meta learner. The basic parameters set for the above models are listed in Table 6 below. Referring to Table 6, number of splits has a value equal to 10 which signifies the value of K in K- fold cross validation and 3 indicates number of repetitions. So, 10-fold cross validation method with 3 repetitions was performed for each method presented above, as this method gives a reliable and robust result for these models [28]. The RMSE obtained by the proposed Optimised model, non optimised Multivariable Linear Regression model (LR) and the other existing methods applied as mentioned above have been tabulated in Table 7. The graphical representation of the table 7 is presented in Fig. 20.

TABLE 6. Parameters set for the performance analysis of existing models on the dataset.

Cross validation	Number of splits	Number of repeats
Repeated K-fold	10	3

Here, the proposed optimised Multivariable Regression model (indicated as opt. model in Fig. 20) is compared with non-optimised Linear Regression model (LR in Fig. 20) and other existing models in literature to analyse the performance on the Diabetic Disease Progression dataset. It can be

TABLE 7. Performance comparison of proposed model with other state of the art models.

Models used	RMSE
LR	54.2
KNN	55
SVR	63
CART	62
RF	58.3
Bagging	60.7
Adaboost	57.8
Xgboost	64.7
Stacking	62
Optimised model	1.5

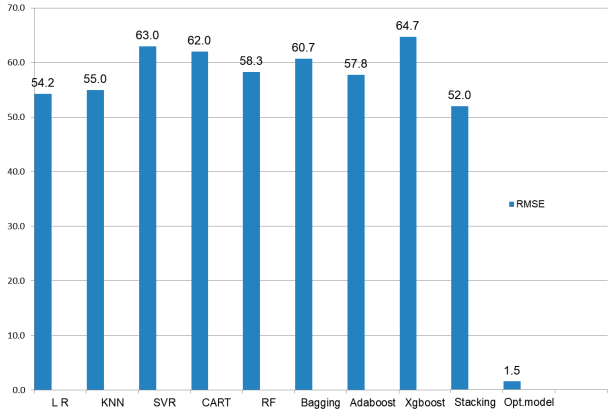


FIGURE 20. Comparison plot of RMSE score of proposed optimised model with other state of the art models.

inferred that, the dataset has given RMSE values in the range of 54 to 65 for the tested existing models, while the optimised Regression model has given RMSE of 1.5 units, which shows the supremacy of the method on the dataset. Papers in literature such as [15], [18] which employ individual models such as KNN, SVR and articles [26]–[29], which handle ensembles, perform well with minimum error with respect to classifying diabetic and non-diabetic outcomes and in predicting short term diabetic information. But these learners, when used for the diabetic progression dataset have not performed so well in comparison with the proposed technique. Referring to Table 1 and 2, it can be inferred that the RMSE values in literature, pertaining to various diabetic datasets and techniques, varies from.005 to 37 units. Hence, it can be observed that RMSE of 1.5 units of the proposed model falls in the low range, even though, it is not the least score obtained. The Ensemble methods, having multiple learning levels are known to perform well with many predictive analysis scenarios. But the proposed model, outperforms them for the regression problem under consideration. The use of Linear Regression as a base model, which is known for its simple and robust mathematical mapping procedure, helps the proposed optimised model to stand out. In addition, the in-depth statistical analysis conducted in this paper, to choose the right optimisation for the dataset considered, makes the proposed model unique and precise.

VI. CONCLUSION AND FUTURE WORK

The optimised Multivariable Regression method used for predictive analysis of Diabetic Disease Progression, one year after the baseline is proven to be efficient, in terms of RMSE, residual plot as well as the linear regression plot between actual and predicted values. The RMSE value obtained in log scale is 0.4, which indicates a variation of 1.5 units from the actual values. The non-optimised regression model produced RMSE of 54.247 only, exposing the superiority of optimised model. Comparing with the existing popular and successful methods in literature, the proposed model has shown significant improvement in performance. In contrast with the diabetic prediction methods in literature, the insight into the diabetic disease progression with respect to 10 input features with a good number of patients, is an effective measure in understanding the individuals' lifestyle and it could help the doctors in providing precise medical advice. However, this model does not consider any change in a person's lifestyle or any other blood sample measurements, after taking the samples at a particular time. The model could be used further in combination with other parameters which are likely to influence diabetic disease progression. It can also be tested for other datasets for finding out the quantitative values of diabetic disease progression.

In future, other machine learning techniques with different combinations will be tested and performance measures will be compared. Since highest accuracy and least error is very important in medical field, further improvement in prediction will be attempted. We are also considering, to introduce various other parameters, which could influence the progression of the disease. Classification of disease progression into high risk and low risk categories by keeping a threshold value for the outcome, will also be carried out; with the help of various classification based machine learning algorithms. After optimising the high performing algorithms, the best model will be identified. This study will certainly provide value addition to the research community, working in the field of machine learning related to healthcare.

REFERENCES

[1] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for Internet of Things data analysis: A survey," *Digit. Commun. Netw.*, vol. 4, no. 3, pp. 161–175, Aug. 2018.
[2] Q. Ang, Z. Liu, W. Wang, and K. Li, "Explored research on data preprocessing and mining technology for clinical data applications," in *Proc. 2nd IEEE Int. Conf. Inf. Manage. Eng.*, Apr. 2010, pp. 327–330.
[3] Y. Roh, G. Heo, and S. Euijong Whang, "A survey on data collection for machine learning: A big data—AI integration perspective," 2018, *arXiv:1811.03402*. [Online]. Available: http://arxiv.org/abs/1811.03402
[4] D. S. Kumar, "A comparative study of various data transformation techniques in data mining," *Int. J. Sci. Eng. Technol.*, vol. 4, no. 3, pp. 146–148, Mar. 2015.
[5] M. Marjani, F. Nasaruddin, A. Gani, A. Karim, I. A. T. Hashem, A. Siddiqa, and I. Yaqoob, "Big IoT data analytics: Architecture, opportunities, and open research challenges," *IEEE Access*, vol. 5, pp. 5247–5261, 2017.
[6] V. K. Daliya and T. K. Ramesh, "A survey on enhancing the interoperability aspect of IoT based systems," in *Proc. IEEE Conf. Smart Technol. Smart Nation*, Aug. 2017, pp. 581–586.
[7] V. K. Daliya and T. K. Ramesh, "Data interoperability enhancement of electronic health record data using a hybrid model," in *Proc. Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Nov. 2019, pp. 318–322.

[8] A. Moraru, M. Pesko, M. Porcius, C. Fortuna, D. Mladenic, M. Serrano, P. Barnaghi, and P. Cousin, "Using machine learning on sensor data," *J. Comput. Inf. Technol.*, vol. 18, no. 4, pp. 341–347, 2010.

[9] P. Kalaiyarasi and J. Suguna, "Prediction of diabetic disease using ensemble classifier," *Int. J. Psychosocial Rehabil.*, vol. 24, no. 7, pp. 91–109, 2020.

[10] A. Z. Woldaregay, E. Årsand, S. Walderhaug, D. Albers, L. Mamykina, T. Botsis, and G. Hartvigsen, "Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes," *Artif. Intell. Med.*, vol. 98, pp. 109–134, Jul. 2019.

[11] F. Allam, Z. Nossai, H. Gomma, I. Ibrahim, and M. Abdelsalam, "A recurrent neural network approach for predicting glucose concentration in type-1 diabetic patients," in *Proc. 12th INNS EANN-SIG Int. Conf. Eng. Appl. Neural Netw. (EANN)*, 2011, pp. 254–259.

[12] F. Allam, Z. Nossair, H. Gomma, I. Ibrahim, and M. Abd-el Salam, "Prediction of subcutaneous glucose concentration for type-1 diabetic patients using a feed forward neural network," in *Proc. Int. Conf. Comput. Eng. Syst.*, Nov. 2011, pp. 129–133.

[13] E. Daskalaki, A. Prountzou, P. Diem, and S. G. Mougiakakou, "Real-time adaptive models for the personalized prediction of glycemic profile in type 1 diabetes patients," *Diabetes Technol. Therapeutics*, vol. 14, no. 2, pp. 168–174, Feb. 2012.

[14] W. Xao, F. Shao, J. Ji, R. Sun, and C. Xing, "Fasting blood glucose change prediction model based on medical examination data and data mining techniques," in *Proc. IEEE Int. Conf. Smart City/SocialCom/SustainCom (SmartCity)*, Dec. 2015, pp. 742–747.

[15] T. Hamdi, J. B. Ali, V. Di Costanzo, F. Fnaiech, E. Moreau, and J.-M. Ginoux, "Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm," *Biocybernetics Biomed. Eng.*, vol. 38, no. 2, pp. 362–372, 2018.

[16] J. I. Hidalgo, J. M. Colmenar, G. Kronberger, S. M. Winkler, O. Garnica, and J. Lanchares, "Data based prediction of blood glucose concentrations using evolutionary methods," *J. Med. Syst.*, vol. 41, no. 9, pp. 1–20, Sep. 2017.

[17] J. I. Hidalgo, J. M. Colmenar, J. L. Risco-Martin, A. Cuesta-Infante, E. Maqueda, M. Botella, and J. A. Rubio, "Modeling glycemia in humans by means of grammatical evolution," *Appl. Soft Comput.*, vol. 20, pp. 40–53, Jul. 2014.

[18] I. H. Sarker, F. Faruque, H. Alqahtani, and A. Kalim, "K-nearest neighbor learning based diabetes mellitus prediction and analysis for eHealth services," *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 7, no. 26, pp. e4-1–e4-9, 2020.

[19] C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli, "Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 6, pp. 1550–1560, Jun. 2012.

[20] C. Zecchin, A. Facchinetti, G. Sparacino, and C. Cobelli, "Jump neural network for Realtime prediction of glucose concentration," in *Artificial Neural Networks*, H. Cartwright, Ed. New York, NY, USA: Springer, 2015, pp. 245–259.

[21] K. Zarkogianni, K. Mitsis, E. Litsa, M. T. Arredondo, G. Fico, A. Fioravanti, and K. S. Nikita, "Comparative assessment of glucose prediction models for patients with type1 diabetes mellitus applying sensors for glucose and physical activity monitoring," *Med. Biol. Eng. Comput.*, vol. 53, no. 12, pp. 1333–1343, 2015.

[22] Y. Wang, X. Wu, and X. Mo, "A novel adaptive-weighted-average framework for blood glucose prediction," *Diabetes Technol. Therapeutics*, vol. 15, no. 10, pp. 792–801, Oct. 2013.

[23] E. Otto, C. Semotok, J. Andrysek, and O. Basir, "An intelligent diabetes software prototype: Predicting blood glucose levels and recommending regimen changes," *Diabetes Technol. Therapeutics*, vol. 2, no. 4, pp. 569–576, Dec. 2000.

[24] J. Li and C. Fernando, "Smartphone-based personalized blood glucose prediction," *ICT Exp.*, vol. 2, no. 4, pp. 150–154, Dec. 2016.

[25] E. I. Georga, V. C. Protopappas, D. Polyzos, and D. I. Fotiadis, "Predictive modeling of glucose metabolism using free-living data of type 1 diabetic patients," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, Aug. 2010, pp. 589–592.

[26] G. Alfian, M. Syafrudin, J. Rhee, M. Anshari, M. Mustakim, and I. Fahrurrozi, "Blood glucose prediction model for type 1 diabetes based on extreme gradient boosting," in *Proc. Int. Conf. Inf. Technol. Digit. Appl.*, vol. 803/012012, 2019, pp. 1–7.

[27] J. Abdollahi and B. Nouri-Moghaddam, "Hybrid stacked ensemble combined with genetic algorithms for prediction of diabetes," 2021, *arXiv:2103.08186*. [Online]. Available: http://arxiv.org/abs/2103.08186

[28] B. A. Tama and K.-H. Rhee, "Tree-based classifier ensembles for early detection method of diabetes: An exploratory study," *Artif. Intell. Rev.*, vol. 51, no. 3, pp. 355–370, Mar. 2019.

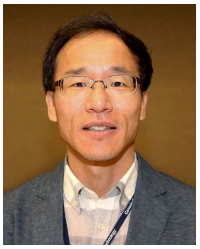
[29] J. Liu, L. Wang, L. Zhang, Z. Zhang, and S. Zhang, "Predictive analytics for blood glucose concentration: An empirical study using the tree-based ensemble approach," Library Hi Tech, Emerald Group Publishing, Bingley, U.K., Tech. Rep., 2020, doi: 10.1108/lht-08-2019-0171.

[30] E. I. Georga, V. C. Protopappas, D. Polyzos, and D. I. Fotiadis, "Evaluation of short-term predictors of glucose concentration in type 1 diabetes combining feature ranking with regression models," *Med. Biol. Eng. Comput.*, vol. 53, no. 12, pp. 1305–1318, Dec. 2015.

[31] U. S. Shanthamallu, A. Spanias, C. Tepedelenlioglu, and M. Stanley, "A brief survey of machine learning methods and their sensor and IoT applications," in *Proc. 8th Int. Conf. Inf., Intell., Syst. Appl. (IISA)*, Aug. 2017, pp. 1–8.

V. K. DALIYA received the M.Tech. degree in embedded systems technology. She is currently a part-time Research Scholar in machine learning, under the guidance of Dr. T. K. Ramesh, with the Department of Electronics and Communication Engineering, Amrita School of Engineering, Bengaluru. She is associated with New Horizon College of Engineering as Senior Assistant Professor in the Department of Electronics and Communication Engineering. She has ten years of teaching experience in various engineering colleges and three years of research experience. She has guided the academic project works of many M.Tech. and B.Tech. students. Her research works and guided projects are published in various IEEE conferences and international journals. Her research interests include artificial intelligence, application of machine learning in healthcare, and data analysis in the IoT. She is a Life Member of the Indian Society for Technical Education.

T. K. RAMESH (Member, IEEE) received the Ph.D. degree in optical networks from Amrita Vishwa Vidyapeetham. He is currently an Associate Professor with the Department of Electronics and Communication Engineering, Amrita School of Engineering, Bengaluru. He has successfully guided four Ph.D. students. He is also guiding eight Ph.D. scholars. He has published over 80 research publications in peer-reviewed international journals and conferences. His research interests include communication networks and its applications, analog and digital devices and circuits, functional safety, artificial intelligence, and network-on-chip. He has 28 years of teaching and research experience. He is a Lifetime Member of the Indian Society for Technical Education and a member of the Institution of Electronics and Telecommunication Engineers.



SEOK-BUM KO (Senior Member, IEEE) received the Ph.D. degree from The University of Rhode Island, USA, in 2002. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Saskatchewan, Canada, and the Division of Biomedical Engineering, University of Saskatchewan. His research interests include computer architecture/arithmetic, efficient hardware implementation of compute-intensive applications, deep learning processor architecture, and biomedical engineering. He is also a Senior Member of the IEEE Circuits and Systems Society and an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS and IEEE ACCESS.