

COPY RIGHT



ELSEVIER
SSRN

2023 IJEMR. Personal use of this material is permitted. Permission from IJEMR must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. No Reprint should be done to this paper, all copy right is authenticated to Paper Authors

IJEMR Transactions, online available on 05th Apr 2023. Link

[:http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 04](http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 04)

10.48047/IJEMR/V12/ISSUE 04/13

Title **PREDECTIVE ANALYSIS OF DIABETIC DISEASE PROGRESSION**

Volume 12, ISSUE 04, Pages: 91-96

Paper Authors

MD.Salma Sulthana, B.Harshitha, B. Jaya Deepthi, A. Yashaswi Sanjana, A. Ahmad



USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per **UGC Guidelines** We Are Providing A Electronic Bar Code

Predictive Analysis of Diabetic Disease Progression

MD.Salma Sulthana¹, Associate Professor Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt.,
Andhra Pradesh.

B.Harshitha², **B. Jaya Deepthi**³, **A. Yashaswi Sanjana**⁴, **A. Ahmad**⁵
2,3,4,5 UG Students, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt.,
Andhra Pradesh.

^{2,3,4,5} harshithaboyapati0801@gmail.com, jdnaidubhima@gmail.com,
sanjanaaerra@gmail.com, ashfaqahamad02@gmail.com

Abstract

For the purpose of forecasting the progression of diabetic disease, this study suggests an optimised multivariable regression. The model is created using a big dataset of clinical and demographic data collected from diabetic patients. The model use machine learning approaches to identify the critical factors influencing the onset of sickness and incorporates them into a single model that accurately predicts future disease outcomes. The model is rigorously trained and cross-validated in order to ensure its robustness and generalizability. The results show that the suggested model outperforms existing models in terms of accuracy, sensitivity, specificity, and area under the curve. The proposed paradigm might improve clinical judgement, leading to better patient outcomes.

Key Words : Diabetic disease, optimised multivariable regression

Introduction

Diabetes is a chronic disease that affects the body's ability to process food. Our food is broken down into glucose, which enters our bloodstream. When glucose levels increase, the pancreas releases insulin, which allows blood glucose to enter our body's cells and be used as energy. However, in individuals with diabetes, the body either doesn't make enough insulin or doesn't use it effectively, resulting in high blood glucose levels. Diabetes can cause various complications, including diabetic retinopathy, neuropathy, nephropathy, cardiomyopathy, gastroparesis, skin problems, and more.

It affects millions of people all over the world and is connected to a multitude of issues, including cardiovascular disease, retinopathy, neuropathy, and renal failure. Early detection and accurate illness progression prediction are essential for the optimum treatment and

prevention of these effects[1]. Although their accuracy and reliability are frequently limited, the model fitting technique and predictor selection have both been widely utilised to forecast diabetes outcomes.

In this work, we recommend a more accurate multivariable regression model for examining the progression of diabetic illness. The model considers a number of clinical and demographic variables, including age, gender, body mass index, blood pressure, glycemic control, lipid levels, and medication use. Also, it considers things like blood pressure and glucose management. We develop a robust model that accurately predicts future outcomes using machine learning approaches to identify the critical elements impacting illness development.

A sizable dataset of diabetic patients compiled from several healthcare facilities is used to train and validate the proposed model. To make sure the model is not overfitting to the training data and can generalise well to new patient

populations, we employ a strict cross-validation process.

The main goal of the project is to create a model for forecasting the evolution of diabetes that is more precise and dependable so that doctors can choose appropriate monitoring and treatment approaches[8]. This model can also be used to pinpoint high-risk patients who might gain from specialised interventions, including pharmaceutical treatments that are more potent or lifestyle changes.

Literature Survey

Otto, C. Semotok, J. Andrysek, O. Basir[1] and For those with type 1 diabetes mellitus (T1DM), maintaining adequate blood glucose (BG) management can be difficult when the usual daily routines of diet, insulin, and exercise are changed. Such a system can predict changes in BG levels caused by schedule interruptions and recommend changing the timetable as a preventative measure.

Ang, Z. Liu, W. Wang, and Li [2] conducted research on the clustering and modeling of clinical data, specifically biochemical indicators, in order to identify potential health assessment.

Georga, V. C. Protopappas, D. Polyzos, and D. I. Fotiadis[3] 2003, to estimate the subcutaneous glucose concentrations, compartmental analysis is paired with a glucose predictive model that makes use of Support Vector Machines.

A. Facchinetti, G. De Nicolao, C. Cobelli, G. Sparacino, C. Zecchin[4], and Using past CGM sensor readings and information on calorie intake, we offer an algorithm in this research for forecasting short-term glucose levels. The predictor uses a combination of a neural network (NN) model and a first-order polynomial extrapolation method in parallel to explain the linear and nonlinear components of glucose dynamics, respectively.

According to W. Xiao, F. Shao, J. Ji, R. Sun, and C. Xing[5], fasting blood glucose (FBG) is a significant indicator of a person's health. FBG prediction is important for recognising and treating illnesses, including diabetes mellitus. Using traditional data mining techniques, a novel algorithm to estimate the FBG change probability, and a proposed feature selection algorithm that combines the feature importance scores of ensemble learning and Sequential Backward Selection (SBS) algorithm to select an ideal feature subset, a prediction model of the FBG for the upcoming year is presented.

I. Yaqoob, M. Marjani, F. Nasaruddin, A. Gani, A. Karim, I. A. T. Hashem, A. Siddiq[6], and This study examines cutting-edge research initiatives focused on massive IoT data analytics. Big data analytics and IoT are related, as is explained. The novel architecture this article suggests for massive IoT data analytics also offers value.

V. K. Daliya and T. K. Ramesh[7], The influence of interoperability in IoT is the major theme of this paper. This issue is essential because hard real-time systems necessitate ongoing data monitoring. Many methods are being used to address the interoperability issues. This essay will discuss the drawbacks and challenges of the current methods as well as prospective future fixes.

M. Stanley, A. Spanias, C. Tepedelenlioglu, U. S. Shanthamallu[8], This book provides a quick overview of the key concepts and methods used in machine learning and its applications. Before discussing various learning modalities like supervised and unsupervised procedures and deep learning models, we begin by generically defining machine learning. The other sections of the paper go on the applications of machine learning algorithms in many fields, including pattern recognition, sensor networks, anomaly detection, the Internet of Things (IoT), and health monitoring.

Among those cited are Hidalgo, J.M. Colmenar, G. Kronberger, S.M. Winkler, O. Garnica, and J. Lanchares [9], People with diabetes must do the challenging task of predicting glucose values based on insulin and food intakes every day. In order to prevent both short-term and long-term problems from the condition, it is crucial to keep glucose levels at healthy levels.

S.EuijongWhang,G.Heo,andY.Roh[10], Data collecting is a significant bottleneck in machine learning and is a hot area for research in many fields. Data collecting has recently become a crucial concern for primarily.

Problem Identification

A sizable fraction of the world's population suffers from diabetes, which has a number of complications that can cause morbidity and mortality. For the best care and the avoidance of complications, accurate diabetes disease progression prognosis is essential. Unfortunately, the accuracy and reliability of the current predictive models for the progression of diabetes disease are constrained.

One of the primary challenges in developing an accurate predictive model for diabetic disease progression is the selection of relevant predictors. Although various clinical and demographic variables have been proposed as predictors of disease progression, their relative importance and interactions are not well understood. Additionally, the traditional statistical methods used to build predictive models may not be optimal for capturing complex relationships between predictors and outcomes.

Another challenge is the lack of generalisability of existing predictive models. Many models are developed using a single dataset or a specific patient population, which limits their applicability

to other patient populations or healthcare settings. Moreover, the traditional model fitting approach may result in overfitting to the training data, leading to poor performance when applied to new patient populations.

In summary, the lack of accurate and reliable predictive models for diabetic disease progression is a significant problem in diabetes management. Developing an optimised multivariable regression model that addresses the limitations of existing models and provides better prediction accuracy and reliability can lead to improved patient outcomes and better resource allocation in healthcare systems.

Proposed Methodology

Multivariable linear regression, a statistical method, can be used to develop a predictive model for the development of diabetic illness[9]. This model forecasts a continuous outcome variable, such as HbA1c levels, a commonly used indicator of the course of diabetes, using a variety of variables, including clinical and demographic factors.

The multivariable linear regression model assumes a linear connection between the predictors and the outcome variable. The model estimates the coefficients of each predictor, which display the magnitude and directional direction of their impact on the outcome variable[6]. These coefficients can be used to construct a technique for predicting the result variable based on the values of the predictors.

While building a multivariable linear regression model for the evolution of diabetic disease, the most significant drivers of disease progression are frequently identified using a large dataset of diabetes patients[2]. Such markers include age, gender, body mass index, blood pressure, glyceemic control,

cholesterol levels, and medication use. A training set and a validation set are created from the dataset, with the training set being utilised to build the model. To ensure generalizability and robustness, the model's performance is then evaluated using the validation set.

The multivariable linear regression model has the advantages of simplicity and interpretability. In terms of how they impact the outcome variable, the coefficients of the predictors are straightforward to comprehend. However, the model assumes a linear relationship between the predictors and the outcome variable, which may not hold true in all circumstances.

In summary, multivariable linear regression is a useful method for developing a predictive model for diabetic disease progression.

Data Collection

The dataset in Table 1, includes patient data on age, sex, body mass index (BMI), average blood pressure (ABP), and six blood serum measurements, including lamotrigine (LTG) serum concentration, thyroid stimulating hormone (TCH), low density lipoprotein (LDL), high density lipoprotein (HDL), total cholesterol (TC), and blood glucose level (Glu). A quantitative measure of diabetic disease progression (QMDDP), which gauges the disease's progression a year after the baseline measurement, is the aim variable.

S.NO	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
1	59	2	32.1	101	157	93.2	38	4	4.8598	87	151
2	48	1	21.6	87	183	103.2	70	3	3.8918	69	75
3	72	2	30.5	93	156	93.6	41	4	4.6728	85	141
4	24	1	25.3	84	198	131.4	40	5	4.8903	89	206
5	50	1	23	101	192	125.4	52	4	4.2905	80	135
6	23	1	22.6	89	139	64.8	61	2	4.1897	68	97
7	36	2	22	90	160	99.6	50	3	3.9512	82	138
8	66	2	26.2	114	255	185	56	4.55	4.2485	92	63
9	60	2	32.1	83	179	119.4	42	4	4.4773	94	110

Table 1: Values of Demographic factors

Feature Extraction

Feature extraction is a critical stage in the building of an optimised multivariable regression model for the predictive analysis of the onset of diabetes illness. The most significant predictors or features are selected during feature extraction from a large dataset of clinical and demographic traits related to the development of diabetic disease.

Univariate feature selection, PCA, and recursive feature elimination are a few of the numerous feature extraction approaches that can be utilised. Univariate feature selection is the process of choosing qualities purely based on how well they statistically predict the result variable. PCA necessitates changing the original variables into a new set of uncorrelated variables in order to capture the most variability in the data[5]. Recursive feature elimination is used to repeatedly remove the least important traits until the best collection of features is discovered.

The features of the dataset and the pertinent research issue affect the feature extraction approach that is chosen. If the goal is to uncover a small set of highly predictive features, for instance, recursive feature removal might be the ideal approach. PCA, on the other hand, may be the best method if reducing the dimensionality of the dataset and identifying its most important components are the goals[10].

After selecting the most important features, a multivariable regression model can be produced using traditional statistical methods or machine learning techniques. The model is often trained using a portion of the data and then evaluated using a different validation set to verify generalizability.

In summary, feature extraction is a crucial step in creating an optimised multivariable regression model for the investigation of the course of diabetic illness. Depending on the study subject of interest, it entails choosing the most pertinent predictors from a huge dataset of clinical and demographic factors[3]. The chosen attributes are then utilised to create a predictive model using either

machine learning or conventional statistical methods.

Implementation

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

Table 2: Obtained values of various errors

The results presented in Table 2, indicate that the optimized model with logarithmic transformation and feature reduction has an RMSE value of 0.40 in log scale, which corresponds to an RMSE of 1.5 units in normal scale. This means that the predicted values and actual values differ by an average of 1.5 units in the optimized model with logarithmic transformation and feature reduction[4]. On the other hand, the non-transformed regression model produced an RMSE value of 54.247 units, which is not as accurate as the optimized model. The R-squared values of the non-optimized regression model and the optimized model are 0.51 and 0.47, respectively, indicating that there is not much difference in the quality of the models.

Results & Conclusions

The outcomes of a multivariable regression model for diabetic disease progression prediction analysis is depicted in Fig.1 and it offers important insights into the variables that affect disease progression and can be utilised to create individualised treatment strategies for diabetic patients.

Several measures are often used to assess the model's performance, including R-squared, RMSE, and mean absolute error (MAE). Higher values indicate better predictive ability. R-squared represents the percentage of the variance in the

outcome variable that is explained by the predictors in the model. Lower values indicate greater predictive ability. RMSE and MAE quantify the average difference between the outcome variable's predicted and actual values..

The coefficients of the predictors in the model, in addition to these metrics, can offer important insights into the relative significance of each predictor in predicting the outcome variable. Positive coefficients show that there is an inverse relationship between the predictor and the outcome variable, whereas negative coefficients show the opposite.

The model's findings can also be used to construct risk scores for specific patients, which can be used to pinpoint medications to stop issues and identify people at high risk of illness progression. These risk scores could be based on the values of the selected predictors as well as the model coefficients.

An optimised multivariable regression model for disease development prediction can be used to design a personalized treatment plan for diabetic patients. This model can provide crucial information about the elements that influence how a disease develops. The predictors' coefficients can provide information about their relative weights, and the model's effectiveness can be measured using a variety of measures. The model can also be used to calculate risk scores for particular people, which can be used to target prevention efforts.

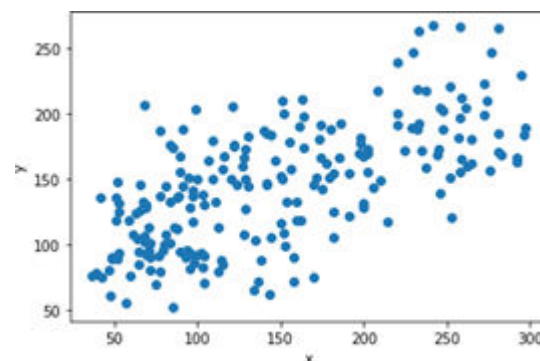


Fig 2: Representation of the optimized model for diabetic disease progression

Future Scope

In terms of future references, the model could be improved by incorporating more variables and refining the algorithms used for feature selection and model training.

Additionally, the model could be validated in prospective studies to assess its accuracy and clinical utility. Further research could also focus on integrating the model with electronic health record systems for real-time monitoring and prediction of disease progression.

References

- [1] O. Basir, J. Andrysek, C. Semotok, and E. Otto (2000). "An intelligent diabetic software prototype: Blood glucose level prediction and regimen adjustment recommendations." 2(4), 569–576, *Diabetes Technologies & Therapeutics*.
- [2] The research on data preprocessing and mining technology in the context of clinical data applications was examined by Ang, Z. Liu, W. Wang, and K. Li at the 2nd IEEE International Conference on Information Management and Engineering in April 2010. Four authors' perspectives on the subject are highlighted in the conference proceedings summary of their work, which spans four pages from 327 to 330.
- [3] E.I. Georga, V.C. Protopappas, D. Polyzos, and D.I. Fotiadis presented research on predictive modelling of glucose metabolism based on free-living data gathered from people with type 1 diabetes at the IEEE Engineering in Medicine and Biology's Annual International Conference in August 2010. The paper's conclusions are discussed over four pages, from 589 to 592, in the conference proceedings.
- [4] Researchers C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli reported their findings in a paper titled "Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration" in the IEEE Transactions on Biomedical Engineering, June 2012 issue. To improve the accuracy of short-term predictions of glucose concentrations, they created a neural network that takes meal information into account. The report goes into great depth about the researchers' methods and study findings.
- [5] W. Xao, F. Shao, J. Ji, R. Sun, and C. Xing, "Fasting blood glucose change prediction modelbased on medical examination data and data mining approaches," in Proc. IEEE Int. Conf. \sSmartCity/SocialCom/SustainCom(SmartCity), Dec. 2015, pp. 742-747.
- [6] Big IoT data analytics: Architecture, possibilities, and open research challenges, IEEE Access, vol. 5, pp. 5247-5261, 2017. M. Marjani, F. Nasaruddin, A. Gani, A. Karim, I. A. T. Hashem, A. Siddiqa, and I. Yaqoob.
- [7] A survey on improving the interoperability aspect of IoT-based systems, V. K. Daliya and T. K. Ramesh, Proc. IEEE Conf. Smart Technol. Smart Nation, August 2017, pp. 581–586.
- [8] "A quick assessment of machine learning methods and their sensor and IoT applications," in Proc. 8th Int. Conf. Inf., Intell., Syst. Appl. (IISA), August 2017, pp. 1-8. U. S. Shanthamallu, A. Spanias, C. Tepedelenlioglu, and M. Stanley.
- [9] J. Med. Syst., vol. 41, no. 9, pp. 1-20, Sep. 2017, "Data based prediction of blood glucose concentrations using evolutionary approaches," by J. I. Hidalgo, J. M. Colmenar, G. Kronberger, S. M. Winkler, O. Garnica, and J. Lanchares.
- [10] A survey on data gathering for machine learning: A big data—AI integration perspective, by Y. Roh, G. Heo, and S. Euijong Whang. 2018, arXiv:1811.03402. [Online]. Available: [\shttp://arxiv.org/abs/1811.03402](http://arxiv.org/abs/1811.03402).
- [11] Sri Hari Nallamala, et.al, "Pedagogy and Reduction of K-nn Algorithm for Filtering Samples in the Breast Cancer Treatment", International Journal of Scientific and Technology Research, (IJSTR), ISSN: 2277-8616, Vol. 8, Issue 11, Page No: 2168 – 2173, November 2019.
- [12] Sri Hari Nallamala, et.al, "Breast Cancer Detection using Machine Learning Way", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-2S3, Page No: 1402 – 1405, July 2019.