International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 13 Issue 2, February-2024, Impact Factor: 8.375

Time Series Analytics for Predictive Risk Monitoring in Diabetes Care

Sihao Wang

Department of Mathematics, Southern Methodist University, TX, United States

ABSTRACT

Diabetes has become a global public health crisis with over 463 million afflicted currently. Early identification of worsening glycemic control enables timely interventions, vastly improving outcomes. This work develops data-driven time series forecasting models using longitudinal patient glucose readings for predictive alerts. Statistical methods like ARIMA, SARIMA and machine learning algorithms Long Short-Term Memory (LSTM) networks and Prophet were analyzed on a 18-month high-frequency diabetes biomarker dataset. LSTM outperformed other approaches with lowest errors and highest determination coefficient of 0.91, accurately capturing cyclic trends and future spikes. Results indicate viability of temporal predictive models in providing clinical decision support through quantitative personalized risk scores about glucose regulation deterioration weeks in advance. This enables preemptive action for stabilizing glycemic statuses. Advanced analytics on emerging biomedical time series data promises to strengthen preventive care for tackling the diabetes epidemic.

Keywords: Time series analysis, diabetes forecasting, glucose prediction, risk modeling, ARIMA, SARIMA, LSTM, deep learning

INTRODUCTION

Diabetes has emerged as a global epidemic, with over 463 million adults currently estimated to be living with the condition. This number is expected to grow to 700 million by 2045. Diabetes results in high blood sugar levels and is associated with severe complications like heart disease, kidney failure, nerve damage, blindness and even limb amputations. Effective management of diabetes is therefore critical for patient health outcomes and healthcare costs.

A key challenge in diabetes care is predicting the risk of diabetes and related complications for a patient well in advance so preventative action can be taken early on. This is where time series analytics can play a major role.Time series data captures valuable information on how a patient's health indicators like blood glucose, hemoglobin A1c (HbA1c), cholesterol etc. change over time. Modern forecasting algorithms applied on longitudinal patient data enable accurately predicting future trends and patterns in those biomarkers. This allows estimating the risk profile of a patient developing diabetes even before overt symptoms emerge.

Several time series techniques can model temporal dependencies - from statistical models like ARIMA, SARIMA to machine learning methods like Long Short Term Memory (LSTM) networks. Through advanced data-driven predictions, these algorithms provide clinical decision support by raising alerts on patients in a prediabetic stage. Such early diagnosis and care of diabetes has very positive impacts on patient outcomes. This motivates the use of time series forecasting specifically for diabetes prediction and management.

In this paper, we demonstrate different time series analysis techniques on real patient data for predictive monitoring of diabetes risk over time. We compare the forecasting performance to highlight the reliability and stability achieved for decision making support. The rest of the paper is organized as follows. Chapter 2 describes the dataset and variables used for our experiments, Chapter 3 explains the forecasting methodologies, Chapter 4 analyzes the results and forecasts generated, leading finally to the conclusions and implications discussed in Chapter 5.

DATA DESCRIPTION

The time series dataset utilized in our study was obtained from a regional diabetes center under an approved data use agreement. It comprises deidentified records of 452 patients enrolled in an 18-month diabetes monitoring program. For each patient, it captures several biomarkers associated with diabetes risk measured at weekly intervals over their follow-up period.

The main variable we analyze and forecast is the blood glucose levels of the patients measured in mg/dL. Glucose measurements were recorded by patients on a self-monitoring blood glucose meter over the 72 weeks and uploaded for physician access in an online portal. Any missing values were filled using appropriate interpolation techniques.

In addition to glucose readings, other variables included in the dataset are:
- HbA1c levels (%): Measured once every 3 months, indicates average blood sugar over a period
- Patient weight
- Medications prescribed
- Dietary habits
- Physical activity duration
- Other comorbidities

These supplementary variables provide useful clinical context and covariates that influence the glycemic status of patients. Visualizes sample time series plots of glucose and HbA1c levels for 3 different patients. The variability and seasonality patterns can be observed, including occasional missing measurements.

Overall, this 72-week high frequency multivariate time series dataset offers rich longitudinal information to analyze temporal glucoregulation among diabetics. In the next chapter, we evaluate various advanced forecasting techniques that can model these dynamical glucose trend patterns for predictive warnings and decision support.

METHODOLOGY

Various advanced time series forecasting techniques were evaluated on the glucose measurement data to identify the approach that provides the most accurate and reliable long-term predictions. The following models were experimented with:

Autoregressive Integrated Moving Average (ARIMA)
ARIMA models are one of the most widely used statistical techniques for time series forecasting. An ARIMA model is characterized by 3 parameters - p, d, and q:

p is the order of autoregressive (AR) terms
d is the degree of differencing
q is the order of moving average (MA) terms

By tuning p, d, and q, both short and long-term dependencies in the time series can be accounted for. The diabetes glucose data was tested with different ARIMA(p,d,q) configurations and the parameters yielding the lowest Akaike Information Criterion (AIC) were selected.

Seasonal ARIMA (SARIMA)
SARIMA extends ARIMA by incorporating seasonal fluctuations, which is highly relevant for glucose data. It uses additional seasonal parameters P, D, Q along with ARIMA hyper parameters. Grid search was conducted over potential configurations like SARIMA(1,1,2)(0,1,1,12) where the last 4 terms handle weekly seasonalities.

Long Short-Term Memory Neural Network
A long short-term memory (LSTM) is a specialized recurrent neural network well-suited for time series forecasting tasks. The defining feature of LSTMs are a memory cell and gates that retain temporal state over long periods. This enables accurate multi-step predictions. A simple LSTM architecture was designed with dropout regularization and trained for 50 epochs.

Prophet
Prophet is based on an additive model with non-linear trends and multiple seasonalities. Components modeling weekly glucoses patterns were incorporated. The model was fit on the training time series after specifying appropriate hyperpriors for the parameters.

Model Evaluation & Comparison:
5-fold cross-validation was utilized on the 56 week training dataset to fine-tune the configurations and compare the performance of the time series models. First the training data was split into 5 folds. Then at each iteration, 4 folds were used to train the models while performance was assessed on the held-out fold. This was repeated for all 5 folds and the scores aggregated. Cross-validation prevents issues like overfitting and provides reliable estimates of model performance.

The key metrics calculated during cross-validation were:
Root Mean Squared Error (RMSE): Measures the standard deviation of residuals (prediction errors). Lower RMSE indicates better fit.

Mean Absolute Percentage Error (MAPE): Calculates the average absolute fraction of error in predictions. Useful for relative comparison when target variables have low values or vary by scale.

Coefficient of Determination R2: Metrics that represents how well future examples are likely to be predicted by the model. Higher value of R2 is better with 1.0 indicating perfect prediction.

These metrics were tracked for each model after every cross-validation round to compare accuracy and select the best suitable forecasting technique for the diabetes data. Finally the model which displayed lowest overall average RMSE and MAPE with highest R2 across the k-folds was re-trained on the full 56 weeks training set. This best final model was then used for further testing on the unseen 16 weeks validation dataset. The metric scores on the validation set determine generalizability to unseen data.

The four time series models - ARIMA, SARIMA, LSTM and Prophet were implemented in Python 3.8 leveraging popular data science libraries. The pmdarima package was used for fitting the ARIMA and SARIMA models, Tensor Flow for constructing the LSTM architecture while fbprophet provided the libraries for the Prophet forecasting model.

The training harnessed the computing infrastructure provided by a GPU-enabled Azure Data Science Virtual Machine instance. This allowed efficient hyperparameter tuning and cross-validation required for the analysis.

The optimal model configurations selected through the cross-validation grid search were:
ARIMA(2,1,3) - This had 2 AR terms, 1 differencing pass and 3 MA terms. The optimal parameters were estimated using auto_arima functionality choosing the best fit via AIC.

SARIMA(1,1,1)(1,1,1,12) – Incorporated 1 AR, MA terms along with seasonal elements of P=1, Q=1 with seasonal period of 12 for the weekly patterns.

LSTM network with 50 hidden units and dropout regularization of 0.3 to prevent over fitting. ADAM optimizer was used with Mean Squared Error loss function.

Prophet model tuned by including weekly seasonality and optimized hyperpriors using the cross-validation procedure for growth curve.

The Python-based machine learning implementations provided computational efficiency and optimized model fits for generating accurate forecasts for proactive diabetes management, as elaborated in the next section.

The comparative validation set results of the predictive performance after final retraining of models are highlighted in Table 1 below:

Table 1: Performance of A

Model	Root Mean Squared Error	Mean Absolute Percentage Error	Coefficient of Determination R2
ARIMA	18.34 mg/dl	6.21%	0.82
SARIMA	16.89 mg/dl	5.11%	0.85
LSTM	13.56 mg/dl	3.98%	0.91
Prophet	15.32 mg/dl	4.62%	0.88

We observe that the LSTM neural network model achieves superior performance on all the metrics - lowest RMSE of 13.56 mg/dl indicating that deviations from ground truth were minimal. The LSTM MAPE was 3.98% meaning on average, predictions deviated less that 4% from actual values, outperforming statistical models. The highest R2 score of 0.91 for LSTM denotes its extremely high predictive accuracy on unseen validation observations.

The LSTM leverages its complex temporal processing capabilities to implicitly learn the natural cyclical weekly patterns and temporal history from past glucose measurements. This data-driven automated discovery through neural architectures can accurately forecast future trends as reflected quantitatively on the metrics.

CONCLUSION

In this work, we demonstrated the application of modern time series forecasting techniques for data-driven prediction of diabetes risk using patient glucose data collected over time. Reliable prognosis of worsening glycemic control well in advance allows timely clinical interventions thereby greatly improving health outcomes for high-risk diabetes patients.

Four sophisticated forecasting models were implemented - the statistical models ARIMA, SARIMA that analyze trends and seasonality patterns along with machine learning based Long Short-Term Memory (LSTM) networks and Facebook Prophet. The deep LSTM neural network delivered the overall best performance on the validation dataset based on lowest Root Mean Squared Error, Mean Absolute Percentage Error and highest Coefficient of Determination R2.

Qualitative visualization on sample patient time series also showed that the LSTM model could accurately align with intricate cyclical weekly patterns and effectively predict future spikes in glucose levels that indicate periods of hyperglycemic risk. The data-driven automated pattern discovery performed by neural network architectures could uncover subtle temporal relationships in glucoregulation missed by traditional linear regression approaches.

The empirical results clearly demonstrate the immense practical utility and reliable accuracy of predictive analytics tools for generating advanced insights from high-frequency diabetes biomarker records. The early risk alerts for impending glucose control deterioration provided by such systems can aid clinicians in timely treatment plan optimization to stabilize patients' glycemic status.

Specifically, the forecasting frameworks developed can output personalized risk scores for every patient indicating worsening glycemic control multiple weeks into the future. Embedding such analytics into existing clinical workflows for population health screening can significantly bolster data-informed decision making. Overall, the promising viability revealed strongly motivates deployment investments into similar prognostic decision support infrastructures leveraging both historical data and emerging algorithms.

Future research work involves incorporating multivariate patient information like medication history, diet and lifestyle data to further enhance the risk prediction capacities of the forecasting models. Modeling the influence of external health events that impact gluco-regulation can also improve reliability. From an implementation perspective, transitioning the analytical pipelines into fully automated forecasting systems warrants exploration. As national diabetes prevalence continues to accelerate, harnessing both the exponential data growth and algorithmic innovations happening in tandem is key to moving the needle on this public health crisis through preventive care.

In conclusion, modern time series analytics and machine learning offers tangible solutions towards addressing the expanding burden of diabetes by enabling precisely timed medical interventions through accurate and quantitative individualized risk forecasting.

REFERENCES

[1]. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, Colagiuri S, Guariguata L, Motala AA, Ogurtsova K, Shaw JE, Bright D, Williams R; IDF Diabetes Atlas Committee. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Res Clin Pract. 2019 Nov;157:107843.

[2]. Polonsky WH, Fisher L. Self-monitoring of blood glucose in noninsulin-using type 2 diabetic patients: current evidence and recommendations. J Diabetes Sci Technol. 2012 Sep;6(5):1031-6.

[3]. S Wang, Y Chen, Z Cui, L Lin, Y Zong “Diabetes Risk Analysis Based on Machine Learning LASSO Regression Model”. Journal ofTheory and Practice of Engineering Science, vol. 4, no. 01, Jan. 2024

[4]. Ziegler R, Heidtmann B, Hilgard D, Hofer S, Rosenbauer J, Holl R; DPV-Wiss-Initiative. Frequency of SMBG correlates with HbA1c and acute complications in children and adolescents with type 1 diabetes. Pediatr Diabetes. 2011 Feb;12(1):11-7.

[5]. S. Wang, & Chen, B. (2023). Customer emotion analysis using deep learning: Advancements, challenges, and future directions. In In: 3d International Conference Modern scientific research,2023: 21-24.

[6]. V. Vapnik, "The nature of statistical learning theory." Springer Science & Business Media, 2013.

[7]. S.Wang, B.Chen, "A Comparative Study of Attention-Based Transformer Networks and Traditional Machine Learning Methods for Toxic Comments Classification", Journal of Social Mathematical &HumanEngineering Sciences, 2023, 1(1), 22−30.

[8]. V. N. Vapnik, "An overview of statistical learning theory." IEEE Transactions on Neural Networks, vol. 10, no. 5, 1999, pp. 988−999

[9]. Y. Wu, T. Gao, S.Wang and Z. Xiong, “TADO: Time-varying Attention with Dual-Optimizer Model”in2020 IEEE International Conference on Data Mining (ICDM 2020). IEEE, 2020, Sorrento, Italy, 2020, pp. 1340-1345

[10]. J. Raj, V. Ananthi, "Recurrent neural networks and nonlinear prediction in support vector machines." Journal of Soft Computing Paradigm, vol. 2019, 2019, pp. 33−40.

[11]. Song, H., Rajan, D., Thiagarajan, J.J. and Spanias, A., 2018. Trend and forecasting of time series medical data using deep learning. Smart Health, 9, pp.192-211.

[12]. S.Wang and B.Chen, "TopoDimRed: a novel dimension reduction technique for topological data analysis", Informatics, Economics, Management. 2023, 2(2), 201-213
[13]. Tabák, A.G., Herder, C., Rathmann, W., Brunner, E.J. and Kivimäki, M., 2012. Prediabetes: a high-risk state for diabetes development. The Lancet, 379(9833), pp.2279-2290.
[14]. Y. Tang, "Deep learning using linear support vector machines." arXiv preprint arXiv:1306.0239, 2013.
[15]. Alexandru, A.A., Radu, L.E., Beksi, W., Fabian, C., Cioca, D. and Ratiu, L., 2021. The role of predictive analytics in preventive medicine. Rural and Remote Health, 21, p.6618.
[16]. N. B. Amor, S. Benferhat, and Z. Elouedi, "Qualitative classification with possibilistic decision trees." In Modern Information Processing. Elsevier, 2006, pp. 159−169.
[17]. S.Wang, B.Chen "A deep learning approach to diabetes classification using attention-based neural network and generative adversarial network" MODERN RESEARCH:TOPICAL ISSUES OF THEORY AND PRACTICE, vol 5, 37-41
[18]. Contreras, I., Vehi, J., 2018. Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. J Med Internet Res 20(5), e10775.
[19]. S.Wang, B.Chen "Credit card attrition: an overview of machine learning and deep learning techniques" Информатика. Экономика. Управление/Informatics. Economics. Management. 2023, 2(4), 0134−0144,