Diabetes Prevalence Forecasting Using GARCH and Deep Learning Models

Aida Abdulhssein Mohammed[1], Maytham Mohammad Bakr[2] and Rdab Tark Abdalla[2]
[1]*Open Educational College - Balad, Ministry of Education, 34001 Balad, Iraq.*
[2]*Department of mathematics, College of Education, University of Samarra, 34010 Salah-Al Din, Iraq*
ayeda.mohmmed@st.tu.edu.iq, Maitham.m@uosamarra.edu.iq, rdab.abdalla@st.tu.edu.iq

Keywords: Type 1 Diabetes, GARCH, LSTM, Time Series, Forecasting, Statistical Stability.

Abstract: The aim of this study is to use cutting-edge statistical and artificial intelligence techniques to assess and forecast the prevalence of type 1 diabetes in Iraqi children. In order to determine the prospective ultimate caseload and the fictitious inflection point, the research used the generalised autoregressive conditional variance (GARCH(p,q)) model to estimate the temporal behaviour of the data. The models under test were subjected to statistical stability conditions, and the AIC and BIC information criteria were used to assess their performance. The findings indicated that, out of all the nonparametric models examined, the GARCH(1,0) model performed the best. However, the paper showed that there were no significant variance fluctuations in the incidence data, which restricts the use of strictly nonparametric models. Deep LSTMNs neural networks were used for time series analysis in order to get around this. Using the ideal parameters for the number of layers and iterations, the data was divided into an 80% training set and a 20% test set. In comparison to other models, the LSTMNs findings showed the lowest mean squared errors, indicating a good capacity to predict future values. This implies that an efficient method for tracking upcoming changes in chronic illnesses is to combine deep learning algorithms with traditional statistical models. To enhance planning and response to the epidemic load of chronic illnesses, the paper suggests using this strategy in medical modelling and health forecasting.

1 INTRODUCTION

One of the most common chronic metabolic diseases affecting kids and teenagers is type 1 diabetes. Chronically high blood glucose levels are the result of the autoimmune death of the beta cells in the pancreas that secrete insulin [1]. The condition is severe because it directly affects the body's metabolic equilibrium, affecting numerous essential processes such as protein, lipid, and glucose metabolism, as well as oxidative stress and inflammatory pathways [2].

Laboratory indicators like fasting glucose, glycated hemoglobin (HbA1c), insulin levels, C-peptide, and inflammatory markers like CRP, triglycerides, and cholesterol are crucial for comprehending the course of a disease and evaluating the efficacy of treatment [3].

Long-term tracking of these markers yields precise data on the physiological and metabolic alterations linked to the illness and aids in the prediction of further consequences, including microvascular and macrovascular disease [1], [2].

Advances in artificial intelligence and statistical analysis techniques have made it possible to use deep neural networks like long-short-term memory networks (LSTMNs), a type of recursive neural network designed to handle time-series data and address the vanishing gradient problem [4], as well as time series models like the generalized autoregressive conditional variance (GARCH) model, which was developed as an extension of the Autoregressive Conditional Heteroskedasticity (ARCH) model proposed by Robert F. Engle in the 1980s [5].

More opportunities for early intervention and better treatment plans are made possible by the development of predictive models that can recognize pivotal moments in the disease trajectory and forecast future biochemical markers thanks to the combination of statistical analysis and artificial intelligence [6].

The purpose of this paper is to forecast future trends and possible changes in biomarkers by analyzing biochemical data from pediatric type 1 diabetes patients in Iraq using a combination of GARCH models and LSTMNs approaches. In order to better understand disease dynamics and enhance healthcare practices, we want to offer a scientific framework that integrates biochemistry and sophisticated statistical analysis.

2 LITERATURE REVIEW

Over the last several decades, time series analysis has developed into a vital tool for researching health issues, especially long-term conditions like diabetes. In order to solve the issue of non-stationary variance in economic data, Engle proposed the ARCH model in 1982, making it one of the first statistical models in this area. It uses the squares of the previous mistakes to depict the conditional variance as a series [7].

The generalized ARCH model (GARCH), which Bollerslev expanded in 1986, is more adaptable when dealing with very volatile data because it combines ARCH components with other components to better correctly describe the conditional variance [5].

Subsequent iterations of this model, including TARCH [8], GJR-GARCH [9], and EGARCH [10], attempted to address certain facets of variance, including the leverage impact and volatility asymmetry. GARCH models have been used in the healthcare industry to analyze data on infectious and non-infectious disorders. For example, [11] used GARCH to Middle Eastern diabetes data and showed that it was successful in describing the disease's temporal patterns.

The analysis of temporal data has been transformed by artificial intelligence in tandem with classical statistical models. This is especially true since the development of LSTMNs by [12], which has demonstrated their effectiveness in processing serial data with long-term temporal correlations [4].

As demonstrated by a study by [13], which compared the effectiveness of LSTMs with conventional statistical models in predicting the prevalence of diabetes and discovered that LSTMNs outperformed them in accuracy and capacity to identify intricate patterns in data, LSTMNss have been widely used in the prediction of chronic diseases.

This research demonstrate that integrating contemporary deep learning methods like LSTMNs with statistical models like GARCH offers an integrated framework for evaluating medical time series, improving prediction accuracy, and assisting with data-driven health planning choices.

2.1 ARCH Model

ARCH statistical model is used to examine time series volatility and forecast future volatility. The ARCH model is used in finance to assess risk by offering a volatility framework that accurately reflects actual market conditions. Higher volatility follows times of high volatility, while lower volatility follows periods of low volatility, according to ARCH modeling [14], ARCH model, which will have the formula [15]:

$$y_t = \sigma_t \epsilon_t \text{ where } \epsilon_t \sim iid\ N(0,1)$$

$$\sigma_t^2 = \omega + \sum_{j=1}^{Q} \varphi_j\, y_{t-j}^2 \qquad (1)$$

Where σ_t^2 is the conditional variance, and $\omega, \sum_{j=1}^{Q} \varphi_j$ are the parameters of the model.

This model is predicated on the martingale difference [15],

$$E(Y_{t+1}^2 / F_t) = \sigma_t^2 \qquad (2)$$

Where F_t is a $\sigma - field$ of a random variables called a sometimes filter $(y_{t-1}, y_{t-2}, \dots, y_{t-Q})$ [8].

The large-scale conditional autoheteroscedasticity (GARCH(Q, P)) model was introduced by Bollerslev in 1986 as an extension of the ARCH model, aiming to address limitations related to time-varying conditional variance and provide a more flexible model for predicting time series volatility. This model is defined by the following system of equations [15]:

$$y_t = \sigma_t \epsilon_t \text{ where } \epsilon_t \sim iid\ N(0,1)$$

$$\sigma_t^2 = \omega + \sum_{j=1}^{Q} \varphi_j\, y_{t-j}^2 + \sum_{i=1}^{P} \gamma_i\, \sigma_{t-i}^2 \qquad (2)$$

Where ω is a constant and φ_j, γ_i a represent the ARCH and GARCH model parameters, respectively.

By imposing certain restrictions on the coefficients of the conditional variance equation, the value of ω must be > 0, to ensure that the variance remains positive and that negative values cannot occur, which maintains the mathematical consistency and validity of the statistical model, $\varphi_j \geq 0\ \forall j,\ j = 1,2, \dots, Q$، $\gamma_i \geq 0\ \forall i, i = 1,2, \dots, P$, a

$\sum_{j=1}^{Q} \varphi_j + \sum_{i=1}^{P} \gamma_i < 1$ then the conditional variance is:

$$\sigma_t^2 = \frac{\omega}{1 - (\sum_{j=1}^{Q} \varphi_j + \sum_{i=1}^{P} \gamma_i)} \quad (3)$$

Despite the merits of GARCH models, they exhibit shortcomings, as noted by Black in 1976, namely the inability to capture the Leverage effect [16].

For a GARCH model to maintain stationarity, the conditional variance must converge to the unconditional variance over time, so that the mathematical condition $\sigma_t \rightarrow \sigma_y$ at infinity is met, ensuring that the variance remains constant and does not deviate toward unstable values [17], [18].

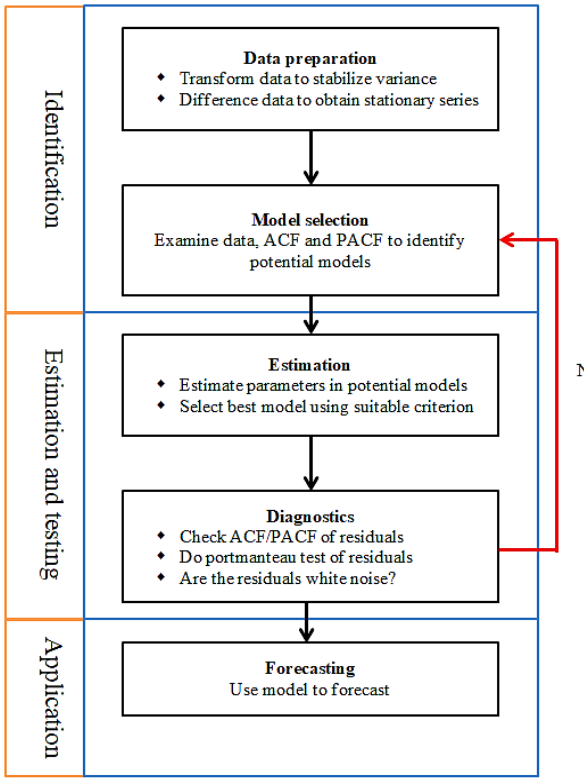


Figure 1: Box-Jenkins flow chart [19].

2.2 Box-Jenkins Method

Box and Jenkins developed this method in 1970 and subsequently underwent further developments between 1970 and 1996 to include a range of complementary techniques as needed. This methodology aims to achieve optimal predictive accuracy through three main stages [15]:

1) Identification. This includes data preparation and selecting the appropriate model for the time series under study.
2) Estimation and Testing. This involves estimating model parameters and performing diagnostics to ensure its suitability and validity for prediction.
3) Application. This is where the developed model is used to make the required predictions and make decisions based on them.

Figure 1 illustrates the various stages of this methodology graphically, making it easier to understand the sequence of processes and steps involved.

2.3 Model Order Selection Criteria

2.3.1 Akaike Information Criterion (AIC)

This model is considered the most widely used in statistical literature, and is represented by the following mathematical function [20]:

$$AIC(p,q) = -2\,log(maximum\ likelehood) + 2k \quad (5)$$

Where k is number of parameters in proposed model.

2.3.2 Bayesian Information Criterion (BIC)

This criteria was introduced similarly to the Akaike criterion to get enhanced convergence features and is expressed by the following function [18], [21]:

$$BIC = -2l(maximum\ likelehood) + k\,ln\,n \quad (6)$$

Where m represents the number of observations or measurements used in the analysis.

2.3.3 Long Short-Term Memory Networks (LSTMNs)

LSTMNs are effective tools for classifying, processing, and predicting time series data, especially when these series contain unspecified time intervals between important events [12].

These networks were developed to address the problem of vanishing gradients that hinders the learning of traditional neural networks. LSTMNs belong to the class of artificial neural networks (ANNs) used in the fields of artificial intelligence. Recurrent neural networks (RNNs) are characterized by their ability to process both individual data points and entire data series [13].

3 ANALYSIS AND RESULTS

3.1 Collection of Data

This study is based on 571 observations of approved diabetes cases, recorded between January 1, 2021, and July 25, 2022, taken from the Iraqi Ministry of Health official website. To achieve optimal predictive accuracy, the GARCH model's stability criteria will be applied to this data. This ensures that the model's conditional variance converges with its unconditional variance, thus achieving model stability and reliable future predictions based on time series data.

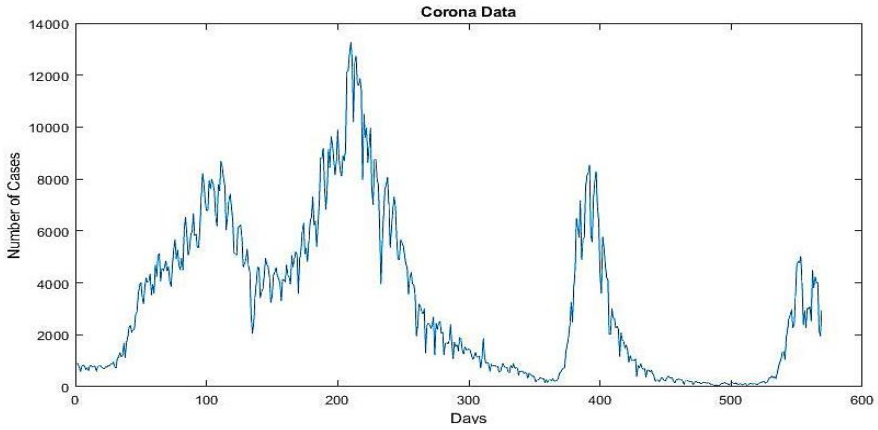
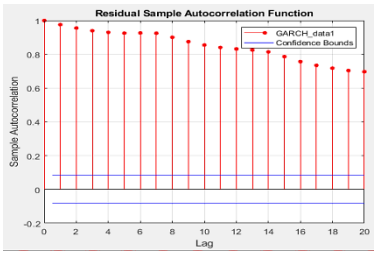
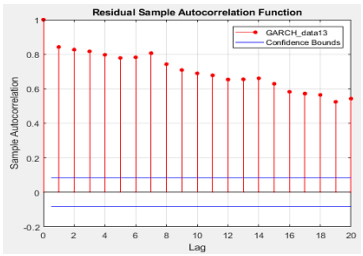


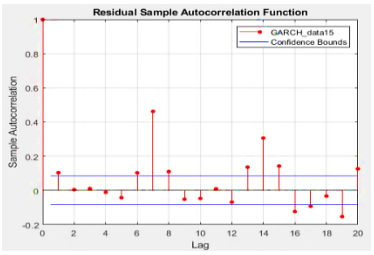
Figure 2: Historical statistics about diabetes from January 1, 2021, to July 25, 2022.



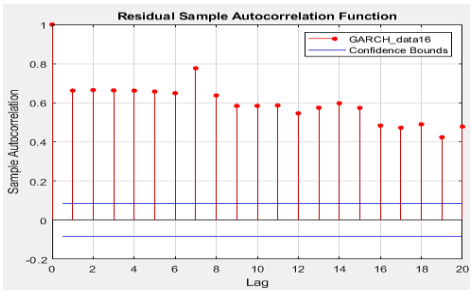
GARCH(0,0)



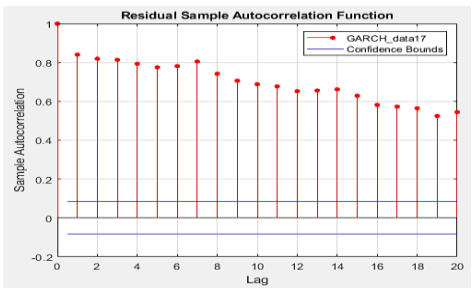
GARCH(1,1)



GARCH(1,2)

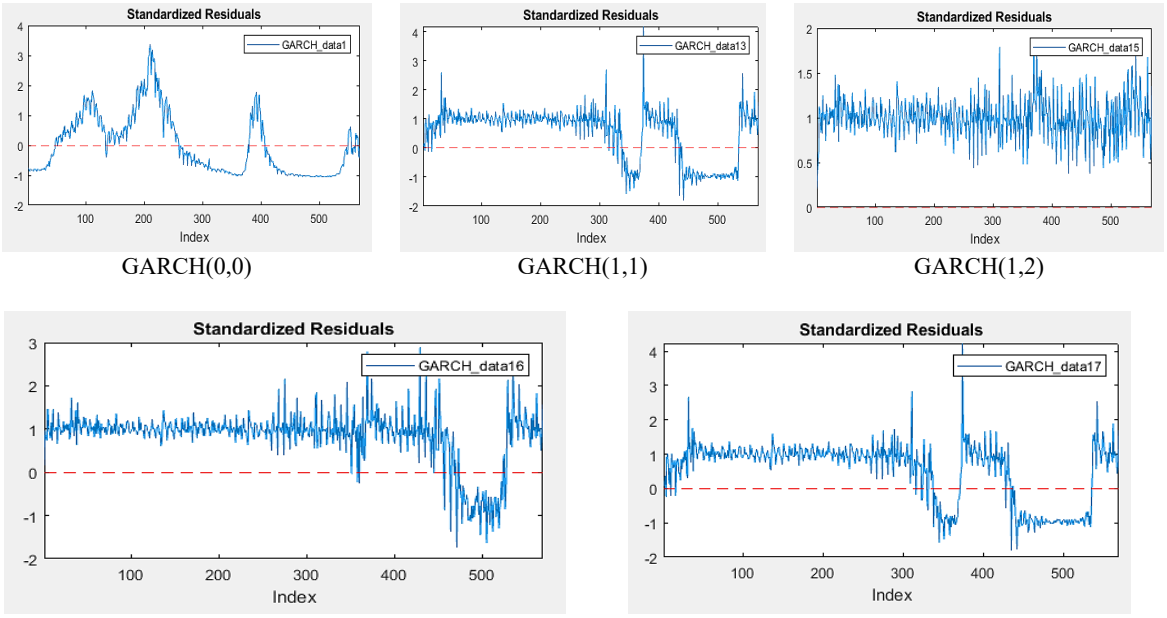


GARCH(0,1)



GARCH(0,2)

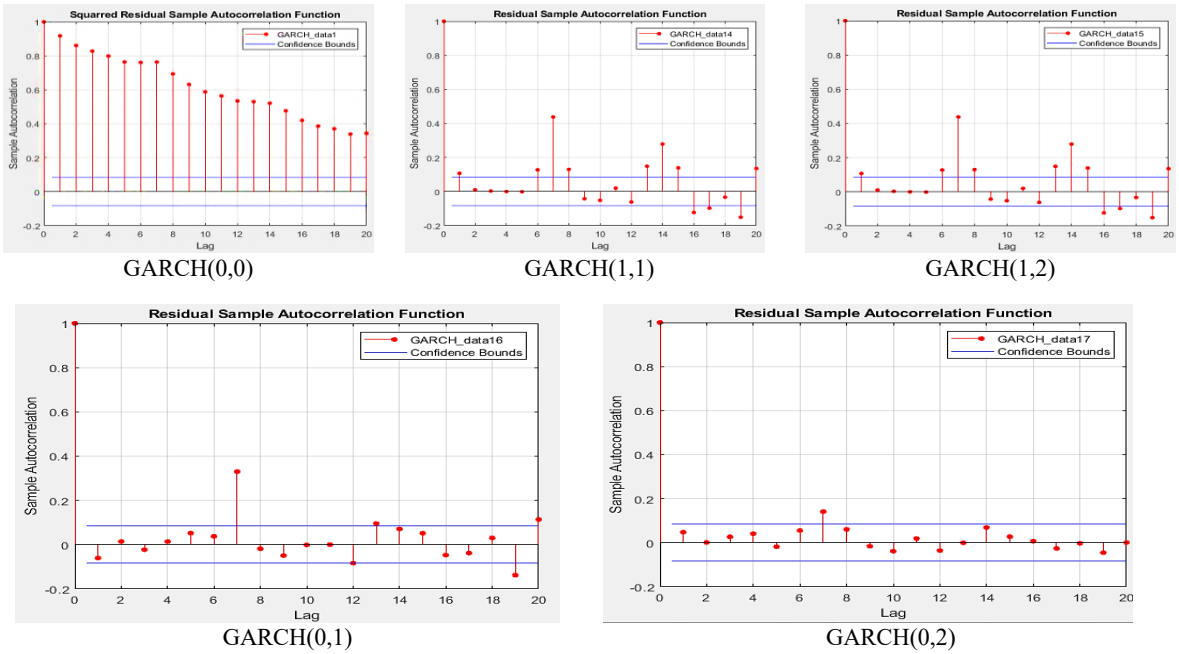
Figure 3: Autocorrelation functions with varying orders.



GARCH(0,0) GARCH(1,1) GARCH(1,2)

GARCH(0,1) GARCH(0,2)

Figure 4: Residual functions with varying orders.



GARCH(0,0) GARCH(1,1) GARCH(1,2)

GARCH(0,1) GARCH(0,2)

Figure 5: Autocorrelation functions of the GARCH model after Residual functions with different orders.

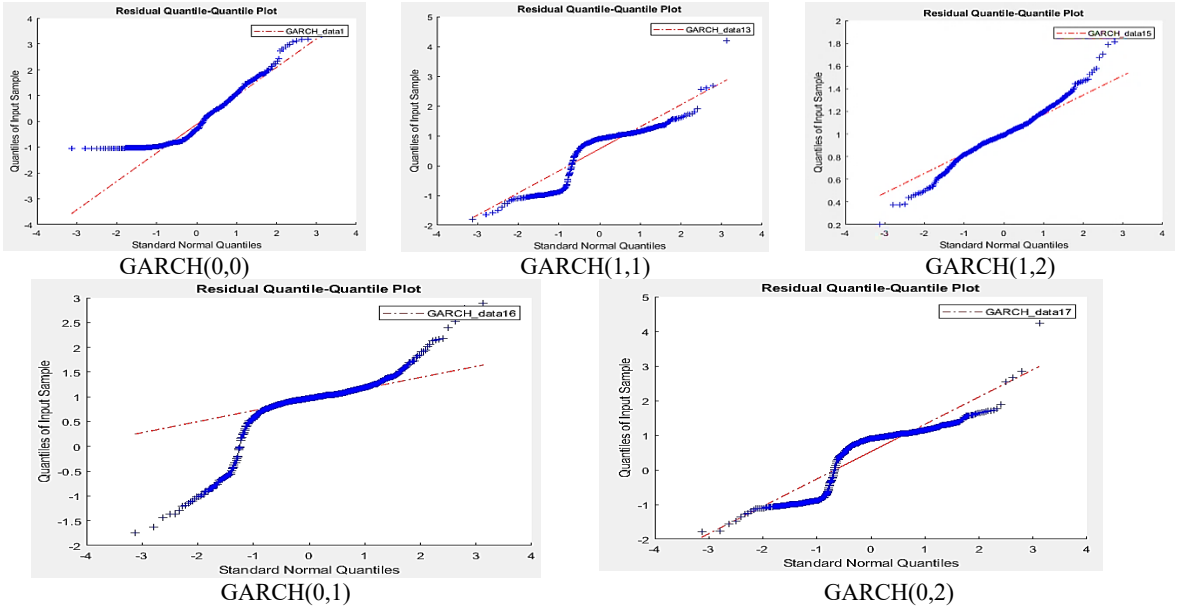


Figure 6: The QQ Plot functions with varying orders.

Table 1: The parameter values of models of varying rank derived from the GARCH model.

GARCH(P,Q)	Constant	ARCH(1)	ARCH(2)	GARCH(1)	GARCH(2)
GARCH(0,0)	$8.99e+06$	0	0	0	0
GARCH(1,1)	$1.708e+04$	0.829	0	0.171	0
GARCH(1,2)	$2.176e+04$	0.796	0.204	0	0
GARCH(0,1)	$3.049e+03$	1.000	0	0	0
GARCH(0,2)	$2.176e+04$	0.796	0.204	0	0

Table 2: The AIC and BIC values of various ranks from the GARCH model.

GARCH(P,Q)	AIC	BIC
GARCH(0,0)	1.073e+04	1.739e+04
GARCH(1,1)	9.899e+03	9.917e+03
GARCH(1,2)	1.009e+04	1.010e+04
GARCH(0,1)	9.848e+03	9.861e+03
GARCH(0,2)	9.898e+03	9.915e+03

3.2 Modeling and Analyzing of Data by GARCH Model

The GARCH model will be applied to the data, observing and confirming that the conditional variance converges to the unconditional variance, which is a prerequisite for ensuring model stability and reliable predictions. The process involves several sequential steps, beginning with identifying changes in the data, then building the model according to appropriate statistical criteria. This is followed by estimating the model parameters using appropriate estimation methods, then assessing the model's suitability through diagnostic tests, and finally forecasting the future conditional variance. Matlab R2020a was used to create and program the time series data, evaluating several models of different orders to find the optimal model that best fits the data.

The data visualization phase is the first and important step in time series analysis, helping to understand the data pattern and identify any potential trends or fluctuations. Figure 2 shows a time series graph displaying historical data on the number of diabetes cases from January 1, 2021, to July 25, 2022.

The next step involves converting the original series into a series of returns to facilitate volatility

analysis, and then converting these returns into a series of squared errors for conditional variance analysis. Autocorrelation functions (ACFs) for the GARCH model were also calculated and displayed at different levels. Figure 3 illustrates the ACFs for a set of models, which helps assess the correlation structure in the data and select the most appropriate model.

After completing this step, we transform the squared error series into a series of residuals by performing the Young-Box test to assess the effect of heteroscedasticity over multiple lags. Figure 4 shows several residual functions for a GARCH model across multiple orders, while Figure 5 displays the autocorrelation functions for these models.

We will now examine the quantile-quantile plot, or QQ Plot, which visually evaluates whether the remaining series adheres to a normal distribution, as seen in Figure 6.

In addition to the graphs presented previously, which illustrate the general trend of the series, the model's suitability for future predictions must be evaluated. Tables 1 and 2 will be presented: the first shows the estimated parameters for each of the proposed models, and the second shows the results of the AIC and BIC values of various ranks from the GARCH model for selecting the optimal model.

A multitude of coefficients with zero values indicates that some model components do not significantly contribute to variance explanation. The conditional variance remains almost constant throughout time, exhibiting little fluctuations, as shown by the GARCH(0,1) model's constant of 1.000. The inadequate time-dependent effect of data variance is shown by the relatively small ARCH and GARCH coefficients in the other models.The GARCH(0,1) model exhibited the optimal fit among the analyzed models, shown by the lowest AIC and BIC values. Despite minor differences in AIC and BIC, GARCH(0,1) is statistically superior than GARCH(1,1). Extremely high GARCH(0,0) scores indicate a suboptimal overall fit to the data.

Thus, the conditional variance of the examined models is shown in Figure 7.

Upon reviewing the preceding stages of data analysis illustrated in graphs and tables, it is evident that the most reliable model for forecasting is the GARCH(1,0) model. However, we must now verify the stability of the selected model by applying its stability conditions and assessing the conditional variance as delineated in (4).

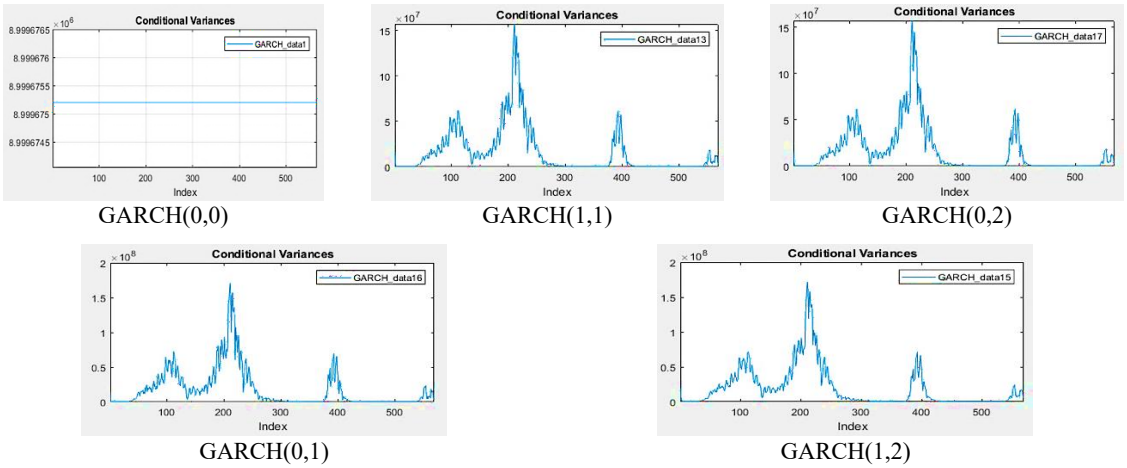
$$\varphi_1 = 1.000$$

After applying stability conditions to the GARCH model, it was confirmed that the model achieves the required stability, which is essential for ensuring the reliability and accuracy of future predictions. This property ensures that the model's conditional variance does not deviate toward unstable values over time but rather converges toward the unconditional variance. The value of the model's unconditional variance is defined as follows:

$$\sigma^2 = \frac{\omega}{1-(\varphi_1+\sum_{i=1}^{2}\gamma_i)} = \frac{3.049e{+}03}{1-1} = \text{unknown value}$$

Therefore, the equation for the GARCH(1,0) model, which is used to predict unconditional variance values, is as follows:

$$y_t = \sigma_t \epsilon_t \text{ where } \epsilon_t \sim iid\ N(0,1)$$
$$\sigma_t^2 = \omega + \varphi_1 y_{t-1}^2 \qquad (7)$$



GARCH(0,0) GARCH(1,1) GARCH(0,2)

GARCH(0,1) GARCH(1,2)

Figure 7: The conditional variance of the GARCH model with varying orders.

3.3 Modeling and Creating by Using LSTM

LSTM, another method for modeling and analyzing time series, will be used in this paper.

The data is segmented into two groups: the training group, including 80% of the observations, and the test group, comprising 20% of the data. Given that the maximum number of hidden layers is 50 and the input consists of a single variable with one lag, after conducting 250 iterations, the minimal mean squared error was 0.0359.

Figure 8 represents the values of the training set, and the Figure 9 illustrates the predicted values of the test set.

And that the value:

$$AIC = 1.1915e + 03$$
$$BIC = 1.1956e + 03$$

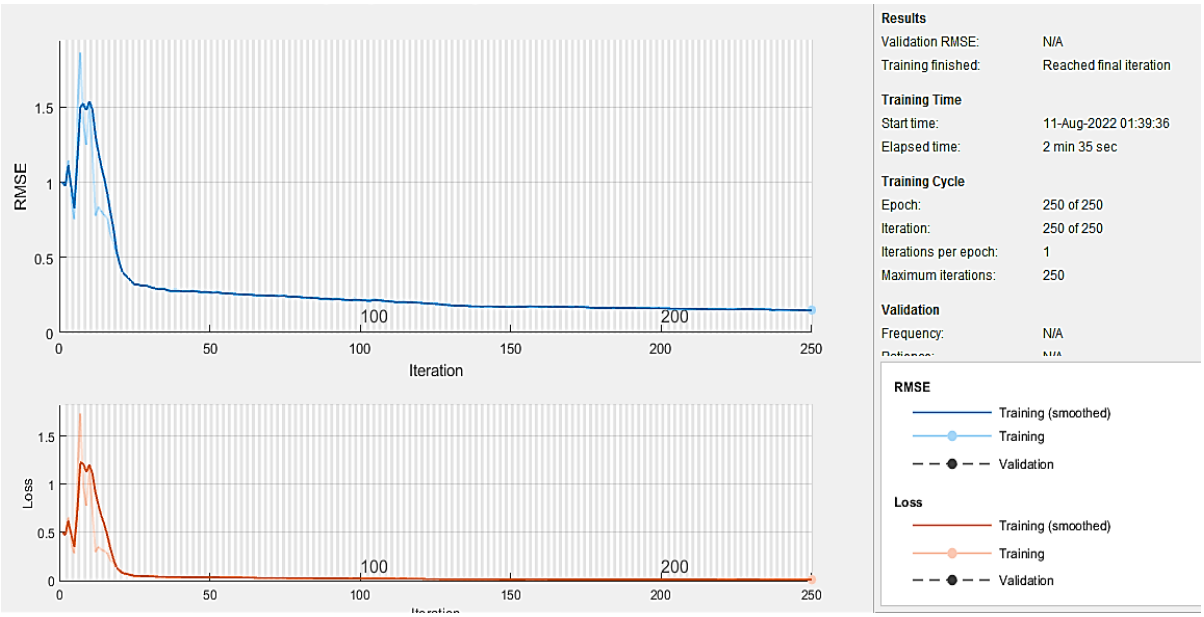


Figure 8: Training progress by using LSTM.

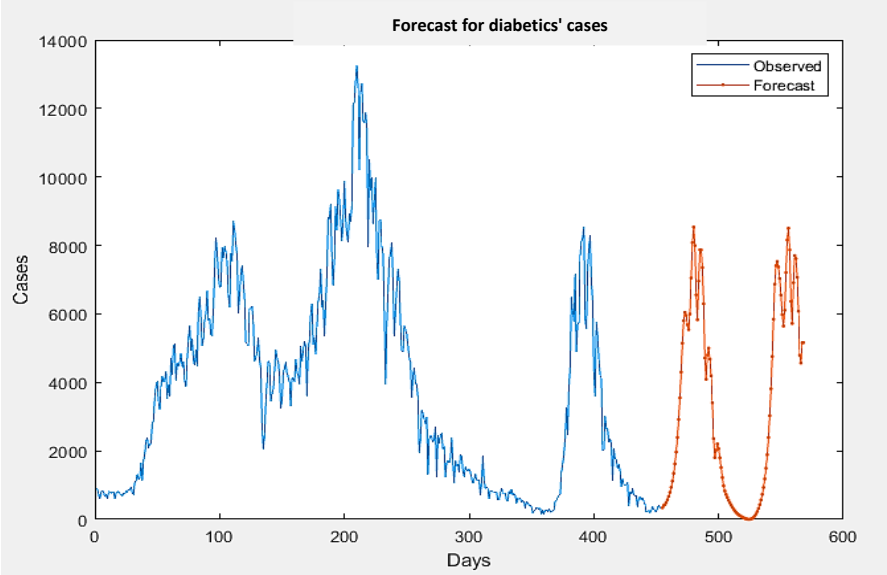


Figure 9: Predicted values for the test set by using LSTM.

Table 3: The estimated values for the training set by using LSTM.

Month	1st week	2nd week	3rd week	4th week
1st month	1.0e+04	0.0638	0.0864	0.0868
2nd month	0.0928	0.1093	0.1113	0.1097
3rd month	0.0645	0.0697	0.0827	0.0876
4th month	0.0556	0.0717	0.0855	0.0891
5th month	0.0770	0.0906	0.0990	0.1114
6th month	0.1142	0.1520	0.1662	0.1869
7th month	0.2191	0.2413	0.2816	0.3214
8th month	0.3410	0.3272	0.3688	0.4079
9th month	0.3670	0.3366	0.3994	0.4174
10th month	0.4275	0.3911	0.4518	0.4843
11th month	0.4218	0.4151	0.4489	0.5073
12th month	0.4666	0.4646	0.5137	0.5586
13th month	0.5359	0.4910	0.5548	0.6284
14th month	0.5397	0.5496	0.6273	0.7160
15th month	0.6388	0.6654	0.7732	0.8326
16th month	0.7046	0.6857	0.7511	0.8121
17th month	0.7513	0.6125	0.6797	0.7346
18th month	0.5528	0.4739	0.5544	0.5809
19th month	0.4740	0.4244	0.4997	0.5175

The numbers are the LSTM model's expected estimations for the training set. From around month 1 to month 15, there is a noticeable increasing trend that suggests a slow rise in the anticipated number of infections. The numbers start to vary and progressively decrease after month 16, demonstrating the model's capacity to identify shifts in trends (Table 3). There is little variation across weeks within a month, which is in line with the overall conclusion that the data is not very volatile.

4 CONCLUSIONS

Since there were few large variance swings in the daily data on confirmed children diabetes cases in Iraq from January 1, 2021, to July 25, 2022, using conventional GARCH models was less successful.

When statistical models were compared, GARCH(0,1) was determined to be the best model in terms of AIC and BIC. However, the data's low conditional variance makes it less reliable for making predictions in the future.

After using GARCH models, certain correlations were still present in the data, according to graphical analysis of residuals and autocorrelation functions (ACFs). This suggests that the models were not able to fully capture the temporal patterns in the series.

Quantile-Quantile (QQ) plots demonstrated that the GARCH models' residuals do not have a perfectly normal distribution, which has an impact on prediction accuracy. With a low mean square error (MSE) of 0.0359 and projected values that were very near to the actual values in the test data, the LSTM network's findings showed a definite advantage over GARCH models in terms of prediction accuracy.

In this instance, LSTM is the best model for short-term forecasts of future pediatric diabetes cases due to its capacity to adjust to temporal patterns, even in the absence of abrupt swings.

Instead of restricting the use of GARCH models to data with significant volatility, the research suggests using AI-based models, such LSTM, to analyze health time series with relative stability of variance.

REFERENCES

[1] A. M. Nour ElDin Abd ElBaky, et al., “Role of epicardial fat thickness and irisin levels in early prediction of cardiac dysfunction in children and adolescents with type 1 diabetes mellitus,” Pediatria Polska - Polish Journal of Paediatrics, vol. 98, no. 4, pp. 278-284, 2023, [Online]. Available: https://doi.org/10.5114/polp.2023.133530.

[2] N. Lubasinski, et al., “Blood Glucose Prediction from Nutrition Analytics in Type 1 Diabetes: A Review,” Nutrients, vol. 16, no. 14, p. 2214, Jul. 2024, [Online]. Available: https://doi.org/10.3390/nu16142214.

[3] K. Pang, "A comparative study of explainable machine learning models with Shapley values for diabetes prediction," Healthcare Analytics, vol. 7, 2025, [Online]. Available: https://doi.org/10.1016/j.health.2025.100390.

[4] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A Novel Connectionist System for Unconstrained Handwriting Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 5, pp. 855-868, May 2009, [Online]. Available: https://doi.org/10.1109/TPAMI.2008.137.

[5] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," J. Econometrics, vol. 31, no. 3, pp. 307-327, 1986, [Online]. Available: https://doi.org/10.1016/0304-4076(86)90063-1.

[6] A. Abuyassin and I. Laher, "Diabetes epidemic sweeping the Arab world," World J. Diabetes, vol. 7, no. 8, pp. 165-174, 2016, [Online]. Available: https://doi.org/10.4239/wjd.v7.i8.165.

[7] R. F. Engle, "Autoregressive Conditional Heteroscedasticity with Estimates variance of United Kingdom Inflation," J. Econometrica, vol. 50, no. 4, pp. 987-1008, 1982.

[8] C. Francq and J. M. Zakoian, GARCH Models: Structure, Statistical Inference and Financial Applications, 2nd ed., John Wiley & Sons, Ltd, 2019, pp. 1-14, [Online]. Available: https://doi.org/10.1002/9781119313472.ch1.

[9] L. R. Glosten, R. Jagannathan, and D. E. Runkle, "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks," J. Finance, vol. 48, no. 5, pp. 1779-1801, 1993.

[10] D. B. Nelson, "Conditional Heteroskedasticity in Asset Returns: A New Approach," Econometrica, vol. 59, no. 2, pp. 347-370, 1991, [Online]. Available: https://doi.org/10.2307/2938260.

[11] I. M. El-Kebbi, et al., "Epidemiology of type 2 diabetes in the Middle East and North Africa: Challenges and call for action," World J. Diabetes, vol. 12, no. 9, pp. 1401-1425, 2021, [Online]. Available: https://doi.org/10.4239/wjd.v12.i9.1401.

[12] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997, [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735.

[13] J. M. Ahn, J. Kim, and K. Kim, "Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting," Toxins, vol. 15, no. 10, 2023, [Online]. Available: https://doi.org/10.3390/toxins15100608.

[14] R. F. Engle III, "Risk and Volatility: Econometric Models and Financial Practice," Nobel Lecture, pp. 326-349, Dec. 2003.

[15] A. A. Mohammad and N. S. K. Aljboori, "Employment of GARCH Model and L.S.M.E Method in Time Series with Application to COVID-19 Virus," J. Algebraic Stat., vol. 13, no. 3, pp. 4856-4867, 2022, [Online]. Available: https://publishoa.com/index.php/journal/article/view/1330.

[16] E. M. Epaphra, "Modeling Exchange Rate Volatility: Application of the GARCH and EGARCH Models," J. Math. Finance, vol. 7, pp. 121-143, 2017, [Online]. Available: https://doi.org/10.4236/jmf.2017.71007.

[17] A. Mohammad and A. J. Salim, "The Analysis and Modeling of the time series of annual mean temperature in Mosul City," Rafidain J. Sci., vol. 7, no. 1, pp. 37-48, 1996.

[18] S. Lee, C. K. Kim, and D. Kim, "Monitoring Volatility Change for Time Series Based on Support Vector Regression," Entropy, vol. 22, no. 11, 2020, [Online]. Available: https://doi.org/10.3390/e22111312.

[19] S. Makridakis, S. C. Wheelwright, and R. J. Hyndman, Forecasting Methods and Applications, 3rd ed., John Wiley & Sons, Inc., 1998.

[20] H. Akaike, "A new look at the statistical model identification," IEEE Trans. Autom. Control, vol. 19, no. 6, pp. 716-723, Dec. 1974, [Online]. Available: https://doi.org/10.1109/TAC.1974.1100705.

[21] J. Fan and Q. Yao, Nonlinear Time Series: Nonparametric and Parametric Methods, Springer-Verlag New York, Inc., 2003.