# Predictive Analysis for Journal Abstracts using Polynomial Neural Networks Algorithm

**Adebola K. Ojo**

**Department of Computer Science**
**University of Ibadan**
**Ibadan, Nigeria**

**Abstract**

Academic journals are an important outlet for dissemination of academic research. In this study, Neural Networks model was used in the prediction of abstracts from The Institute of Electrical and Electronics Engineers (IEEE) Transactions on Computers. Simulation of results was done using the Polynomial Neural Networks algorithm. This algorithm, which is based on Group Method of Data Handling (GMDH) method, utilizes a class of polynomials such as linear, quadratic and modified quadratic. The prediction was done for a period of twenty-four months using a predictive model of three layers and two coefficients. The performance measures used in this study were mean square errors, mean absolute error and root mean square error.

***Keywords:*** *Polynomial Neural Networks, IEEE, GMDH, mean square errors, mean absolute error, Root mean square error*

## 1. Introduction

The Institute of Electrical and Electronics Engineers (IEEE) Transactions on Computers is a highly rated journal and a monthly publication with a wide distribution to researchers, developers, technical managers, and educators in the computer field. It publishes papers on research in areas of current interest to the readers. These areas include, but are not limited to, the following: computer organizations and architectures; operating systems, software systems, and communication protocols; real-time systems and embedded systems; digital devices, computer components, and interconnection networks; specification, design, prototyping, and testing methods and tools; performance, fault tolerance, reliability, security, and testability; case studies and experimental and theoretical evaluations; and new and important applications and trends.

Abstracts provide a comprehensive yet concise overview of an article [1] [2]. Abstracts are much shorter than the full text, which minimises the influence of data noise. Therefore, this study focused on prediction of abstracts of a journal article. These data were retrieved from the Institute of Electrical and Electronics Engineers (IEEE) Transactions on Computers, a monthly publication with a wide distribution to researchers, developers, technical managers, and educators in the computer field. It publishes papers on research in areas of current interest to the readers. Journal of Institute of Electrical and Electronics Engineers (IEEE) Transactions on Computers was chosen because it is one of the highly rated Computer Science journals with ISI indexed ranking and impact factors [3] [4] [5].

In the recent times and from the data gathered, the number of abstracts has been increasing. The indication of this is that new areas in the field of Computer Science are emerging. Furthermore, it shows how popular the IEEE journal has been. Predicting the number of abstracts will guide the management of IEEE on making strategic decision in the preparation of this journal articles. In this study, we shall employ neural network model for the prediction because of its prediction accuracy and the speed of its convergence.

## 2. Literature Review

A neural network is an interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on the animal neuron. The processing ability of the network is stored in the interunit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns [6].

Neural Networks (NN) is information-processing models inspired by the way the densely interconnected, parallel structure of the brain processes information. In other words, neural networks are simplified mathematical models of biological neural networks [7]. It can be viewed as weighted directed graphs in which artificial neurons are nodes and directed edges (with weights) are connections between neuron outputs and neuron inputs. Based on the connection pattern (architecture), ANNs can be grouped into two categories: Feed-forward networks (in which graphs have no loops, and Recurrent/Feedback) networks (in which loops occur because of feedback connections). In the most common family of feed-forward networks, called multilayer perceptron, neurons are organized into layers that have unidirectional connections between them [8].

ANN are used for statistical analysis and data modelling, in which their role is perceived as an alternative to standard nonlinear regression or cluster analysis techniques [9]. Thus, they are typically used in problems that may be couched in terms in terms of classification, or forecasting.

Artificial neural networks are forecasting methods that are based on simple mathematical models of the brain. They allow complex nonlinear relationships between the response variable and its predictors.

Polynomial Neural Network (PNN), a GMDH-type algorithm, is one of the useful methods for modelling nonlinear systems. It has architecture like feed forward neural networks, whose neurons are replaced by polynomial nodes. The output of the each node in a PNN structure is obtained by using several types of high-order polynomials such as linear, quadratic, and modified quadratic of the input variables. These polynomials are referred to as Partial Descriptions (PDs) [10]. This employs combinatorial algorithm for optimization of neuron connection. The algorithm iteratively creates layers of neurons with two or more inputs. Every new layer is created using two or more neurons taken from any previous layers [11].

There have been various research studies on predictions. Some of these include weather forecasting, market predictions [12], stock market price [13] [14] [15].

However, little work has been done on predictions of abstracts of academic publications of a journal.

## 3. Methodology

In this section, the various steps in achieving our results were discussed. This involved the data collection or gathering, prediction model formation and, predictive analysis and results. This is depicted in Figure 1. Each of these prediction processing steps is discussed in the following section.

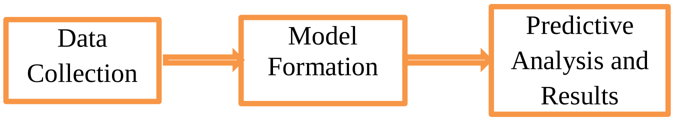

Figure 1: The Prediction Process

### 3.1 Data Collection

The dataset used in this study consisted of the number of abstracts of the academic journal (Journal of Institute of Electrical and Electronics Engineers Transactions on Computers.), between 1997 and 2017 data set. The data were extracted based on the number of abstracts in each issue of a volume.

### 3.2 Model Formation

Given a set of $n$ samples, $\{y(t_1), y(t_2), \ldots, y(t_n)\}$ in a time sequence, $t_1, t_2, \ldots, t_n$, the task is to predict the sample $y(t_{n+1})$ at some future time $t_{n+1}$. Prediction (or forecasting) has a significant impact on decision-making in business, science and engineering.

A neural network can be thought of as a network of “neurons” organised in layers. The predictors (or inputs, 1, t-1, t-2) form the bottom layer, and the forecasts (or outputs, t) form the top layer. The intermediate layer, which is the nonlinear part, is called the hidden layer. [16]. The nonlinear part here is described as a "black box", as shown in Figure 2.

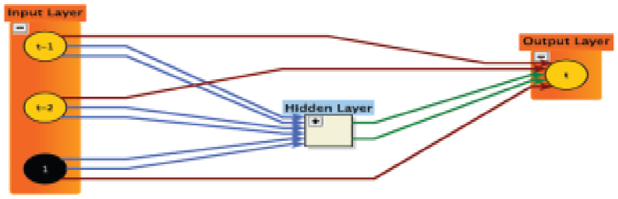

Figure 2: The Neural Networks
[17]

The nonlinear part contains *h*, called hidden neurons, transforms the input variables, weighted by parameters $\gamma_{ij}$ plus a bias $\gamma_{0j}$ via a nonlinear activation function $\psi(.)$. Let *i* indicate the number of lags and *j* the number of hidden neurons. A hidden neuron is denoted by

$$\psi\left(\gamma_{oj} + \sum_{i=1}^{n} \gamma_{ij} X_{t-i}\right) \qquad (1)$$

In the most cases $\psi(.)$ is the same for all hidden neurons, but it also can be chosen to be different for each hidden neuron.

Polynomial Neural Networks (PNN) algorithm is based on GMDH method and utilization a class of polynomials such as linear, quadratic and modified quadratic. The steps involved in the algorithm are: [10]

i. Determine system’s input variables. The input variables such as $x_1$, $x_2$, …$N_i$ are defined, related to output variables $y_i$, where N and i are the numbers of the entire input variables and input-output data set.

ii. Form training and testing data. The input-output data set is separated into training ($n_{tr}$) data set and testin ($n_{te}$) data set. Then, $n=n_{tr} + n_{te}$.

iii. Choose a structure of the PNN. The structure of the PNN is strongly depending on the number of input variables and the order of Partial Description (PD) in each layer.

iv. Determine the number of input variables and the order of the polynomial forming a PD. The total number of PDs in the current layer is equal to the combination ${}_NC_r$, that is.

$$\frac{N!}{r!(N-r)!} \qquad (2)$$

where N is the number of nodes in the preceding layer.

v. Estimate the coefficients of the PD. The vector of the coefficients of the PDs is determined by using the standard Mean Squared Error (MSE) obtained by minimizing the following index:

$$E_k = \frac{1}{n_{tr}} \sum_{i}^{n_{tr}} (y_i - Z_{ki})^2, k = 1, 2, \ldots, \frac{N!}{r!(N-r)!} \qquad (3)$$

$Z_{ki}$ is the output of the k-th node with respect to the i-th data

$N_{tr}$ is the number of training data subset.

Step (v) is completed repeatedly for all the nodes in the current layer and, for all layers of the PNN starting from the input to the output layer.

vi. Select the PDs with the good predictive capability. The predictive capability of each PD is evaluated by the performance index using the testing data set.
vii. Check the stopping criterion. The PNN algorithm terminates when the number of layers predetermined by the designer is reached.
viii. Determine new input variables for the nest layer. If the stopping criterion is not satisfied, the next layer is constructed by repeating Step (iv) through Step (viii).

## 4. The Predictive Analysis and Results

Figure 3 shows the number of abstracts from the period of 1997 to 2016.

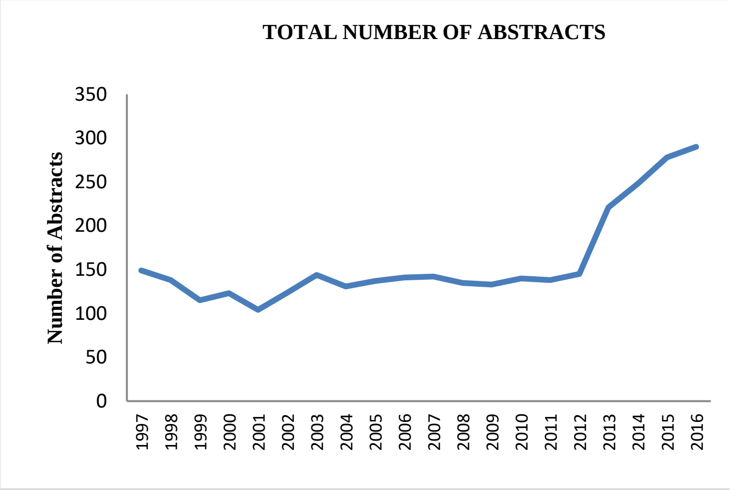

Figure 3: Total Number of Abstracts

The prediction model is given in equation (4).

$$Y1[t] = 0.530899 + cycle*0.00361419 + cycle*(-0.186178) \qquad (4)$$

Model complexity is 3 of 2 and the criterion value is 0.16382. The model formula is the expression that relates input variables and predicted variable. This is composed of coefficients and input variable names. Model complexity is the number of coefficients in the model and the number of layers. The above model has 3 layers and 2 coefficients. Criterion value informs about the value of Validation criterion configured in the Solver module. Top-ranked model has the smallest criterion value.

Table 1 presents the forecast result of twenty four months.

Table 1: Forecast Report

| # | 1 |
|---|---|
| Target name | Abstracts |
| Confidence band | 4.985372 |
| +1 | 26 |
| +2 | 26 |
| +3 | 27 |
| +4 | 27 |
| +5 | 27 |
| +6 | 27 |
| +7 | 27 |
| +8 | 28 |
| +9 | 28 |
| +10 | 28 |
| +11 | 28 |
| +12 | 28 |
| +13 | 29 |
| +14 | 29 |
| +15 | 29 |
| +16 | 29 |
| +17 | 29 |
| +18 | 29 |
| +19 | 29 |
| +20 | 29 |
| +21 | 29 |
| +22 | 28 |
| +23 | 27 |
| +24 | 32 |

*Model plots*

This is the time series chart used to analyse models visually. This is shown in Figure 4

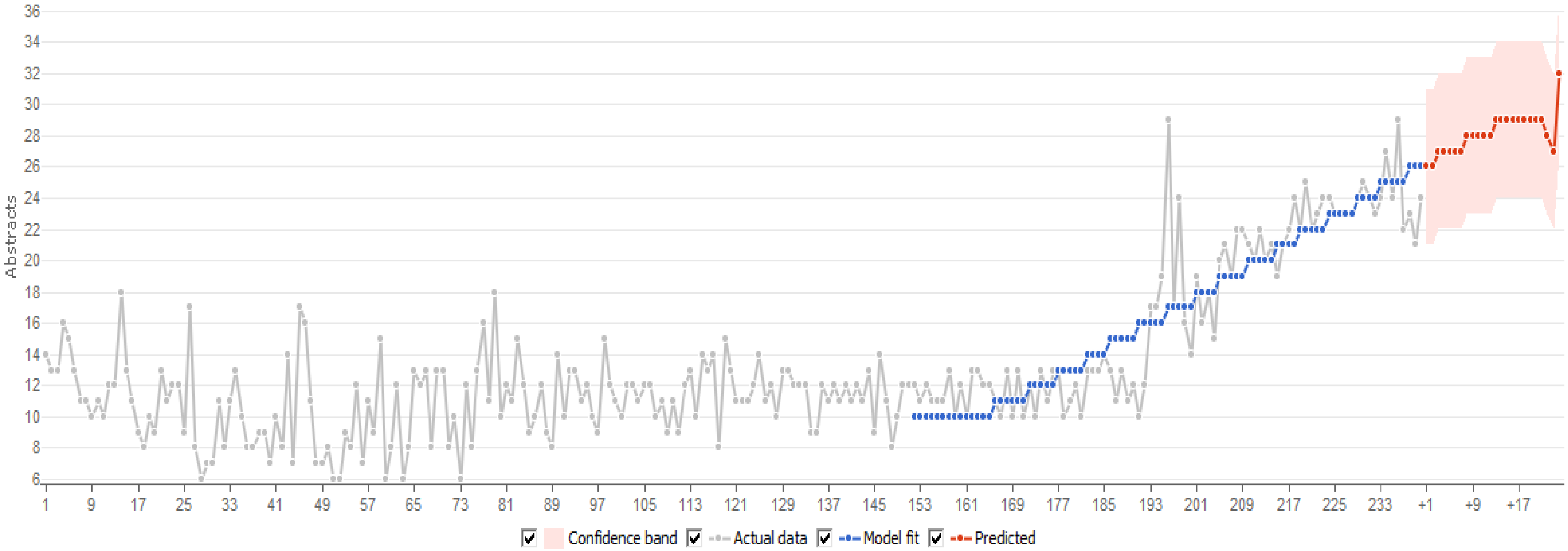

Figure 4: The Model Prediction

https://doi.org/10.20943/01201704.2934

In Figure 4, the. model predictions are red. Model fit are model values fitted to the data, they are blue. Actual data, the data available from the initial dataset is gray. Confidence band is the 95% confidence band calculated for predictions. Confidence band calculation uses model values fitted to the data (blue curve), it equals to two standard deviations (2*sigma) of model residuals. If predicted values have no their own IDs or future timestamps cannot be automatically generated, the forecast values will be marked with +1, +2, +3 on the horizontal axes. The Plot outputs post-processed (final) predictions using a thick line with dots. Raw models are plotted with a thin line.

Table 2 shows the accuracy metrics or the performance measures of the model, based on mean square errors, Mean absolute error and Root mean square error.

Table 2: Accuracy Metrics

| **Post-processed results** | **Model fit** |
|---|---|
| Mean absolute error (MAE) | 1.74157 |
| Root mean square error (RMSE) | 2.49269 |
| Standard deviation of residuals | 2.48698 |
| Coefficient of determination ($R^2$) | 0.797423 |
| Correlation | 0.89424 |

## 5. Conclusion and Future Work

In this study, the prediction of journal abstracts has been done using Polynomial Neural Networks based on Group Method of Data Handling method. The performance measures were mean absolute error and root mean square error, among others. Future work will be based on the comparative analysis of polynomial neural networks and time series model.

### Acknowledgment

The acknowledgement goes to the journal of The Institute of Electrical and Electronics Engineers Transactions on Computers from where the online data used for this study have been extracted.

### References

[1] Carroll, Leah, "HOW TO WRITE AN ABSTRACT: Tips and Samples," pp. 1-4.

[2] Hartley, James and Cabanac, Guillaume, "ThirteenWays to Write an Abstract," *MDPI*, vol. 5, no. 11, 2017.

[3] Lowry, P. B., Moody, G. D., Gaskin, J., Galletta, D. F., Humpherys, S. L., Barlow, J. B., & Wilson, D. W., "Evaluating journal quality and the association for information systems Senior Scholars'journal basket via bibliometric measures: Do expert journal assessments add value?," *MIS Quarterly*, vol. 37, no. 4, pp. 993-1012, 2013.

[4] Ware M., Mabe, M., The STM report: an overview of scientific and scholarly journal publishing, Oxford International Association of Scientific, Technical and Medical Publishers, 2009.

[5] Andonie, R., Dzitac, I., "How to write a good paper in computer science and how will it be measured by ISI Web of Knowledge.," *International Journal of Computers, Communications & Control*, vol. 4, pp. 432-446., 2010.

[6] K. Gurney, An introduction to neural networks, UCL Press Limited, 11 New Fetter Lane, London EC4P 4EE, 1997.

[7] Ojo, A.K and Adeyemo, A.B., "A Comparison Of The Predictive Capabilities of Artificial Neural Networks And Regression Models for Knowledge Discovery," *Computing, Information Systems, Development Informatics & Allied Research Journal*, vol. 4, no. 2, pp. 2167-1710, 2013.

[8] K. Gurney, An Introduction to Neural Networks., CRC Press, 2003.

[9] J. Wu, "An Effective Hybrid Semi-parametric Regression Strategy for Artificial Neural Network Ensemble and Its Application Rainfall Forecasting," in *Fourth International Joint Conference on Computational Sciences and Optimization*, 2011.

[10] Saeed Farzi, "The Design of Self-Organizing Evolved Polynomial Neural Networks Based on Learnable Evolution Model 3," *The International Arab Journal of Information Technology*, vol. 9, no. 2, pp. 124-132, March 2012.

[11] A. Abd Aziz, "PV boost converter conditioning using neural network.," Diss. Universiti Tun Hussein Onn Malaysia, Malaysia, 2013.

[12] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Waha, David Chek Ling Ngo, "Text mining for market prediction: A systematic review," *Elsevier: Expert Systems with Applications 41 (2014)* , p. 7653–7670, 2014.

[13] Jageshwer Shriwas, Shagufta Farzana, "Using Text Mining and Rule Based Technique for Prediction of Stock Market Price," *International Journal of Emerging Technology and Advanced Engineering*, vol. 4, no. 1, pp. 246-250, 2014.

[14] Enric Junqué de Fortuny,⇑, Tom De Smedt, David Martens, Walter Daelemans, "Evaluating and understanding text-based stock price prediction," *Elsevier: Information Processing and Management* , vol. 50, p. 426–441, 2014.

[15] Yoosin Kim, Seung Ryul Jeong, Imran Ghani, "Text Opinion Mining to Analyze News for Stock Market Prediction," *Int. J. Advance. Soft Comput. Appl.*, vol. 6, no. 1, pp. 2074-8523, 2014.

[16] Rob H Hyndman and George Athana-sopou-los, Forecasting: Principles and Practice, Online, Open-Access Textbooks, 2017.

[17] Sebastian Dietz, "Autoregressive Neural Network Processes.Univariate, Multivariate and Cointegrated Models with," 2010.

## Authors Profile

Adebola K. OJO is a lecturer and a researcher in the Department of Computer Science, University of Ibadan, Nigeria. She is a registered member of the Computer Professional of Nigeria (CPN). She had her Master of Science and PhD Degrees in Computer Science from University of Ibadan, Nigeria. She had her Bachelor of Science (BSc) in Computer Engineering from Obafemi Awolowo University (OAU), Ile-Ife, Nigeria. Her research interests are in Digital Computer Networks, Data and Text Mining, and Computer Simulation. She is also into data warehouse architecture, design and data quality via data mining approach.