https://doi.org/10.46344/JBINO.2025.v14i02.07

DEVELOPMENT OF DEEP LEARNING MULTI – LAYER PERCEPTRON MODEL USING NEGATIVE BINOMIAL REGRESSION APPROACH FOR WAITING TIME PREDICTION

[1]Megbatighon John Oke, [2]Emmanuel ShammaChaku, &[2]Bilkisu, Maijamaa

[1]Research Student Department of Statistics and Data Abalytics, Nasarawa State University Keffi, Nasarawa State
[2]Lecturer Department of Statistics and Data Abalytics, Nasarawa State University Keffi, Nasarawa State
maijamaab@nsuk.edu.ng

ABSTRACT

Evidently queuing model has been very useful in identifying appropriate levels of staff, equipment, and beds along with in making decisions about resource allocation and the design of new services as well as Waiting Time (WT) estimation and predictions. However, the traditional Queuing Theory approach has been known not to be sufficient in real life applications because the methodology is limited, for instance, unrealistic assumptions of the time distribution it requires to do queuing analysis. Thus, the goal of this study is to develop deep learning multi – layer perceptron model using negative binomial regression approach for waiting time prediction and examine the existing used ML algorithms for WT prediction at different sample sizes of queuing and make an essential tool for responsive actions for any organization (i.e. healthcare centers) reporting long waiting times. The study sets to follow Monte Carlos Simulation process and utilization of real-life data. Sequel to the developed multi-layer poisson regression (MLP-PR) and following the shortcomings of poisson regression to account for overdispersion in count data as well as to regularize the problem of over-fitting in MLP approach, this pursues to introduce a *Novel Multi-Layer Perceptron Negative Binomial Regression (MLP-NBR)*for WT Prediction. The Novel model is expected to handle problem of overdispersion, autocorrelation and over-fitting more effectively compare to the MLP-PR.

Key word: Waiting Time (WT), Deep learning multi – layer perceptron model, *Negative Binomial Regression (NBR),* Multi-layer poisson regression (MLP-PR), and Monte Carlos Simulation.

1. INTRODUCTION

Queuing theory also known as QT is an old-style mathematical method used to analyze queuing arrangements for decades (Gupta 2013). QT was developed by Erlang in 1904 in order to determine the capacity requirements of the Danish telephone system (see Brockmeyer *et al.* 1948). Ever since, the theory has been applied to a enormous range of service organizations like airlines, banks, supermarkets, telephone call centers and so on (Brewton 1989, Holloran and Byrne 1986, Brusco *et al* 1995, and Brigandi *et al* 1994) along with emergency systems for example police patrol, fire and ambulances (Larson 1972, Kolesar *et al* 1975, Chelst and Barlach 1981, Green and Kolesar 1984, Taylor and Huxley 1989). The theory has also been used in various healthcare settings as it will be discussed and applied later in this study. Queuing models can be very useful in identifying appropriate levels of staff, equipment, and beds along with in making decisions about resource allocation and the design of new services as well as Waiting Time (WT) estimation and predictions.

QT is the customary approach underlying the design of service processes, which relies on assumptions about the arrival process, service time and distributions (i.e. patients' distributions), the service machinery, and the queue discipline (Elisheva *et al.*, 2022). Studies such as Gupta (2013), Mahadevan (2015), Pianykh and Rosenthal 2015), and Elisheva *et al.*, (2022) have long established QT to violate the aforementioned restrictive assumptions, which often occur in practice leading to poor WT approximations. Also, QT models were known to

not capture human factors that may affect the WT. Thus, the traditional QT approach may not be sufficient in real life applications because the methodology is limited, for instance, unrealistic assumptions of the time distribution it requires to do queuing analysis (Mahadevan 2015; Pianykh and Rosenthal 2015). Consequently, recent studies such as Hassan and Richard (2021), Enrico *et al.* (2021), Elisheva *et al.*, (2022) to mention but few have investigated alternative prediction methods to the widely used QT-based ones, such alternative techniques are Machine Learning (ML) algorithms. Therefore, our focus is on increasingly popular ML algorithms in queuing analysis, which are exclusively suited to extracting patterns from large amounts of information stored in service system event logs. Studies have shown that modelling and predicting WT by means of ML algorithms which reflect more features than the traditional QT models, lead to more favorable WT predictions compared to the traditional QT models (Gal *et al.*, 2015; Mourõo*et al.*, 2017). Taking into consideration the claimed inadequacy of QT models for real-life service systems alongside the existing alternative ML algorithms, we believe that it is important to investigate to what extent the existing ML algorithms can best handle QT assumptions that are commonly seen in real-life systems. More specifically how best are the performances of these ML algorithms in modelling and predicting WT. Thus, it is against this backdrop this research is proposed.

2. **Statement of the Problem**

Single queue many-server service systems in which people queue until being served are common in many application areas (Elisheva *et al.,* 2022), such as healthcare centers, banking, supermarkets, call and chat centers, post offices etc. Effective and accurate WT modelling and predictions play a vital role in the management and design of service systems. For instance, planning decisions about the type and number of servers can be perfected when the WT modelling and predictions are taken into deliberation as can planning service or appointment scheduling (Pan *et al.,* 2020). While communicating to the people as delay announcements, improved WT predictions can affect vital system characteristics, such as clients' or customers' or patients' abandonment. In this perspective, inaccurate WT predictions, specifically underestimating WTs, may lead people to perceive their service experiences as inadequate, or cripple organizations' system performances. Therefore, accurate WT prediction is of great practical or real-world significance and is a vital motivation for this research. Consequently, recent studies have demonstrated more effective WT prediction of queuing using ML algorithms compare to the traditional QT modelling. Additionally, studies such as Hassan and Richard (2021), Enrico *et al.* (2021) and Elisheva *et al.*, (2022) have established effective performance of ML algorithms in reducing human error and better accuracy compared with traditional methods. Thus, the goal of this study is to compare the existing used ML algorithms for waiting time prediction with the pursuit to deliver optimal models for waiting time prediction at the different sample sizes of queuing and make an essential tools for responsive actions for any organization (i.e. healthcare centers) reporting long waiting times. This goal was motivated by high error rates in waiting time predictions as established queuing empirical studies.

3. **LITERATURE REVIEW**

This chapter reviews appropriate extant literature to the study. The chapter begins with the conceptual clarifications, followed by theoretical framework adopted in the study. The last part of the chapter discusses the empirical literature related to the study in order to identify the gaps in the literature.

Deep Learning: This area of Machine Learning deals with the design of programs that can learn rules from data, adapt to changes, and improve performance with experience. In addition to being one of the initial dreams of Computer Science, Machine Learning has become crucial as computers are expected to solve increasingly complex problems and become more integrated into our daily lives (Mahesh, 2018). Writing a computer program is a bit like writing down instructions for an extremely literal child who just happens to be millions of times faster than you. Yet many of the problems we now want computers to solve are no longer tasks we know how to explicitly tell a computer how to do. These include identifying faces in images, autonomous driving in the desert, finding relevant documents in a database (or throwing out irrelevant ones, such as spam email), finding patterns in large volumes of scientific data, and adjusting internal

parameters of systems to optimize performance (Jafar *et al.*, 2018).

Researchers have formally defined ML across pertinent literature. The term was coined by Samuel Arthur in 1959, who defined ML as a field of study that provides learning capability to computers without being explicitly programmed (Samuel, 1959). In the 1990s, Tom Mitchell gave a "well-posed" definition that has proven more useful to engineering set-up: "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E (Mitchell, 1997). Machine learning is a multi-disciplinary field having a wide-range of research domains reinforcing its existence. The simulation of ML models is significantly related to Computational Statistics whose main aim is to focus on making predictions via computers. It is also co-related to Mathematical Optimization which relates models, applications and frameworks to the field of statistics.

According to Mahesh (2018), machine learning (ML) is used to teach machines how to handle the data more efficiently. Sometimes after viewing the data, we cannot interpret the exact information from the data. In that case, we apply machine learning. With the abundance of datasets available, the demand for machine learning is in rise. Many industries apply machine learning to extract relevant data. The purpose of machine learning is to learn from the data. Many studies have been done on how to make machines learn by themselves without being explicitly programmed. Many statisticians and programmers apply several approaches to find the solution of this problem which are having huge data sets. Machine Learning relies on different algorithms to solve data problems. Data scientists like to point out that there is no single one-size-fits-all type of algorithm that is best to solve a problem. The kind of algorithm employed depends on the kind of problem you wish to solve, the number of variables, the kind of model that would suit it best and so on. Below are brief of some commonly used algorithms in ML.

In order to forecast patient waiting times in a system either in addition to or instead of queueing theory (QT), Hassan Hijry and Richard Olawoyin use deep learning methods for historical queueing variables. Four optimization techniques were used by them: SGD, Adam, RMSprop, and AdaGrad. To determine which model has the lowest mean absolute error (MAE), the algorithms were compared. For further comparisons, a conventional mathematical simulation was employed. The findings demonstrated that the DL model may be used to estimate patients' waiting times using the SGD algorithm, with the lowest MAE of 10.80 minutes (24% error reduction) activated (Hijry&Olawoyin, 2021).

With data from a variety of heterogeneous sources, such as electronic patient records and external non-hospital data, M. Dashtban and Weizi Li research attempts to develop an advanced predictive model for forecasting non-attendance with respect to the entire spectrum of likely supporting factors. They presented a novel deep neural network and machine learning model for non-attendance

prediction. By learning the underlying manifold of the data, the suggested method uses sparse stacked denoising autoencoders (SDAEs) to condense information and produce a better representation that may subsequently be used by other learning models. The suggested method is assessed using actual hospital data and contrasted with a number of popular and expandable machine learning methods. The evaluation findings show that the suggested strategy using logistic regression and the soft max layer performs better than alternative techniques.(Dashtban& Li, 2022).

4. RESEARCH METHODOLOGY

Our focus is on predicting a patient's WT at the moment they enter healthcare centers. Therefore, our first step is feature selection. We shall use a set of features that were available to us within the healthcare premises and that reflect the state of the system at time *t*. Specifically, we classified the features into the following four categories.

i- **Arrival-related features**: To describe the patient's arrival time, we used the day of the week and the arrival time during the day;

ii- **Service-related features**: The number of patients in service for each type of patient and the number of available and occupied doctors/nurses/health-consultant in the system at the patient's arrival time.

iii- **Queue-related features**: The number of patients in the queue for each type of patient.

iv- **Short-term history-related features**: This category contains features, calculated within a time window of 45min that capture the transient nature of the service system. We selected a time window of 45 min based on simulation results by fine-tuning it to capture system state changes. Making the window smaller than 45 min led to poor parameter estimation due to insufficient service time samples resulting in noisy service rate estimation. For the selected time window, we calculated the total WTs, the number of abandonments, the service time of the last customer in the system, and the mean service rate.

4.1 Technique for Data analysis and Model Specification:

After defining the features for the WT prediction, we performed a MinMax normalization for range rescaling to [0,1]. The MinMax normalization brings all features to the same scale and is defined as $y_i = \frac{y_i^0 - min\{y^0\}}{max\{y^0\} - min\{y^0\}}$ for each original feature y^0 and for $i = 1, \ldots, n$, where i is the instance number and there are n data instances for the feature. Afterwards the study, explores the count data model (i.e. Poisson Regression

4.2 Model Specification: Waiting times are a form of count data as time waited can only take non-negative integer values, and these integer values arise from counting rather than ranking (Cameron and Trivedi, 1998).

The benchmark model for count data is the Poisson model. This model assumes that the dependent variable, y_i, follows a Poisson distribution with mean μ_i. The Poisson probability distribution is given by: $P(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$ where μ_i represent the mean of y_i. This distribution is characterized by equidispersion, such that

the mean is equal to the variance i.e. $E(y) = var(y) = \mu$. The Poisson regression model incorporates observed heterogeneity into the Poisson distribution function, such that $E[y_i|x_i] = var[y_i|x_i] = \mu_i = \exp(x_i\beta)$. Thus, it is against this bedrock this research pursue to improve the WT prediction by the modification of the Poisson Regression (PR) approach to WT. The statistical model assumed for the data is that the values of the dependent variable Y follow a Poisson distribution of the form; $P(Y_i) = \frac{e^{-\lambda_i t_i}(\lambda_i t_i)}{Y_i!}$, where $\lambda_i$is the Poisson rate parameter at the settings of the predictor variables corresponding to the ith observation. It is further assumed that the rate is related to the predictor variables through a log-linear link function of the form;

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad 3.1$$

By taking exponential of both sides of equation (3.1), we have;

$$\lambda = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \quad 3.2$$

Further, the likelihood function (Poisson Distribution) is; $L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{y_i!} e^{-\lambda_i}$.

Take the logarithm of the likelihood function; we obtain the log-likelihood function,

$$l = \log(L) = \sum(y_i \beta' X_i - e^{\beta' X_i} - \log(y_i!)) \quad 3.3$$

One can see the relationship between the log of the Poisson distribution and the regression model (3.1).

Negative Binomial Regression Model

In practice, the variance is usually greater than the mean (overdispersion), and thus the Poisson model rarely fits the data well. To overcome the problems of the Poisson model, the negative binomial model can be used. This directly takes into account overdispersion, through an inclusion of an additional parameter. The negative binomial distribution is given by: $P(y_i|\mu_i, v_i) = \frac{\Gamma(y_i + v_i)}{y_i!\Gamma(v_i)} \left(\frac{v_i}{v_i + \mu_i}\right)^{v_i} \left(\frac{\mu_i}{v_i + \mu_i}\right)^{y_i}$, where Γ represents the Gamma probability distribution, $\mu_i = t_i\mu$, $v_i = \frac{1}{\alpha_i}$ determines the degree of dispersion and a > 0 defines the overdispersion parameter. The parameter μ is the mean incidence rate of y per unit of exposure. Exposure may be time, space, distance, area, volume, or population size. Because exposure is often a period of time, we use the symbol t_i to represent the exposure for a particular observation. When no exposure given, it is assumed to be one. The negative binomial regression model incorporates both observed and unobserved heterogeneity into the conditional mean such that $E[y_i|x_i] = \mu_i = \exp(x_i\beta + e_i)$. The parameter μ may be interpreted as the risk of a new occurrence of the event during a specified exposure period, t.The results below make use of the following relationship derived from the definition of the gamma function

$$\ln\left(\frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})}\right) = \sum_{j=0}^{y_i - 1} \ln(j + \alpha^{-1}) \quad 3.4$$

The Combining Deep Learning and Count Data Model for WT Predictions

Several researchers have applied ML algorithms for WT prediction as detailed in Senderovich et al. (2015), Ang et al. (2015), Mourõo et al. (2017), Sanit-in and Saikaew (2019) and so on. However, this study is limited to explore the Deep Learning ML

algorithms in the pursuit to produce a novel Negative Binomial Regression and Deep Learning Machine Learning model that would improve the WT prediction.

4.3 Justification Of Methods: The Multi-Layer Perceptron Poisson Regression and the Novel Multi-Layer Perceptron Negative Binomial Regression for WT Prediction

Following the development of Multi-Layer Perceptron Poisson Regression (MLP-PR) by Nader et al. (2009) and extended by Osval et al. (2021), the ϕ_h can be fixed as tangent hyperbolic function in hidden layer and exponential function in output layer(ϕ_o). Denote the inputs as x_i's and the outputs t_i' for MLP with one hidden layer.

$$t_k = \phi_0\left(\alpha_k + \sum_{j\to k}\omega_{jk}\phi_h\left(\alpha_j + \sum_{i\to k}\omega_{ij}x_i\right)\right)$$

i.

An improved Poisson Deep neural network model using the MLP framework was developed by Osval*et al.* (2021). The MLP-PR was established by substituting Poisson probability function in equation (3.9) and using equation (3.6) as Poisson means, the negative log-likelihood criterion can be obtained as:

ii.

$$E = -\sum_{n=1}^{N}[-t_n + y_n log t_n - ln y_n!]$$

3.10

iii.

Eliminating the last term which is not related to the model fitting, we have:

$$E = -\sum_{n=1}^{N}[-t_n + y_n log t_n]$$

3.11

Sequel to the developed MLP-PR and following the shortcomings of Poisson Regression to account for overdispersion in count data as well as to regularize the problem of over-fitting in MLP approach, this pursues to introduce a ***Novel Multi-Layer Perceptron Negative Binomial Regression (MLP-NBR)*** for WT Prediction by substituting negative binomial probability function in equation (3.9) using the ***Ensemble Learning model feature fusion algorithm***. We compare the performances of the models (i.e. MLP-PR and MLP-NBR) using simulations and healthcare real-life data.

5. KEY FINDINGS:

We present the summary of all the simulations conducted with brief discussion. Standard tabular formats will be used to capture all necessary information needed to build up comprehensive analyses that will showcase the evidences to support the results.

Based on the simulation results, the following are deduced:

i. Hybrid MLP-PR is optimal for patients WT prediction for 30 sample size and below;

ii. Hybrid MLP-NBR is optimal for patients WT prediction for sample sizes between 31 and 100; and

iii. NBR is optimal for patients WT prediction when sample sizes are large precisely when sample size is above 100.

References

Abir, M., Goldstick, J.E., Malsberger, R. *et al.* (2019). Evaluating the impact of emergency department crowding on disposition patterns and outcomes of discharged patients. *Int J Emerg Med,* 12,(4). https://doi.org/10.1186/s12245-019-0223-1

Adekitan A.I. (2018). Data Based Analytical Identification of Vehicle Maintenance Cost Components and Usage Data Trends. *International Journal of Mechanical Engineering and Technology*, 9(8), 691–701.

Alnuaim, A. A., Zakariah, M., Shukla, P. K., Alhadlaq, A., Hatamleh, W. A., Tarazi, H., Sureshbabu, R., & Ratna, R. (2022). Human-Computer Interaction for Recognizing Speech Emotions Using Multilayer Perceptron Classifier. *Journal of Healthcare Engineering*, *2022*. https://doi.org/10.1155/2022/6005446

Ang E., Kwasnick S., Bayati M., Plambeck E., &Aratow M. (2015). Accurate emergency department wait time prediction. *Manuf. Serv. Operations Manage*, 18(1), 141–156.

Ansari M.D.F., Khan A.H., Kathula P., & Attar E.F. (2021). Predict Queue Wait Time in A Hospital Environment. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 8(5).

Argawu, A. S., &Mekebo, G. G. (2022). Risk factors of under-five mortality in Ethiopia using count data regression models, 2021. In *Annals of Medicine and Surgery* (Vol. 82). Elsevier Ltd. https://doi.org/10.1016/j.amsu.2022.104764

Arha, G. (2017). *Reducing Wait Time Prediction In Hospital Emergency Room: Lean Analysis Using a Random Forest Model.* Master's Thesis, University of Tennessee.

Baggio, S., Iglesias, K., & Rousson, V. (2018). Modeling count data in the addiction field: Some simple recommendations. *International Journal of Methods in Psychiatric Research*, *27*(1). https://doi.org/10.1002/mpr.1585

Barach P. & Johnson J.K. (2006). Understanding the complexity of redesigning care around the clinical microsystem. *QualSaf Health Care*, i10–i16. doi: 10.1136/qshc.2005.015859

Benevento, E., Aloini, D., & Squicciarini, N. (2023). Towards a real-time prediction of waiting times in emergency departments: A comparative analysis of machine learning techniques. *International Journal of Forecasting*, *39*(1), 192–208. https://doi.org/10.1016/j.ijforecast.2021.10.006

Bergen J., Nister D. &Naroditsky O. (2006). Visual odometry for ground vehicle applications. *Journal of Field Robotics*.https://doi.org/10.1002/rob.20103

Bergen mar. et al.. (2006). The Content of a Primary Care Population: Including the Patient Agenda. *Journal of the American Board of Family Practice,* 16(4), 279-283.

Breiman L. (1996). Bagging predictors. *Mach. Learn.,* 24(2), 123–140.

Brewton G.W. (1989). A pilot study of diethyldithiocarbamate in patients with acquired immune deficiency syndrome (AIDS) and the AIDS-related complex. *Life Sciences,* 45(26). https://doi.org/10.1016/0024-3205(89)90234-8

Brockmeyer F., Halstrom H.L. & Jensen A. (1948). The Life and 3 Works of A.K. Erlang. *Trans. Dan. Acad. Tech. Sci.,* No. 2.

Brusco M.J., Jacobs L.W., Bongiorno R.J., Lyons D.V. & Tang B. (1995). Improving Personnel Scheduling at Airline Stations. *Operation Research*, 43(5). https://doi.org/10.1287/opre.43.5.741

Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N. (2016). *A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection.* In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9908. Springer, Cham. https://doi.org/10.1007/978-3-319-46493-0_22

Cameron A & Trivedi P (1998). Regression analysis of count data, vol. 30. Cambridge University Press.

Chelst K.R. &Barlach Z. (1981). Multiple Unit Dispatches in Emergency Services: Models to Estimate System Performance. *Management Science,* 27(12). https://doi.org/10.1287/mnsc.27.12.1390

Curtis F.E., Bottou L. &Nocedal J. (2018). Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2).

Dashtban, M., & Li, W. (2022). Predicting non-attendance in hospital outpatient appointments using deep learning approach. *Health Systems*, *11*(3), 189–210. https://doi.org/10.1080/20476965.2021.1924085

Di S. S., Paladino, L, V., Lalle, I., Magrini, L., & Magnanti, M. (2015). Overcrowding in emergency department: an international issue, Internal and emergency medicine, 10,(2), 171-175.

Dong P., Wang H., Tingting Fang a, Yun Wang b c, Quanhui Ye (2019). Assessment of extracellular antibiotic resistance genes (eARGs) in typical environmental samples and the transforming ability of eARG. *Environment International,* 125, 90-96.

Du, J., Park, Y. T., Theera-Ampornpunt, N., McCullough, J. S., &Speedie, S. M. (2012). The use of count data models in biomedical informatics evaluation research. In *Journal of the American Medical Informatics Association* (Vol. 19, Issue 1, pp. 39–44). https://doi.org/10.1136/amiajnl-2011-000256

Eiset, A.H., Kirkegaard, H. & Erlandsen, M. (2019). Crowding in the emergency department in the absence of boarding – a transition regression model to predict departures and waiting time. *BMC Med Res Methodol* 19, 68. https://doi.org/10.1186/s12874-019-0710-3

Elisheva C., Izack C. & Paul F. (2022). Delay Prediction for Managing Multiclass Service Systems: An Investigation of Queueing Theory and Machine Learning Approaches. IEEE *Transactions On Engineering Management*, 0018-9391.

Enrico B., D. Aloini, & N. Squicciarini (2021). "Towards a real-time prediction of waiting times in emergency departments: A comparative analysis of machine learning techniques," *Int. J. Forecasting, to be published*, doi: 10.1016/j.ijforecast.2021.10.006.