Dr. Kusumadhara S, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 13, Issue 3, March 2023, pp. 43-50

RESEARCH ARTICLE OPEN ACCESS

Machine Learning Based Genetic Algorithmic Optimization Model for Inpatient Length-of-Stay Prediction

Dr. Kusumadhara S[1], Mr. Jagadeesh M[2], Mr. Sathyajith M[3]
Electronics and Communication Engineering
K.V.G. College of Engineering, Sullia

ABSTRACT
European nations have a plethora of Home Health Care (HHC) providers that visit patients in their homes to aid in their recovery from sickness or accident. As transportation expenses are a significant part of any business operating budget, reducing such costs is crucial, especially in the Residential Health Care Industry. This study examines the Route Planning Scheduling issue from the point of view of HHC organizations, with the goal of optimizing transportation costs. A study of HHC providers found that the amount of medicine needed by each patient is unpredictable when it comes to planning delivery routes. For the sake of realism, this research treats uncertainties as a fuzzy variable. We consider fuzzy demand in the home healthcare scheduling problem and develop a fuzzy probability constraint model. To address the suggested paradigm, we advocate for a hybrid evolutionary algorithm that incorporates stochastic simulation techniques. First, we reduce it to the well-known vehicle routing issue under time constraints. The suggested technique is shown to perform well in experiments on both the Solomon and Homburger benchmark examples. The Dispatch Preference Index (DPI) variable is then set to values within the range [0, 1] to conduct further trials on the fuzz version model. Finally, the optimal value of DPI is determined via the use of stochastic simulation, after which the impact of DPI on the problem's goal and its indicators is addressed. The findings of this study will aid HHC providers in making informed judgements on the scheduling of their vehicles.

Keywords: Home Health Care, DPI, HHC, Hospital, Genetic Algorithm, Random Forest Algorithm, XGB Genetic Algorithm.

I. INTRODUCTION

The duration of hospital (LOS) has been the subject of research and study since the 1970s, with the goal of improving hospital quality and efficiency. The goal of every hospital is to improve patient care while spending as little money as feasible. Reducing healthcare spending in developed nations without lowering quality is mostly achieved by focusing on the length of stay (LOS) as a major performance metric [1]. Problems with hospital bed management have arisen as a direct consequence of the rising number of admitted patients and the associated costs associated with the inpatient units. Hospital bed management is complicated by several factors, including patient duration of stay and a lack of knowledge about expected release time [2]. While it may be difficult to anticipate the duration of stay due to the many elements that contribute to it, knowing the true value may greatly aid in the scheduling of beds and personnel [3]. Hospital quality, productivity, and performance may be measured in part by looking at LOS. This suggests that LOS prediction may be useful in addressing issues such as resource allocation, capacity planning, and personnel manning. It improves service delivery, patient safety, healthcare efficiency, and cost savings [4]. Waste and blocked bed days might result from inaccurate LOS forecasts. Dissatisfaction among both patients and medical staff might ensue, which in turn disrupts the delivery of care.

This paper proposes a healthcare-related application of agent-based techniques in an effort to bridge the gap between these two fields. Here, we use an agent-oriented paradigm to examine an ED's management structure as an example of a healthcare business process. We used **NetLogo,** a free, open-source, programmable modelling environment for many agents that serves as a type of benchmark toolkit [15].

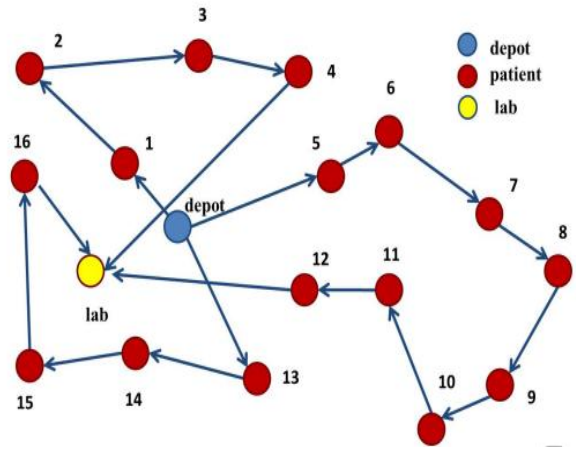

Figure No. 1: *The BPMN Supply Chain's Unique Architecture*

Formalisms like Directly Follows graphs, Petri nets, and the Business Process Modeling Notation (BPMN) have long been considered in the field of business process analysis [3]. Doctors and nurses, both important members of the healthcare team, sometimes lack the formal education necessary to comprehend this kind of modelling. Since the notion of the agent is intelligible even for not competent experts of the AI sector [19], we made sure to adopt a clear representation of agents working in the environment. Part of our work involves proposing an agent-based model for healthcare that can be used to conduct scenario analysis within the context of business process management. Second, we investigate a parameter-variable, agent-oriented decision-support system (parameter sweeping) with the aim of optimizing it via genetic algorithms (GA) [13]. We investigate a five-dimensional space with the goals of decreasing patient wait times in the emergency department (ED) while increasing throughput and maintaining high quality standards. We show that an advisor medical decision support system is possible by using stochastic optimization methods to agent interactions.

II. LITERATURE SURVEY

Recent work has focused on the function of autonomous entities and multi-agent networks in healthcare [15], in the context of advisor applications to healthcare [2]. When it comes to agent-based simulations, however, it was logistics and marketing that got the most attention. In addition, "agent-based care platforms and simulation" (which accounts for 32.60%) is a prominent use of agent-based modelling in the healthcare area [8]. Planning accounts for just 9% of the content we produce, while "Decision Support Systems" only accounts for 12% [12]. Current research has focused on ER simulation for both operational management [12] and patient care [13]. Using AI methods for ABM in BPM is an intriguing new direction.

The route 90 optimization issue has been studied extensively in the past fifty years. The traditional "travelling salesman" issue is the simplest example of a routing problem (TSP). The goal of this challenge is to minimize total distance a salesperson must travel between his several stops before returning to his starting point (**Ursani et al., 2011**). When each city (hence referred to as a client) has a demand and each vehicle has a fixed capacity, Service Problem (VRP) (**El-Sherbeny, 2010**).

Few studies address HHC issues, despite their growing significance (Liu et al., 2014). Services provided by HHCs have been examined from a variety of angles, including policy, health 120 treatment center, logistics (**Harris, 2015**), etc.

Regression models, machine learning, and deep learning (a kind of ML) are the three main groups of LOS prediction approaches [8, 9]. When Baek et al. [13] used a multivariate regression approach to all available hospital inpatient data, they found an R2 value of 0.267. In a similar vein to **Beak et al., Ray-Zack et al.** used a multivariate regression approach to foretell the LOS of radical laparoscopic procedure for patients with muscle-invasive bladder cancer. It was found that the regression model had an R2 of 0.048 [15]. After heart surgery, **Meadows et al. [14]** developed a logistic regression approach for predicting whether intensive care unit (ICU) patients will need short-term or long-term hospitalization.

To predict how long diabetes patients will be in the hospital, **Alahmar et al.** [10] used the stacked-ensemble technique. When compared to lack of efficiency models such as regression-based, tree-based, and artificial neural network (ANN) models, their newly suggested technique performed the best (accuracy 0.81). The ensemble technique outperformed the random forest approach (accuracy 0.80) and the xgboost method (accuracy 0.80), but the difference was not statistically significant. They manually ran HPO to fine-tune the parameters of interest.

Often used is a technique called "grid search," in which the user specifies a small selection of hyperparameters for the machine learning algorithm of interest and the approach iteratively searches through those parameters. Grid search is a reliable approach in low-dimensional spaces (i.e., 1D or 2D) due to its easy implementation and parallelism capabilities; nevertheless, the computational cost skyrockets with the number of hyperparameters [13].

The genetic algorithm (GA) is an example of the metaheuristic optimization algorithms based on a population and motivated by the concept of natural selection. This method iteratively applies genetic operators to each member of the population to produce a new population. The chromosomes, the selection, the crossover, the mutation, and the fitness function make up the core components of this algorithm. A random population of Y (where Y is the total number of solutions) with n chromosomes (where n is the total number of issue parameters). From the population, the two most fit chromosomes, C1 and C2, are used to form the solution. With the crossover operator, the offspring of C1 and C2 will be O. The parameter CP representing the likelihood of such an action occurring is called the crossing probability. All a population must be formed by selection, crossover, and mutation. The GA can continuously search for and obtain the best solution [14] due to the frequency of crossover and mutation.

III. PROPOSED SYSTEM

Our paper discusses the emergency department of a hospital in a highly populated region of northern Italy. There are seven emergency department nurses, four medical physicians, two scanning nurses, and two social services working in the department. The department offers the following services (or tests): a blood analysis, radiography, and imaging. Over 47,000 people visit the clinic each year, with an average of 125 new patients each day. Mondays and Fridays have the highest volume of patients (16.9% and 15.3% respectively in 2019), while Saturdays and Sundays see the lowest (13.4% and 14.5% respectively). Tuesday had the highest daily arrival rate at 15%, followed by Wednesday at 15.4%, and Thursday at 14.6%. From 2019, Italy has used a four-point Emergency Severity Index (ESI) ranging from 1 (extremely severe) to 4 (low) (very low). As of 2020, the international 5-point scale will be mandatory under Italian law. In ED, activities often follow a predictable pattern. Although emergencies are sent immediately to Stockroom, regular patients go through the steps of Registration, Triage, and Visit, which may include tests and consultations with doctors. Department heads at hospitals want to learn more about the ED model so they may implement reforms inside their own institutions.

IV. METHODS AND MATERIALS

Source of Data:This study's hospital has 300 beds and employs 1055 doctors and other medical professionals. The hospital has 19 different inpatient departments and offers both clinical and non-clinical care. There is a database for storing information about patients. All the data used in this analysis came from the internal medicine division. Similar research was evaluated to identify potential factors influencing LOS and to compile the relevant data. The information system was randomly sampled to get two hundred records, one hundred male and one hundred female. Age, sex, insurance, marriage, medical advice line, and physician specialty are only few of the characteristics included in Table 2 below.

Sex	Sex_Female	Sex_Male
Female	1	0
Male	0	1

Marital Status	Marital Status_Married	Marital Status_Single
Married	1	0
Single	0	1

Main insurance type	Main insurance type_ Ordinary SSI	Main insurance type_Occupational SSI	Main insurance type_Employee HI	Main insurance type_Special SSI	Main insurance type_ Ordinary SSI	Main insurance type_Without insurance
Ordinary Social Security Insurance (Ordinary SSI)	0	0	0	0	0	0
Occupational Social Security Insurance (Occupational SSI)	0	0	0	0	0	0
Employee Health Insurance (Employee HI)	0	0	0	0	0	0
Special Social Security Insurance (Special SSI)	0	0	0	0	0	0
Without insurance (Without insurance)	0	0	0	0	0	0

Data Preprocessing: After double-checking the numbers, we found that there were no blanks. Patients' average age was 63, and their standard deviation was 19. There were as many female data points as male ones. Around 90% of patients had spouses while the other 10% were single. Two was the mean, with three the standard deviation, when it came to calling for medical guidance. There was a mean LOS of 2 weeks and a variance of 4.4 days. Variables like main insurance and physician specialty were not distributed normally. Ninety percent of the patients were covered by regular Social Security insurance, with the remainder covered by various additional plans. Over half (45%) of patients were seen by family doctors, another 54% by experts, and the other 2% by subspecialists. The distributions of the variables are shown in Table 1.

The one-hot encoding method converts categorical data into binary variables that may be analysed by ML systems. Then, the ML algorithm may operate on these updated binary variables [3].

One-hot encoding involves the generation of a new binary feature for each tier of categories [5]. Figure 1 depicts a one-hot coding of four category variables. One more dimension (variable) is added to the factors with each classification of a variable, with the value of this dependent addition in each row being 0 or 1. When the original category variable corresponds to the dummy variable, the value of the regression coefficient is 1, and otherwise it is 0. At last, the entries containing the original category variable are deleted.

Model Training:To evaluate the efficacy of the enhanced model, many models were used, including KNN [11], multiple regressions [6], logistic regression [7], randomized forest [4], artificial neural network [38], and XGBoost [9]. Python 3.8.5 was used for the model construction. An estimate of 12 factors in the KNN framework was made. The K-CV technique, set to a k10, was used to make the estimate. There were two versions of the regression model created. The initial version was constructed with LOS as the sole determinant of all other variables. A natural logarithm of LOS was computed and included in the data set after the regression assumptions were verified. Transformed LOS served as the basis for a second multivariate regression model, hereafter referred to as a converted regression model. Models of regression and modified regression were reevaluated on the test set after being reconstructed using t-test findings with a significance level of less than 0.05. The two versions of this model will be referred to as Lm and Lm transformed. Several models were constructed using the log transformation of LOS since doing so improved the plug and got the data closer to a normal distribution.

Table No. 1*: A single code for each of four classes.*

Physical Expertise Level
General Practitioner
Specialist
Subspecialty Physician

→

Physician Expertise Level_ General Practitioner	Physician Expertise Level_ Speciality	Physician Expertise Level_ Subspecialty Physician
1	0	0
0	1	0
0	0	1

Default hyperparameter settings were used to construct decision trees (DT default), random forests, and XGBoost (XGB default) using the training dataset. The artificial neural network model was developed using a 2-tier architecture. There are twelve neurons in the first stratum and six in the second. The models were then tested on a dataset and their performance was assessed. Table 1 presents the specifics of the default hyperparameters of forest models.

Optimization using Genetic Algorithm: The tree-based models' hyperparameters are initially set to the default values provided by the Python modules used to create the models. The models may be run with a wide variety of possible settings for the hyperparameters. In this work, the GA was implemented in HPO using the Python-based PyGAD module and the PyGAD.GA class. Determining the fitness function, establishing the range of hyperparameters for each model to be assessed in the GA, and setting the parameters of the GA are the three fundamental phases in implementing the GA for each model. For each model, the mean squared error (MSE) is used as the fitness function to reduce the model's overfitting on the training data using the K-CV technique and the parameter value k5.

For each model, the GA will need to verify the following hyperparameter space. The decision tree framework allows for a maximum tree depth between 1 and 1000. If you increase max depth, your tree will grow deeper, and you will overfit the data. Between 1 and 50 samples are required for each node. While calculating alpha, a number between 0 and 1 is used. If ccp alpha is set to 0, no trees are removed, whereas increasing its value results in more trees being removed.

V. MODELING TECHNIQUE

As such, we are interested in learning more about agent-based modelling of a business operations. Agents use the departmental map as a basis for their actions in a 3D setting that facilitates the flow of work across various tasks (Fig. 2). Based on their current condition, agents either search for the next patient or begin doing the activity, wait for it to begin, perform it, and then look for the next client again. Actions are things that include the factors of study. Agents engage in dynamic interactions with one another and their surroundings/activities. In this paper, we provide a formal definition of the links between nodes in a network. The arc's density represents the typical time required to travel between the two hubs. Characteristics of the employees, such their proficiency and speed of action, are modelled in this agent-based simulation.Every agent in the simulation acts initially in accordance with its current condition. When a slot opens, the highest priority patient is chosen. They get to work on the project for whatever long the Duration agents' variable says they should. After the allotted time has passed, the patient moves to the waiting area and the next-task variable is updated with the title of the next activity.

Design: This model presents a means through which the ED's overall operation may be influenced by modifying the basic principles governing agent behavior. Patients are prioritized by I urgency and

(ii) length of stay in the ED before being seen. Some patients may leave the emergency room before being treated by a doctor; therefore, we take into account the agent's urgency level (very low or low), the number of many other agents in the waiting area, and the amount of time since arrival. To accurately portray patients' daily arrivals, stochasticity is necessary. Every day, we choose a new random sample size of patients to represent the population and generate new numbers that are distributed realistically. The daily arrival totals were also modelled in a similar fashion.

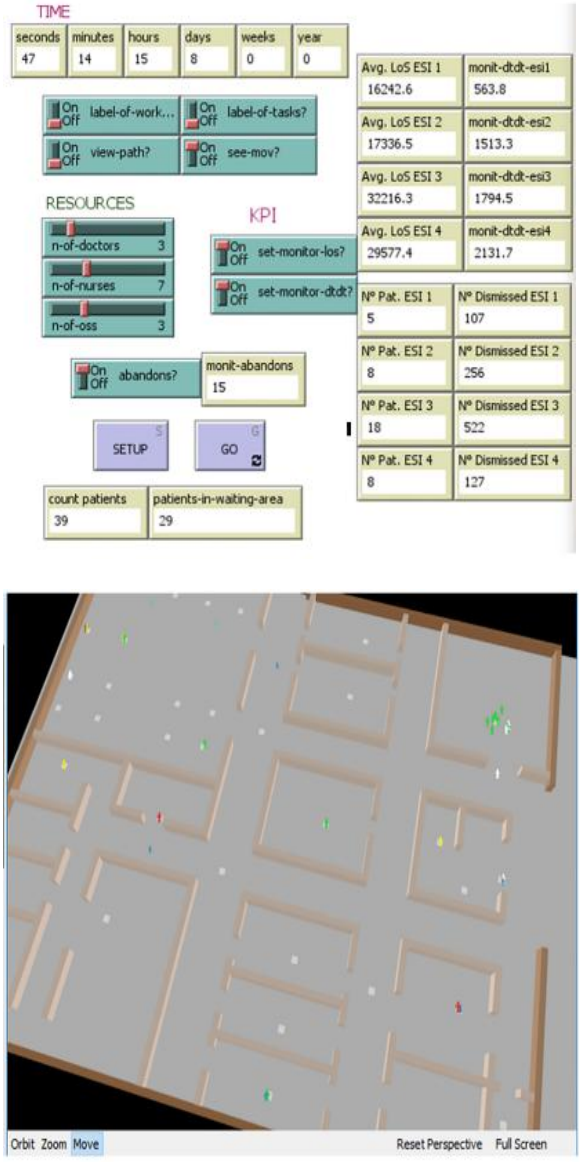

Figure No. 2:*Users may interact with a 3D model of the ER's buttons, screens, and output area for a more nuanced understanding of the ways in which medical staff and patients move.*

Detail:The model starts off in the same place as a Monday night when the department is vacant. We think a day of pre-management is necessary for properly handling business process indicators. The ED's route is brought in from an external file containing the network's graph. The name of the task, the number of operators necessary, the kind of operators needed, and the time required to complete the job for each operator are all included in the data for each node.

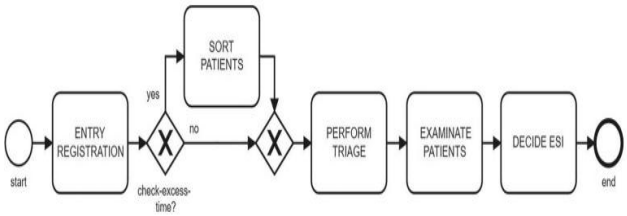

Figure No. 3:*The BPMN Triage and Registration Method*

Performance Process Indicator:There are two predicators available in the literature [4] that may be used to evaluate the reliability of simulation findings. Length of Stay (LoS), the average time a patient spends in the hospital from admission to discharge, and Door-to-Doctor-Time (DTDT), the time from admission to the patient's first scheduled doctor's appointment, are two such metrics.

Performance Indicator		***ESI1***	***ESI2***	***ESI3***	***ESI4***
DTDT	Avg	12.3	17.1	22.3	23.4
	Avg Dev	0.9	0.3	0.7	1.4
LOS	Avg	220.2	259.3	431.25	435.2
	Avg Dev	20.3	21.3	12.9	24.3

Table No. 2:Simulation output of two leading performance indicators (Door-to-Doctor-Time, DTDT and Length-of-Stay, LOS) in four weeks, patientsby ESI (times in minutes)

Resource Utilization: Resource usage, in addition to achievement indicators, is of special relevance. Our methodology calculates two variables related to the time that physicians and nurses spend at work. During the four-week simulation, physicians spent around 83.5% of their time working directly with patients, while nurses spent approximately 48.2%. Staff members attest to the truthfulness, because nurses do a variety of things outside of patient care (e.g., manage drugs, prepare tools, talk to relatives etc.)

VI. ENROLLMENT PROCESS USING GENETIC ALGORITHM TECHNIQUE

Registration: Patient scheduling is an important issue for management to consider in addition to the process's overall functioning. We investigated using the GA method with the activities shown in

Fig. 2. Priority of care and length of wait time are the two most important factors in determining the next patient. The less pressing matters may have to wait too long if we consistently give precedence to the more pressing ones,

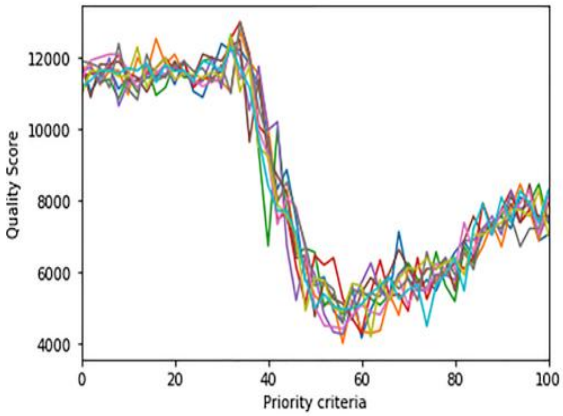

Figure No. 4:The outcomes of 10 iterations of sweeping the parameters by changing the "prioritycriteria" to reduce QS

There is a danger of not getting to the most urgent patients in time if they are treated based only on the time, they were brought in. We offer a method for assessing the effectiveness of the proposed remedies, with special emphasis on the fact that more pressing problems (ESI 1) need prompt treatment. Yet, less pressing matters may be put off, but only up to a point. To prioritize patients based on these criteria, the check-excess-time gateway was developed. Experts in the field told us to calculate a Quality Score (QS) by multiplying the number of critically ill and injured patients by four. We can now evaluate the accuracy of the models.

Parameter Sweeping: To study parameter sweeping, we utilised a programme that was built into NetLogo (BehaviorSpace). 2 In a first step, we determine the optimal value for a parameter by lowering the QS value while just considering the selection criteria, such as whether the next patient is chosen based on urgency or waiting time. We describe the choice that free workers make between patients as a percentage chance we label "prioritycriteria." Our model was run repeatedly, each time with a different combination of parameters to track. By testing several configurations of the model, one may isolate the parameters that get the desired results. We observed a U-shaped curve (Fig. 5) when the crucial "prioritycriteria" parameter was changed, which led us to conclude that the optimal value of this parameter, which minimizes QS, is in the range of 55 to 60.

VII. RESULTS

Then, we use a problem specification for which we know the brute-force solution to put the GA's construction to the test. We started using a third-party programme that communicates with NetLogo. 3 Our first test used a 50-person population with a 0.7 crossover rate and a 0.3 mutation rate. The objective is to decrease QS by changing the criterion within a narrow range of 5 values, from 0 to 100. We may evaluate the quality of our model using the QS value of 60 achieved with GA and the highest accuracy of 4,373, which is quite close to the number produced via parameter sweeping.

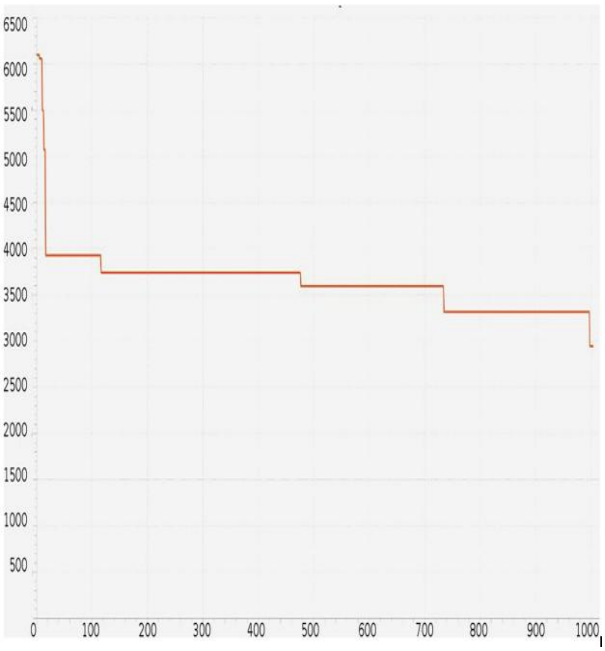

Figure No. 5:The minimum GA fitness Quality is 2,941, which is a significant increase above the results achieved using a purely brute-force strategy. After the model is verified, a more intriguing area of study involves determining the "priority-criteria" selection criterion, in a range of 0 to 100, and the maximum waiting parameters (in seconds) for each kind of emergency (from ESI 2 to ESI 5, since ESI 1 instances are processed instantly).

As a result, there are five dimensions to explore. With a starting overall population of 100, a coefficient of skewness of 0.7, and a mutation rate of 0.3, GAs will always search for the minimum of QS. Eventually, a fitness value of 2,941 produced from the following five values is highly interesting, suggesting that GA produces interesting outcomes. 55 is the priority criterion, 1,380 is the second threshold, 2,280 is the third, 9,060 is the fourth, and 7,200 is the fifth.

CONCLUSION

In this paper, we looked at how hospital administration may benefit from adopting an agent-based modelling strategy. While considering a range of KPIs and scenario analysis, the problems tackled are fairly typical of any business process study. We used GA, an AI method for exploring parameter spaces, to look at the issue of order entry as a parameter-searching optimization problem.

Using an ABM as a model, we demonstrate how GA may be used to evolve the parameters that ultimately determine the behaviour of a system. After ensuring the model is accurate, the suggested method first contrasts parameter sweeping with GA. This verifies the right GA configuration since a brute-force technique also gets the same answer with a single parameter. In order to determine the thresholds (i.e., the maximum waiting parameters established by kind of urgency) that may be advised for medical reasons to choose the next client in the admission process, we did an investigation of a five spatial dimensions issue. In order to enhance the quality of the process in accordance with criteria specified by medical management, the GA findings offer parameter values to decision-making.

REFERENCE

[1]. A. Morton, E. Marzban, G. Giannoulis, A. Patel, R. Aparasu, and I. A. Kakadiaris, "A comparison of supervised machine learning techniques for predicting short-termin-hospital length of stay among diabetic patients," in Proceedings of the 2014 13th International Conference on Machine Learning and Applications, pp. 428–431, Detroit, MI, USA, December 2014.

[2]. A. Alahmar, E. Mohammed, and R. Benlamri, "Application of data mining techniques to predict the length of stay of hospitalized patients with diabetes," in Proceedings of the 2018 4th International Conference on Big Data Innovations and Applications (Innovate-Data), pp. 38–43, Barcelona, Spain, August 2018.

[3]. W. E. Muhlestein, D. S. Akagi, J. M. Davies, and L. B. Chambless, "Predicting inpatient length of stay after brain tumor surgery: developing machine learning ensembles to improve predictive performance," Neurosurgery, vol. 85, no. 3, pp. 384–393, 2019.

[4]. G. Danilov, K. Kotik, M. Shifrin, U. Strunina, T. Pronkina, and A. Potapov, "Prediction of postoperative hospital stay with deep learning based on 101 654 operative reports in neurosurgery," in ICT for Health Science Research, pp. 125– 129, IOS press, Amsterdam, Netherlands, 2019.

[5]. B. Mahboub, M. T. A. Bataineh, H. Alshraideh, R. Hamoudi, L. Salameh, and A. Shamayleh, "Prediction of COVID-19 hospital length of stay and risk of death using artifcial intelligence-based modeling," Frontiers of Medicine, vol. 8, Article ID 592336, 2021.

[6]. B. Alsinglawi, O. Alshari, M. Alorjani et al., "An explainable machine learning framework for lung cancer hospital length of stay prediction," Scientifc Reports, vol. 12, no. 1, pp. 607–610, 2022.

[7]. A. Paleyes, R. G. Urma, and N. D. Lawrence, "Challenges in deploying machine learning: a survey of case studies," ACM Computing Surveys, vol. 55, no. 6, pp. 1–29, 2022.

[8]. S. Putatunda and K. Rama, "A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost," in Proceedings of the 2018 International Conference on Signal Processing and Machine Learning, pp. 6–10, Shanghai, China, November 2018.

[9]. Y. Chen, "Prediction and analysis of length of stay based on nonlinear weighted XGBoost algorithm in hospital," Journal of Healthcare Engineering, vol. 2021, pp. 2021–9, Article ID 4714898, 2021.

[10]. K. Budholiya, S. K. Shrivastava, and V. Sharma, "An Optimized XGBoost Based Diagnostic System for Efective Prediction of Heart Disease," Journal of King Saud UniversityComputer and Information Sciences, vol. 34, no. 7, pp. 4514–4523, 2020.

[11]. J. H. Wasfy, K. F. Kennedy, F. A. Masoudi et al., "Predicting length of stay and the need for postacute care after acute myocardial infarction to improve healthcare efciency: a report from the national cardiovascular data registry's action registry," Circulation: Cardiovascular Quality and Outcomes, vol. 11, no. 9, p. e004635, 2018.

[12]. Wang, T.-C., Taheri, J., Zomaya, A.Y.: Using genetic algorithm in reconstructing single individual haplotype with minimum error correction. J Biomed Inform 45(5), 922–930 (2012). doi:10.1016/j.jbi.2012.03.004

[13]. Chen, Z.Z., Deng, F., Wang, L.: Exact algorithms for haplotype assembly from whole-genome sequence data. Bioinformatics 29(16), 1938–1945 (2013). doi:10.1093/bioinformatics/btt349

[14]. Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., Rice, P.M.: The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38(6), 1767–1771 (2009). doi:0.1093/nar/gkp1137

[15]. Y. Jiang, G. Tong, H. Yin, and N. Xiong, "A pedestrian detection method based on genetic algorithm for optimize XGBoost training parameters," IEEE Access, vol. 7, pp. 118310–118321, 2019.