© Universiti Tun Hussein Onn Malaysia Publisher's Office

JSCDM

http://penerbit.uthm.edu.my/ojs/index.php/jscdm

e-ISSN : 2716-621X

Journal of Soft Computing and Data Mining

A Comparative Study of Artificial Neural Network and Genetic Algorithm in Search Engine Optimization

Mizani Mohamad Madon[1*], **Suhaila Mohd. Yasin**[1]

[1]Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, Parit Raja, 86400, MALAYSIA

*Corresponding Author

DOI: https://doi.org/10.30880/jscdm.2023.04.01.004
Received 15 February 2000; Accepted 15 April 2023; Available online 25 May 2023

Abstract: Search engine optimization applies search principles in search engines to assign a higher ranking to the most suitable webpage. Nowadays, information searching is done ubiquitously on the World Wide Web with the help of search engines. However, the process needs to be efficient and produce accurate results simultaneously. In this research, the objectives are to implement and evaluate the Artificial Neural Network and Genetic Algorithms. The accuracy result for both algorithms is compared by implementing keyword ranking, Search Engine Result Page visibility, and time retrieval for document-based and e-commerce websites. To achieve them, firstly, the problem and data are defined. Next, two datasets are imported from Kaggle and transformed into a more helpful format. Then, the Artificial Neural Network and Genetic Algorithms are implemented on these datasets in Python using Jupyter Notebook tools. Subsequently, the accuracy of these datasets keyword ranking, Search Engine Result Page visibility, and time retrieval are observed based on the output and graph. Lastly, an analysis of the results is performed. Conclusively, the Genetic Algorithm demonstrates a higher percentage of accuracy results than the Artificial Neural Network algorithm in keyword ranking and SERP visibility. However, the accuracy results of time retrieval are vice versa. The results in Genetic Algorithm show 9.0%, 9.0%, and 3.0% in the e-commerce dataset for keyword ranking and 4.0%, 51.0%, and 1.0% in the document-based dataset for SERP visibility. Next, the Artificial Neural Network algorithm shows results of 8.0%, 7.0%, and 7.0% in the e-commerce dataset and 3.0%, 50.0%, and 4.0% in the document-based dataset for time retrieval. Therefore, the results validated the ability of the Genetic Algorithm as one of the most applied algorithms in the search engine optimization field.

Keywords: Search Engine Optimization, machine learning, Artificial Neural Network, Genetic Algorithm

1. Introduction

Since the World Wide Web (WWW) was introduced, the traditional browsing method is not acceptable anymore for users as the method could be more efficient [1]. According to Siteefy, there are around 1.17 billion websites as of now [2]. The abundant numbers of websites lead to a large volume of users and data need to be processed. Therefore, the search engine was developed. Nowadays, an extensive amount of data is used in the WWW and has been in use until now to ease the task of searching for information on the Internet [1]. Consequently, it causes the search engine to perform complex procedures of sorting information on the web pages. So, the result will come out with the most suitable web pages based on the search input given by the user. Searching tasks can be efficient if the process is smooth and the fast speed searching with accurate results is possible. By optimizing the search engine process, complex procedures can be achieved to receive a precise keyword ranking and visibility of the search engine result page (SERP).

Artificial Neural Network (ANN) is a global optimization for analog circuits or components [3]. It acts as the

**Corresponding author: author@organization.edu.co* performance evaluator in order to find an optimal design by using a separate set of data [3]. However, to get an effective result, the design capacity needs to be in a small space and the nonlinearity capacity is not strong [3].

Meanwhile, Genetic Algorithm (GA) has been proven to optimize web efficiency [4] while ANN algorithm is used to estimate either the websites rank gives good result for the test data or vice versa [5]. GA is known as a heuristic algorithm [3]. This popular optimization algorithm is still being used till now although it cannot give assurance to solve complicated optimization problem [3]. Normally GA algorithm were used in parallel computation because they are independent to each other. However, it still needs more time for reasonable coverage because they need bigger population size [3].

In this research, the Multi-layer Perceptron (MLP) classifier is implemented in both algorithms. Keras is one of the deep learning API which the model was built by applying the Sequential API and the API have a function to build the model in a sequential style. Multi-layer Perceptron (MLP) classifier is used in ANN and GA algorithm implementation to perform classification task. MLP have three layers of nodes which are the input layer, hidden layer and output layer. Hidden and output nodes are known as neurons which act as supervised learning technique [37]. Each perceptron layers in MLP have activation functions known as rectified linear unit [17]. Gender classification and emotion classification were used in MLP architecture [17]. MLP classifier was chosen because based on past research, MLP have already proven that it is the most accurate and consistent technique in machine learning classifier [37]. The MLP classifier can be implemented in any dataset that is required for binary classification.

Furthermore, a complicated process to receive accurate keyword rank data occurred as many data were produced day by day. Some users face hard situations to get relevant data on the web pages. Usually, they use in range of two to three keywords during searching in search engines [6]. The actual keyword typed in the search engine will affect the keyword ranking in the website while SEO is accomplished when the number of visitors to the website give result to the higher visibility of the search engine result page [6]. Meanwhile, the issues of time retrieval during keyword searching also occurred when the webpage took some time to return the search results. When the browsing time is increasing, the time taken to retrieve the SERP is increasing, too [7]. Furthermore, in a wide area webpage, users always take time to search required information [8]. So, the time delay during searching can give effects to user satisfaction and engagement of the webpage [8]. Therefore, to evaluate the accuracy of keyword ranking, SERP visibility and time retrieval, both ANN and GA algorithms will be implemented and compared in order to determine which algorithm is the best.

The objectives of this research are to implement ANN and GA for obtaining the accuracy of keyword ranking, SERP visibility, and time retrieval for document-based and e-commerce websites. Second objective is to evaluate and compare ANN and GA algorithms in terms of accuracy of keyword ranking, SERP visibility and time retrieval in search engine optimization (SEO) field.

The research has been carried out using Artificial Neural Networks and Genetic Algorithm by using document-based and e-commerce websites. Then, evaluate the accuracy of keyword ranking, SERP visibility and time retrieval for both domains selected. The datasets are obtained from Kaggle and use Python to carry out the research experiment.

2. Related Work

There are some techniques and algorithms implemented by researchers to evaluate the performance of SEO. Table 1 shows the implementation of various algorithms in SEO used by different researchers.

Table 1 - Related works on search engine optimization

Researchers	Techniques/ Algorithms	Summary
[9]	On-Page Optimization and Off- Page Optimization	Link building elements were used to optimize and give higher page rank level where the higher page rank result off-site will be obtained.
[5]	Artificial Neural Network and Multilayer Perceptron Neural Network (MPNN)	During dataset process, each keyword is separately searched and parsed using crawling and parser. The result shows that both ANN and MPNN technique can produce a high accuracy in predicting rank of site based on keywords which helps in increasing the rank and website validity. Therefore, it proves that ANN and MPNN algorithm has a very good performance.
[10]	Genetic Algorithm	GA act as an automatic web page categorization and updation to get optimal query requested by the user. The queries requested will explore different areas to retrieve the best classification

		performance.
[11]	Page Ranking Algorithm	Page Ranking Algorithm acts to calculate the rank using the links and contents of the web page.
[12]	HITS Algorithm	The algorithm is known as an effective algorithm used for rating and ranking documents in websites based on the link information insert by user in certain areas.
[13]	Artificial Neural Network	ANN algorithm was implemented to do prediction and optimized the system performance.
[14]	Artificial Neural Network and Genetic Algorithm	GA algorithm combined with developed ANN model to get optimal design energy performance.
[15]	Multilayer Perceptron Neural Network	The implementation of MPNN algorithm in obtaining the Diabetes Incidence prediction shows that the prediction get high accuracy with a processing time less than one second.
[16]	Panda and Penguin Algorithm	Panda Algorithm targeted less quality of content with lower SERP displayed while Penguin Algorithm focused on web spamming that affected ranking.
[17]	Convolutional Neural Networks and Multi-layer Perceptron Neural Network	CNN shows better performance than MPNN on experimenting the speaker emotion recognition.
[18]	Range-based Sequential Search (RSS)	RSS is an efficient algorithm that act as to deal for best formal in data segmentation.

Based on Table 1, there are 11 algorithms have been recently and mostly applied in SEO. Firstly, On-Page Optimization and Off-Page Optimization is one of the techniques that have been used by many researchers in SEO. On-page Optimization in SEO helps search engine crawlers to read the content of the website where a readable content means that the website has quality and give higher rank pages [19]. As for Off- page Optimization, it can be done directly on the website in order to get higher ranking. For example, through social networking, article submission, forum and blog marketing [20]. Both techniques are important to determine the success of SEO strategy which include to increase the page rank and data traffic of the website.

Next, the combination of Artificial Neural Network and Multilayer Perceptron Neural Network has been implemented to identify factors impact for rank on websites by search engine of Google. Some steps were needed to be done in order to know the webpage rank. Firstly, the number of parameters of keywords in search engine rank was determined [5]. Then, the number of words that need to be searched is written in the search engine [5]. Next, the results were parsed using crawling method and the parameters is extracted [5]. Therefore, both combined ANN and MPNN algorithms shows a good performance in terms of webpage rank in search engine Google.

Other than that, Genetic Algorithm was implemented in a document-based domain to get the optimized weights from the cited words and it shows good result [21]. GA known as search algorithm to solve optimization issues. GA also can combine with other algorithms to get better results [22]. GA algorithm is the most commonly used in search-based software engineering (SBSE) field [23]. The searching process using GA need to evolve in multiple iteration to get a better solution [24].

Next, Page Rank Algorithm also has been implemented by few researchers in search engine [25]. Page Rank Algorithm has page importance based on the occurrence of web page to calculate the web page rank. The link structure is crucial since a website's page rank increases when the number of incoming and outgoing connections is bigger. However, Page Rank Algorithm needs multiple iteration to get an accurate rank because the calculation considered not an exact answer if it occurred only once [26].

In previous work, Hyper Induced Topic Search (HITS) algorithm have been implemented using link structure on web pages in order to know the rank of the web page. HITS algorithm also used Hubs and Authority to analyze the web page structure. Based on the user query by the user, there are meaningful information in an authority page while hubs pages are used to give links for authority pages [26]. Then, both Hubs and Authority will take sample pages based on ranking and calculate using incoming and outgoing links to provide user query in an efficient time [26].

Next, Artificial Neural Network known as an efficient and accurate fitting technique [22]. ANN has self-learning function in prediction and capable to provide optimal solution in a short period of time [22]. It also acts like a black box model because ANN not considering the modeling objects itself [27]. In ANN, the first step to process is by randomly setting the weights and thresholds [28]. The input layer in ANN act as receiving training data. Then, it multiplied and merged in different ways till meet the output layer [28].

Based on research work by Zhang et al., the combination of Artificial Neural Network and Genetic Algorithm can solve optimization problems [22]. By applying ANN-GA algorithm, the optimization process can be faster than normal process [29]. ANN-GA algorithm mainly to find the best weights, find optimal solutions and give high accuracy results for ANN [24]. Furthermore, it also calculates fitness function for each chromosome in ANN-GA combined algorithm [24].

Multilayer Perceptron Neural Network is referred as a non-recursive neural network that make use of a supervised learning technique [5]. This MPNN algorithm have multi-layered network which starts from input layer to output layer. The output layer is the real answer of network. Moreover, there are also hidden layer neurons where can be obtained using attempt and error. If there are not enough neurons, it will not be able to get an accurate mapping [5]. MLP models shows better generalization performance in regression cases where it also easy to access and scalability in machine learning [15].

Next, Panda Algorithm and Penguin Algorithm are Google Algorithms that focuses on quality of user experience. Panda Algorithm which also known as filter algorithm have function as demoting the rank for low quality websites [16]. The websites affected by the Panda Algorithm with thin content, duplicate content, poor user experience, plagiarism and keyword stuffing [16]. Website pages with low quality content will give poor result in search engines. The second algorithm, known as the Penguin Algorithm, works to stop search engine spams such as keyword stuffing, deceptive linking tactics, invisible content on web pages and pirated content with a high rank and number of likes [16]. For manipulative link scheming, some tactics were implemented like link farming, page cloaking, site mirroring and URL redirection while for keyword stuffing, the techniques implemented are manipulating the HTML meta tag, create doorway pages and scraper sites [16].

Other than that, CNN is a deep neural network where usually used in PC vision while MLP have layers of feed-forward neural networks category contain activation functions. Both algorithms are deep learning techniques. In the research done by Mishra and other researchers, they found out that CNN gave better result performance than MLP [17]. The dataset used are Ravdees, Savee, Tess and Crema-d where MLP have 0.92 value for F1 score and accuracy and CNN have accuracy results of 92.283% respectively [17].

Last algorithm is Range-based Sequential Search (RSS) algorithm. RSS have function to deal with segmentation problem during data pre-processing based on multiple sources [18]. It also can discover the best format in sequence length during data segmentation [18]. In the study by Lin with other researchers, they found out that RSS have successfully can detect new attack variants in webpages [18].

Therefore, in this study, ANN is chosen because during dataset process, each keyword is separately searched and parsed using crawling and parser. The result shows that ANN technique can produce a high accuracy in predicting rank of site based on keywords which helps in increasing the rank and website validity. Therefore, it proves that ANN algorithm has a very good performance. Next, GA is chosen because GA acts as an automatic web page categorization and updation to get optimal query requested by the user. The queries requested will explore different areas to retrieve the best classification performance.

3. Methods

In this study, the research phases are divided into four phases. Fig. 1 shows the flow of the research.

3.1 Phase 1: Problem Definition and Data Definition

Firstly, the first phase in research methodology is problem and data definition. The research starts with identifying the research background about SEO, problem statements, research aim and objectives, scope of research and scope contribution. The problem identified by investigating past and most recent work that relates to the algorithms in SEO. Subsequently, the problem identified in this research is the accuracy of keyword ranking which is a challenging issue in SEO. Next, there is also the problem in obtaining the SERP visibility and time retrieval based on the user's search query.

Moreover, this research is carried out using document-based and e-commerce websites. The dataset is obtained from Kaggle [30]. For the document-based website the experiment will be conducted using Video Ranking datasets. Next, for e-commerce website the dataset from Kaggle is about keyword search by users from the Flights and Tickets SERPs and Landing Pages datasets [30].

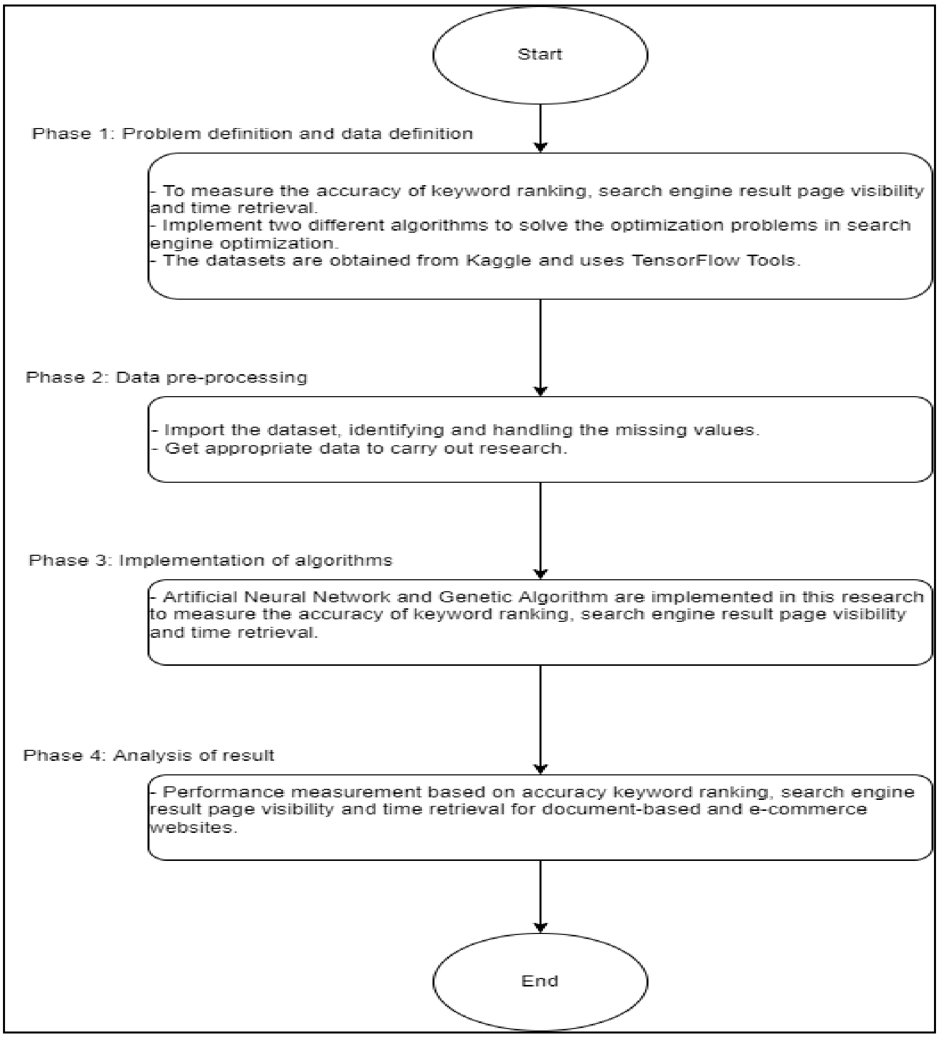
Start
Phase 1: Problem definition and data definition
- To measure the accuracy of keyword ranking, search engine result page visibility and time retrieval.
- Implement two different algorithms to solve the optimization problems in search engine optimization.
- The datasets are obtained from Kaggle and uses TensorFlow Tools.
Phase 2: Data pre-processing
- Import the dataset, identifying and handling the missing values.
- Get appropriate data to carry out research.
Phase 3: Implementation of algorithms
- Artificial Neural Network and Genetic Algorithm are implemented in this research to measure the accuracy of keyword ranking, search engine result page visibility and time retrieval.
Phase 4: Analysis of result
- Performance measurement based on accuracy keyword ranking, search engine result page visibility and time retrieval for document-based and e-commerce websites.
End

Fig. 1 - The research phases

3.2 Phase 2: Data Pre-Processing

Next, the second phase of this research is data pre-processing. Data pre-processing helps to improve the data quality in order to get the extraction of meaningful perceptions based on relevant data. To eliminate noise and resolve any discrepancies in the data, data cleaning is done. SERP visibility is measured based on user input query to retrieve the information. There is a calculation to calculate the SERP visibility score, as shown in equation (1).

$$\text{Search Volume x } 0.1935 = \text{SERP visibility [35]} \qquad (1)$$

In this research, these three parameters were used to calculate the accuracy results for keyword ranking, SERP visibility and time retrieval. The implementation of the Classification Report function was used, and the formula to calculate the accuracy, precision, recall and f1-score was shown in Equation (2), equation (3), equation (4) and equation (5).

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}\ x\ 100\ [36] \qquad (2)$$

$$Precision = \frac{TP}{(TP+FP)}\quad [36] \qquad (3)$$

$$Recall = \frac{TP}{(TP+FN)}\quad [36] \qquad (4)$$

$$F1\ score = \frac{2*(Precision*Recall)}{(Recall+Precision)}\quad [36] \qquad (5)$$

TP: True Positive
TN: True Negative
FP: False Positive
FN: False Negative

3.3 Phase 3: Implementation of Algorithms

The following phases consists of implementation of algorithms which is ANN and GA for SEO. These two algorithms are implemented to measure the accuracy of keyword ranking, SERP visibility and time retrieval.

There are some basic requirements required to conduct the research successfully. A computer with relatively sufficient processing power, memory and storage is used as the hardware. It is preferable to process power at or above 1.8GHz, with a minimum memory of 4 GB and storage at more than 28 GB. This is because the processing time is necessary for the feature selection. The study was carried out using Windows 10 and web tool environment which is Jupyter Notebook. Microsoft Word is used to do the proposal writing while Microsoft Excel is used to sort the data.

This program is implemented in Python [31]. Python commonly implement C as the programming language. Additionally, it has a sizeable standard library with modules that is focused on general programming which are OS, threading, networking and database-specific modules [31]. [32], a data scientist discovered that Python is simpler to learn since the code is more similar every day human speech. Python is also appropriate for usage in both large and small projects because it is an adaptive and advanced programming language [31]. The programmers become productive when using Python because they manage to develop the program better and efficiently. By running the commands in Python for the document-based website and e-commerce website, the dataset simply can be read in different statistical software [31].

Neural networks have recently been regarded as universal function approximation. They are free model estimators. It is feasible to estimate a function without knowing the type of the function. The problem-solving process is defined as the mapping of the issue domain, problem knowledge and solution space into the network's input state space, synaptic weights space and output space [33]. Fig. 2 shows the code segmentation for ANN algorithm.

```
split_ratio = 0.2
for rank in document["keyword rank"].unique():
    temp = document[document["keyword rank"] == rank].reset_index(drop =
True)
    train_data = len(temp) - int(len(temp) * 0.2)

    if len(temp) > 1 and rank == document["keyword rank"].unique()[0]:
        train = temp[:train_data]
        test = temp[train_data:]
    elif len(temp) > 1:
        train = pd.concat((train, temp[:train_data]), axis = 0)
        test = pd.concat((test, temp[train_data:]), axis = 0)
    else:
        pass
model = Sequential()
model.add(Dense(units=10, input_dim = X_train.shape[1], activation='relu'))
model.add(Dense(units=6, activation='relu'))
model.add(Dense(units=len(y_train.unique()) + 1, activation='sigmoid'))
model.compile(loss='sparse_categorical_crossentropy', metrics=['accuracy'])
history = model.fit(X_train_scaled, y_train, batch_size=10 , epochs=100,
verbose=1)
```

Fig. 2 - Code segmentation of ANN algorithm

Firstly, splitting the dataset variables by implementing `split_ratio = 0.2` as shown in Fig. 2. Train and test the dataset also implemented in this segment code to find a pattern that met the best data points with less error. Other than that, the implementation of Sequential Model also shown in Fig. 3. The classification was done on train data and the test data was predicted for accuracy. Next, the Sparse Categorical Cross-Entropy were used because the encoded data were in integer. Lastly, the Classification Report displayed the accuracy result with precision, recall, f1-score and support as well. Fig. 3 shows the code segmentation for GA algorithm.

```
model = Sequential()
model.add(Dense(units=10, input_dim = X_train.shape[1], activation='relu'))
model.add(Dense(units=6, activation='relu'))
model.add(Dense(units=len(y_train.unique()) + 1, activation='sigmoid'))
model.compile(loss='sparse_categorical_crossentropy', metrics=['accuracy'])
model.summary()
```

Fig. 3 - Code segmentation of GA algorithm

In Fig. 3, Sequential Model was applied when building the Keras Models. Implementation of MLP Classifier starts here using Keras. Then, Keras Genetic Algorithm module were also imported to build initial population of solutions to holds all parameters in Keras model. Each layer was created using `tensorflow.keras.layers` module. Then, the `Sequential()` class and `add()` method is used to add the layers to the model**.** In `pygad.kerasga.Kerasga` class creates 2 instance attributes which are model and num_solutions. The num_solutions assigned with value of 10 which means the population has 10 solutions. The code shown in Fig. 4.

```
keras_ga = pygad.kerasga.KerasGA(model = model, num_solutions = 10)
def fitness_func(solution, sol_idx):
    global X_train_scaled, y_train, keras_ga, model
    model_weights_matrix = pygad.kerasga.model_weights_as_matrix(model =
model, weights_vector = solution)
    model.set_weights(weights = model_weights_matrix)
    predictions = model.predict(X_train_scaled)
    bce = tf.keras.losses.BinaryCrossentropy()
    error = bce(to_categorical(y_train), predictions).numpy()
    fitness = 1.0 / (error + 0.00000001)
    return fitness
ga_instance = pygad.GA(num_generations = 100, num_parents_mating = 5,
fitness_func = fitness_func,
initial_population = keras_ga.population_weights)
ga_instance.run()
fig = ga_instance.plot_fitness()
solution, solution_fitness, _ = ga_instance.best_solution()
print("Parameters of the best solution: \n {solution}".format(solution =
solution), end = "\n\n")
#trainable parameters
print("Len of the solution is:", len(solution), end = "\n\n")
print("Fitness value of the best solution: \n
{solution_fitness}".format(solution_fitness = solution_fitness), end =
"\n\n")
```

Fig. 4 - Code segmentation of GA algorithm

Next, `model_weights_as_matrix()` function is used to restore the Keras Model's parameters from the chromosome. Fitness function as shown in Fig. 4 is a maximization function for GA algorithm. The fitness value is determined as the loss value reciprocated. In order to calculate the fitness value, there are some steps to be followed. The steps are to configure the model parameters, generate predictions, compute the loss value, calculate the fitness value and return the fitness value after recovering the model parameters from a 1-D vector. When the classification problem is in binary, binary classification is employed in the code to calculate the binary cross-entropy loss.

Next, instance is created in the `pygad.GA` class as shown in Fig. 4. The number of generations, the number of parents to mate, the starting population of Keras model paramaters, the fitness function and the generation callback function make up the minimum amount of arguments supplied to the code. The instance of `pygad.GA` class runs by calling the `run ()` method. The three parameters implemented in the code which are keyword ranking, SERP visibility and time retrieval used fitness function in Keras to train the optimized weights. The parameters of the best solution, trainable parameters and fitness value of best solution is displayed before the Classification Report displayed the results.

3.4 Phase 4: Analysis of Result

Lastly, phase 4 concerns the analysis of the results to measure the accuracy of the keyword ranking, SERP visibility and time retrieval. The experiment was conducted and the obtained results were analysed with both ANN and GA algorithms applied in this research. The accuracy results were calculated using Classification Report and equation (2), (3), (4) and (5).

Prioritizing keyword in webpage titles is one of the page elements based on SERPs. The sites listed in SERPs mostly are found in titles although the snippets, URLs and links apparently were evaluated. These three factors are the most important keyword in Google's algorithm [34]. Therefore, the web domain name, directories and files based on keywords and keyword prioritizing in titles need to be taken seriously in order to get the SERPs visibility in website.

Each parameters gives accuracy results after running the code in Python. The results between e-commerce based and document-based dataset in ANN and GA algorithm is compared in terms of accuracy in order to measure which algorithm serves the best in SEO.

4. Results

The accuracy of keyword ranking, SERP visibility and time retrieval for e-commerce based and document-based dataset were measured by using Classification Report. In Classification Report, the accuracy, precision, recall and f1-score is calculated and displayed. The accuracy is a ratio of correctly predicted observation to the total observations. The formula to calculate accuracy, precision, recall and f1-score shown in equation (2), equation (3), equation (4) and equation (5). The overall accuracy results based on implementation of ANN and GA algorithm is stated in Table 2.

Table 2 - Analysis of performance measures

Algorithm	Dataset	Accuracy of Keyword Ranking	Accuracy of SERP Visibility	Accuracy of Time Retrieval
ANN	E-commerce	8.0	7.0	7.0
	Document-based	3.0	50.0	4.0
GA	E-commerce	9.0	9.0	3.0
	Document-based	4.0	51.0	1.0

Based on Table 2, it is observed that in the e-commerce dataset and document-based dataset, GA surpasses ANN for keyword ranking and SERP visibility parameters. However, for accuracy results of time retrieval, ANN shows getting higher percentage accuracy results for both the e-commerce dataset and document-based dataset.

ANN algorithm implemented in the e-commerce dataset presents a result of 8.0% accuracy for keyword ranking, 7.0% accuracy for SERP visibility and 7.0% accuracy for time retrieval, respectively. Meanwhile, the document-based dataset presents 3.0% accuracy for keyword ranking, 50.0% accuracy for SERP visibility and 4.0% accuracy for time retrieval, respectively. Next, for GA algorithm implemented in the e-commerce dataset give result 9.0% accuracy for keyword ranking, 9.0% accuracy for SERP visibility and 3.0% accuracy for time retrieval. In the document-based dataset, it gives result 4.0% accuracy for keyword ranking, 51.0% accuracy for SERP visibility and 1.0% accuracy for time retrieval.

In this research, the keyword rank and time retrieval were calculated based on the e-commerce-based and document-based datasets given from the Kaggle website. The e-commerce-based and document-based dataset's author state that the keyword ranking in both datasets is calculated using Rank Tracer by SEO Power Suite. For SERP visibility, the dataset gave search volume and the SERP visibility is calculated using the formula as stated in equation (1). The keyword rank, SERP visibility and time retrieval is classified using MLP classifier in order to measure the accuracy of the parameters.

5. Discussion

In this research, ANN and GA have been implemented on two datasets: e-commerce and document-based website. A comparative analysis has been performed to evaluate both algorithms with respect to accuracy in search engine optimization field. Three parameters have been implemented to evaluate the accuracy of keyword ranking, SERP visibility and time retrieval.

In order to achieve the accurate results, datasets from the document-based website which is Video Ranking dataset and the e-commerce website which is Flights and Tickets SERPs and Landing Pages dataset from Kaggle were used [30]. 1882 data were implemented for both datasets. The document-based dataset chose 3 variables which are keyword ranking, SERP visibility and search time to be evaluated. For e-commerce dataset, 3 chosen variables are rank, formatted search time and SERP visibility. For SERP visibility, the search volume given in the dataset were calculated by using equation (1). Then, both datasets were implemented in ANN and GA algorithm in Python language using

Jupyter Notebook software. Lastly, the Classification Report function were used to get the accuracy results which stated in equation (2), (3), (4) and (5).

Furthermore, this research is one of the earliest attempts in SEO by implementing ANN and GA algorithm for e-commerce and document-based dataset. ANN algorithm shows that the accuracy, efficiency, classification and reorganization of large datasets are applicable. GA algorithm also shows that it gives optimal performance results in terms of accuracy in the SEO field. Furthermore, completing this research, it means that the preliminary result has been provided in order to apply SBSE in optimizing search engine results by using GA algorithms. This research has extended the knowledge in SEO where SEO is regarding helping the search engine to understand and display the best result to the user. So, by combining both different domain which is document-based and e-commerce website, new knowledge on the better performance between ANN and GA are known.

6. Conclusion

In conclusion, GA shows better performance than the ANN algorithm in terms of accuracy of keyword ranking and SERP visibility, while vice versa results in time retrieval. The accuracy results validated the ability of the GA as one of the most applied algorithms in the SEO field for keyword ranking and SERP visibility. Meanwhile, time retrieval shows that the ANN algorithm is more accurate in both datasets. This research focuses in comparing the analysis of performance measures between ANN and GA. Although GA shows better accuracy results than the ANN algorithm for keyword ranking and SERP visibility, there are some recommendations for future work. Firstly, implement and evaluate the GA with Convolutional Neural Network (CNN) algorithm. The CNN algorithm focuses primarily on media reconstruction, recommendation systems, picture and video recognition, image analysis and classification and natural language processing. Therefore, another possible comparative analysis can be done to evaluate the accuracy of GA and CNN algorithm. Next, implement the GA in the SBSE field based on accuracy results obtained from the three parameters in this research, which are keyword ranking, SERP visibility and time retrieval for black-box and white-box optimization problems. The last future work is to implement Support Vector Machine Classifier in the ANN and GA for document-based and e-commerce datasets. Then, compare the performances between both algorithms in terms of accuracy, precision, and f1-score.

Acknowledgement

The author would like to thank the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia for the support and encouragement throughout the process in this research.

References

[1] Veglis, A., & Giomelakis, D. (2020). Search engine optimization. *Future Internet*, *12(1)*, pp. 506-510. https://doi.org/10.3390/fi12010006.

[2] Huss, N. (2022). How Many Websites Are There in the World? Retrieved on April 26, 2022, from https://siteefy.com/how-many-websites-are-there/.

[3] Li, Y., Wang, Y., Li, Y., Zhou, R., & Lin, Z. (2020). An artificial neural network assisted optimization system for analog design space exploration. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and* Systems, *39(10)*, pp. 2460-2653.

[4] Ulah, A., Nawi, N. M., Sutoyo, E., Shazad, A., Khan, S. N., & Aamir, M. (2018). Search engine optimization algorithms for page ranking: Comparative study. *International Journal of Integrated Engineering, 10(6)*, 19-25. https://doi.org/10.30880/ijie.2018.10.06.00.

[5] Banaei, H., & Honarvar, A. R. (2017). Web page rank Estimation in search engine based on SEO parameters using machine learning techniques. *International Journal of Computer Science and Network Security*, *17(5)*, pp. 95-100.

[6] Matosevic, G., Dobsa, J., & Mladenic, D. (2021). Using machine learning for web pages' classification in search engine optimization. *Future Internet*, *13(9)*, pp. 1-20. *https://doi.org/10.3390/fi13010009*.

[7] Yamada, J. & Kitayama, D. (2021). The Analysis of Web Search Snippets Displaying User's Knowledge. *15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. Seoul, Korea (South): IEEE. pp. 1-8, doi: 10.1109/IMCOM51814.2021.9377354.

[8] Yan, X., Zhang, J., Elahi, H., Jiang, M & Gao, H. (2021). A Personalized Search Query Generating Method for Safety-Enhanced Vehicle-to-People Networks. *IEEE Transactions on Vehicular Technology, 70(6)*, pp. 5296-5307, doi: 10.1109/TVT.2021.3075626.

[9] Khan, M. N. A., & Mahmood, A. (2018). A distinctive approach to obtain higher page rank through search engine optimization. *Sadhana - Academy Proceedings in Engineering Sciences*. Pakistan: Institute of Science and Technology, Islamabad. pp. 1-12. *43*(3). https://doi.org/10.1007/s12046-018-0812-3.

[10] Chawla. S. (2016). A novel approach of cluster based optimal ranking of clicked URLs using genetic algorithm for effective personalized web search. *Applied Soft Computing*, *46(0)*, pp. 90-103. doi:

http://dx.doi.org/10.1016/j.asoc.2016.04.042.

[11] Sharma, D., Shukla, R., Giri, A. K., & Kumar, S. (2019). A brief review on search engine optimization. *Proceedings of the 9th International Conference On Cloud Computing, Data Science and Engineering*, Confluence 2019, Noida, India: IEEE. pp. 687-692. https://doi.org/10.1109/CONFLUENCE.2019.8776976.

[12] Singh, B. & Singh, G. (2017). A Study of Ranking Algorithm Used by Various Search Engine. *International Journal of Recent Trends in Engineering & Research*, *3(12),* pp. 201-207, doi: 10.23883/IJRTER.2017.3556.P5OAE.

[13] Ali, L., Al-Sakhnini, M. M., Kalra, D, Afzaal, F., Pervaiz, M. & Khan, M. F. (2022). Distributed Search Engine Query Optimization Using Artificial Neural Network. *2022 International Conference on Cyber Resilience (ICCR)*. Dubai, United Arab Emirates. pp. 01-05, doi: 10.1109/ICCR56254.2022.9995958.

[14] Saryazdi, S. M. E., Etemad, A., Shafaat, A. & Bahman, A. (2022). Data-driven performance analysis of a residential building applying artificial neural network (ANN) and multi-objective genetic algorithm. *Building and Environment, 225(0)*, pp. 1-17. https://doi.org/10.1016/j.buildenv.2022.109633.

[15] Song, H & Lee, S. (2021). Implementation of Diabetes Incidence Prediction Using a Multilayer Perceptron Neural Network. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Houston, TX, USA: IEEE. pp. 3089-3091, doi: 10.1109/BIBM52615.2021.9669583.

[16] Patil, A., Pamnani, J., & Pawade, D. (2021). Comparative Study of Google Search Engine Optimization Algorithms: Panda, Penguin and Hummingbird. *International Conference for Convergence in Technology, (I2CT)*. Pune, India: IEEE. pp. 1-5. https://doi.org/10.1109/I2CT51068.2021.9418074.

[17] Mishra, E., Sharma, A.K., Bhalotia, M. & Katiyar, S. (2022). A Novel Approach to Analyse Speech Emotion using CNN and Multilayer Perceptron. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. Pradesh, India: IEEE. pp. 1157-1161, doi: 10.1109/ICACITE53722.2022.9823781.

[18] Lin, Y. -D., Liu, Z. -Q., Hwang, R. -H., Nguyen, V. -L., Lin, P. -C. and Lai, Y. -C. (2022). Machine Learning with Variational AutoEncoder for Imbalanced Datasets in Intrusion Detection. *IEEE Access*, 10(0), pp. 15247-15260, doi: 10.1109/ACCESS.2022.3149295.

[19] Gupta, S., Agrawal, N., & Gupta, S. (2016). A review on search engine optimization: basics. *International Journal of Hybrid Information Technology*, 9(5), pp. 381-390. https://doi.org/10.14257/ijhit.2016.9.5.32.

[20] Dinesh, P., & SenthiMurugan, S. (2018). A survey on search engine optimization, its techniques, tools and algorithms. *International Journal of Scientific & Engineering*, 9(4), pp. 163-168.

[21] Qiu, Y., Wang, D & Yan, H. (2021). Research on Application of Genetic Algorithms in Corporal Portal Search Engines. *IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. Chongqing, Chine: IEEE. pp. 1310-1314, doi: 10.1109/IAEAC50856.2021.9390879.

[22] Zhang, Q., Deng, D., Dai, W., Li, J. & Jin, X. (2020). Optimization of culture conditions for differentiation of melon based on artificial neural network and genetic algorithm. *Scientific Reports, 10(3524),* pp. 1-8, doi: https://doi.org/10.1038/s41598-020-60278-x.

[23] Mezhuyev. V, Al-Emran. M, Ismail. M. A, Benedicenti. L and Chandran. D. A. P. (2019). The Acceptance of Search-Based Software Engineering Techniques: An Empirical Evaluation Using the Technology Acceptance Model. *IEEE Access*, 7(0), pp. 101073-101085. doi: 10.1109/ACCESS.2019.2917913.

[24] Nezamoddini, N. & Gholami, A. (2019). Integrated Genetic Algorithm and Artificial Neural Network. *IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*. New York, USA: IEEE. pp. 260-262, doi: 10.1109/CSE/EUC.2019.00057.

[25] Kushwaha, A. K., & Chopde, N. (2014). A comparative study of algorithms in SEO & approach for optimizing the search engine results using hybrid of query recommendation and document clustering, Genetic Algorithm, 5(2), pp. 1800-1802.

[26] Lemos, J. Y. and Joshi, A. R. (2017). Search engine optimization to enhance user interaction. *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).* Palladam, India: IEEE. pp. 398-402, doi: 10.1109/I-SMAC.2017.8058379.

[27] Zhao, J., Ma, Y., Zhang, Z., Wang, S. & Wang, S. (2019). Optimization and matching for range-extenders of electric vehicles with artificial neural network and genetic algorithm. *Energy Conversion and Management*, 184(0), pp. 709-725 https://doi.org/10.1016/j.enconman.2019.01.078.

[28] Alatrany, A. S., Hussain, A. J., Mustafina, J. & Al-Jumeily, D. (2022). Machine Learning Approaches and Applications in Genome Wide Association Study for Alzheimer's Disease: A Systematic Review. *IEEE Access, 10(0)*, pp. 62831-62847, doi: 10.1109/ACCESS.2022.3182543.

[29] Li, Y., Jia, M., Han, X. and Bai, X. -S. (2021). Towards a comprehensive optimization of engine efficiency and emissions by coupling artificial neural network (ANN) with genetic algorithm (GA). Energy, 225(0), pp. 1-13. https://doi.org/10.1016/j.energy.2021.120331.

[30] Dabbas, E. (2019). Flights and Tickets SERPs and Landing Pages. Retrieved on December 01, 2022, from https://www.kaggle.com/datasets/eliasdabbas/flights-serps-and-landing-pages.

[31] Ozgur, C., Colliau, T., Rogers, G., & Hughes, Z. (2021). MatLab vs. Python vs. R. *Journal of Data Science*, *15*(3), pp. 355-372. https://doi.org/10.6339/jds.201707_15(3).0001.

[32] Markham. K. (2016). Should you teach Python or R for data science? Retrieved on November 1, 2022, from: http://www.dataschool.io/python-or-r-for-data-science/

[33] Shanmuganathan, S. (2016). Artificial neural network modelling: An introduction. In Shanmuganathan, S. *Artificial Neural Network Modelling.* Auckland, New Zealand. pp. 1-14. https://doi.org/10.1007/978-3-319-28495-8_1.

[34] John, B. (2013). How to use search engine optimization techniques to increase website visibility. *IEEE Transactions on Professional Communication*, 56(1), pp. 50-66. https://doi.org/10.1109/TPC.2012.2237255.

[35] Marketing Miner (2020). What is SERP visibility. Retrieved on October 13, 2022, from https://help.marketingminer.com/en/article/what-is-serp-visibility/.

[36] Vujovic, Z. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, *12(6)*, pp. 599-606. doi: 10.14569/IJACSA.2021.0120670.

[37] Gaikwad, V., Naik, A., Rane, M. & Jalnekar, R. (2021). Daily Stock Price Direction Prediction using Random Multi-Layer Perceptron. *International Conference on Artificial Intelligence and Machine Vision (AIMV*). Gandhinagar, India: IEEE. pp. 1-9, doi: 10.1109/AIMV53313.2021.9670927.