Institut Riset dan Publikasi Indonesia (IRPI)
IJATIS: Indonesian Journal of Applied Technology and Innovation Science
Journal Homepage: https://journal.irpi.or.id/index.php/ijatis
Vol. 1 Iss. 1 February 2024, pp: 9-15
ISSN(P): 3032-7466 | ISSN(E): 3032-7474

Application of Artificial Neural Network, K-Nearest Neighbor and Naive Bayes Algorithms for Classification of Obesity Risk Cardiovascular Disease

Aulia Wulandari[1], Anggi Mulya[2], Tri Dermawan[3], Ryando Rama Haiban[4], Aghnia Tatamara[5], Habibah Dian Khalifah[6]

[1,2,3]Department of Information System, Fakulty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia,
[4]Management Information System, Faculty of Administrative Sciences from Economics,
Universitas Dokuz Eylul, Turkey
[5]Mangement Bussines, Faculty of Administrative Sciences from Economics,
Universitas Dokuz Eylul, Turkey
[6]Ushuluddin, Fakultas Syariah, Universitas Yarmouk, Jordan

E-Mail: [1]12150321995@students.uin-suska.ac.id, [2]12150312142@students.uin-suska.ac.id,
[3]12150311139@students.uin-suska.ac.id, [4]habibahdian7@gmail.com,
[5]aghniatatamara4@gmail.com, [6]ryandohaibaan@gmail.com

Received Dec 02nd 2023; Revised Dec 31th 2023; Accepted Jan 4th 2023
Corresponding Author: Aulia Wulandari

Abstract

The rate of obesity sufferers continues to increase every year. This happens due to improper lifestyle and diet, as well as various physical conditions. This research aims to analyze the level of obesity using data mining techniques with classification algorithms. This research was conducted on people from countries on the American continent between the ages of 14 and 61 years. Data is collected and information is processed using a web platform that includes surveys where anonymous users answer each question to obtain 17 attributes and 2111 records. This research uses 3 algorithms, namely the Artificial Neural Network algorithm, K-Nearest Neighbor and Naive Bayes. People who are obese are also at higher risk of experiencing health problems, such as asthma, stroke, heart disease, diabetes and cancer. The results after comparing the three algorithms, it is better to use the k-nearest neighbor algorithm compared to Artificial Neural Network and Naive Bayes because the accuracy is 95.74%. Therefore, the K-Nearest Neighbor algorithm is very suitable to use when classifying data.

Keyword: Artificial Neural Network, Classification, K-Nearest Neighbor, Naive Bayes, Obesity

1. INTRODUCTION

A person's body health is considered optimal when body weight and height are in balance. Diet has been established as one of the most important strategies for the prevention of cardiovascular disease (CVD) in the population, as well as the leading cause of death in America and worldwide [1]. Modern lifestyle changes, especially technological advances, have a major impact on diet, physical activity levels and the body's energy balance [2]. Dependence on modern technology, such as the use of electronic devices and easy access to information, has an impact on lifestyle habits and creates an environment that encourages inactive behavior and indirectly causes an increase in obesity rates.

Obesity as a global health problem continues to increase, becoming a major concern in public health research [3]. Obesity, it is widely defined as excess amounts of body fat and is a global epidemic that can have quite serious consequences, including increased risk of disease and decreased life expectancy [4]. Measuring the level of obesity can be done through parameters such as waist and hip circumference [5]. The waist to hip circumference ratio is a very significant factor related to the risk of coronary heart disease. In this research, the data mining process is carried out by finding patterns, information and important knowledge hidden in large data sets [6]. This study aims to analyze obesity levels at ages 14 to 61 years using data mining techniques with classification algorithms.

In recent years, data science has emerged as a new and important scientific field. It can be seen as a fusion of classical disciplines such as statistics, data mining, databases, and distributed systems[7]. Machine learning algorithms have been used for prediction and classification in the healthcare field[8]. Data mining is

DOI: https://doi.org/10.57152/IJATIS.v1i1.1095

a technique in data processing that aims to find relationships or patterns from data that were not previously known to the user. Data mining is also referred to as a useful method for looking for patterns, hidden trends from large amounts of data and has been widely used [9]. Data mining is divided into several categories based on the tasks that can be performed, including description, association, prediction, estimate, clustering, and classification[10].

Classification is a method used to assign a new data record to one of several previously defined categories. Classification in data mining plays a role in developing classification functions and models for an object in the data, which is adapted to the characteristics of the data. The main goal is to map objects into groups that correspond to the characteristics of each class. This research uses three classification algorithms, namely Artificial Neural Network, K-Nearest Neighbor and Naïve Bayes Classifier. Based on previous research in 2017, the Naïve Bayes Classification algorithm method has been proven suitable for use in this case [11].

Based on the problems that exist in analyzing this research data, data mining techniques are used to classify data on the level of possible risk of obesity. Classifying the level of possible risk of obesity or cardiovascular disease can be done using these three algorithms. Of the three algorithms that have been mentioned, each has different characteristics. Therefore, the aim of this research is to compare accuracy levels to determine which classification method is the best and optimal among the three algorithms.

Based on this background, this research classifies obesity datasets by comparing three algorithms, namely K-NN, Naive Bayes, and ANN. This is different from previous research which only used one algorithm. Previous research focused more on basic concepts, whereas currently, technological developments allow the application of three classification algorithms that advance this field. The use of classification algorithms in the current research provides a new dimension to data analysis. By applying these algorithms, research is not only more accurate but also more efficient in identifying underlying patterns and trends in the data.

2. MATERIAL AND METHOD

2.1. Artificial Neural Network

ANN is a mathematical model inspired by the structure and function of human biological neural networks. It is used in machine learning to process information and perform certain tasks. ANN is considered an ideal solution for modeling systems with non-linear relationships and is often used for both regression and classification problems. The main advantage of ANN lies in its ability to produce solutions quickly and reliably, even in data sets that contain noise or missing information [12][13]. ANN is a form of data storage system as well as an algorithm that stores the functions of biological neurons. ANN improves its performance by learning from data in the training step. ANN is an accurate and fast tool for solving complex problems in areas such as science, engineering, and manufacturing [18].

The working principle of ANN takes inspiration from the function of artificial neural network systems in humans. Generally, data in ANN is broken down into three different subgroups [20]:

1. Training: The training process involves using a subset of data to train an ANN, where learning occurs through examples, similar to the functioning of the human brain. Training sessions are repeated iteratively until an acceptable level of model precision is achieved.
2. Validation: This subset determines the range of the trained model and the estimation of model characteristics, such as classification error, mean error for numerical estimation, and so on.
3. Testing: This subgroup is useful for confirming the performance of the training subset that has been applied to the ANN model.

The following is a simple formula used in the classification of neural network algorithms, as for the formula is as follows:

$$N_1 \qquad = (X_1 \times W_1 + X_2 \times W_2 + \dots \ \mathrm{X}n \times Wn) - \mathrm{b} \tag{1}$$

$$Out \qquad = (N_1 \times W_1 + N_2 \times W_2 + \dots \ \mathrm{N}n \times Wn) - \mathrm{b} \tag{2}$$

Information :

N	: Neuron
X	: Input
W	: Weight
Out	: Output
b	: Bias

2.2. Naive Bayes

Naive Bayes is one of the most popular data mining classification algorithms. The Naive Bayes algorithm is a classification method based on Bayes' probability theorem with the assumption that each feature of data is independent of each other. Naïve Bayes is a very competent classifier in many real-world applications[14]. The Naive Bayes algorithm is widely used in text classification, sentiment analysis, and system suggestions . Naive Bayes often provides good performance and is relatively fast in most cases. The Naïve Bayes algorithm is used to help classify classes or levels of public sentiment [15].

Naive Bayes is a data mining classification algorithm calculated based on the probability theorem. Even though it is simple, this algorithm is often used because it is very effective and fast. Bayes' theorem is a mathematical formula for determining probability conditions. This theorem is named after the English mathematician Thomas Bayes in the 18th century [19].

$$P(A \mid B) = \frac{P(B \mid A) . P(A)}{P(B)} \tag{3}$$

Information :

B : Data with unknown classes
A : Data hypothesis B is a specific class
P(A|B) : Probability of hypothesis A based on condition B
P(A) : Probability of hypothesis A
P(B|A) : Probability B is based on these conditions
P(B) : Probability of B

2.3 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is recognized as one of the simplest non-parametric supervised classification approaches and has been around for a long time in the machine learning reading literature. By determining a specific k value in the entire data set, the mean/mode class of the nearest neighbors is identified, and the new object is attributed to the class closest to its neighbors. This method shows a robust structure to dense training data when the k value is large enough. As the data set and size k grow, the required processing time increases rapidly, and in this approach, all distance calculation results must be stored in memory. Therefore, the choice of k value becomes important in the balance between accuracy and efficiency.

$$dis\ (x_1, x_2)\ =\ \sqrt{\sum_{i=0}^{n}\quad (x_{1i} -\ x_{2i})^2} \tag{4}$$

2.4. Google Colab

Google Colab is also known as the free Jupyter Cloud computing platform provided by Google. Used by writing explanatory text and Python code via a web browser and also widely used to teach Machine Learning. This makes it easy to train machine learning models that require high computing power without requiring special hardware. Experiment sharing can be accessed on the Web via Google Colab [16]. In addition, Google Colab also allows users to create and share documents containing narrative text, images, code, and mathematical formulas.

2.5. Obesity

Obesity is excess body fat, is a global epidemic that can lead to serious health problems, including increased mortality and decreased quality of life. Between 1980 and 2015, obesity rates doubled in more than 70 countries. Recent data shows that worldwide, about 108 million children and 604 million adults are obese [17]. The emergence of obesity is influenced by various factors such as environment, lifestyle and genetic factors, but it is also closely related to various diseases such as heart disease.

2.6. Methodology

The planning stage used in this research is data on the level of possible risk of obesity or cardiovascular disease. The planning stage includes three stages, namely identifying problems, determining research objectives, and determining research boundaries. Next, the data collection stage is carried out using the data that will be used and conducting a literature study. The next stage of data processing is preprocessing and dividing training data and test data using Artificial Neural Network, K-Nearest Neighbor and Naïve Bayes. In the Analysis and Results process, accuracy calculations are carried out between the Artificial Neural Network, K-Nearest Neighbor and Naïve Bayes on the classification results and the final analysis is carried out. Finally, documentation of the research is carried out. The following is the Research Methodology in Figure 1 below.

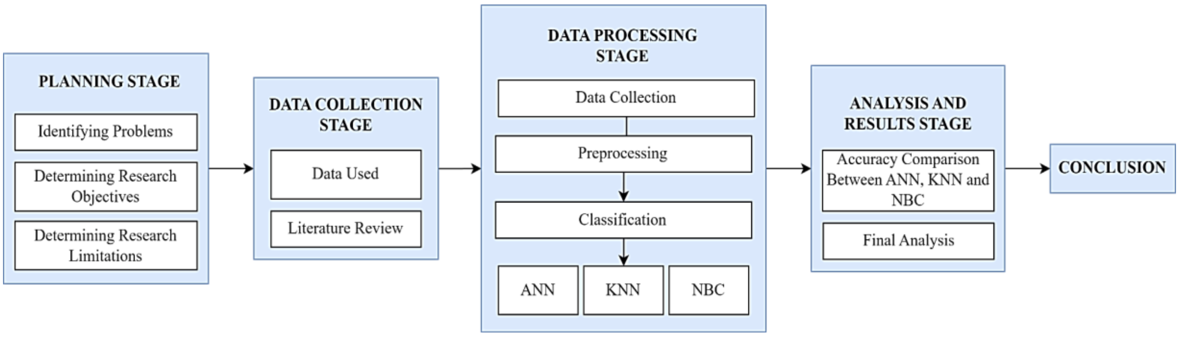

Figure 1. Research Methodology

3. RESULTS AND DISCUSSION

The results and analysis of the data consist of estimates of obesity levels the age range between 14 to 61 years as well as various eating habits and physical conditions.Data is collected through a web platform using surveys, where anonymous users provide answers to each question. Then, this information is processed to get 2111 records with 17 attributes.

3.1. Artificial Neural Network

In this research, the classification process was carried out using an Artificial Neural Network algorithm model. The results of testing the Artificial Neural Network algorithm show that the classification accuracy level reached 65.61%. This shows that the Artificial Neural Network algorithm is not appropriate to use for classification. The following are the results of testing the Artificial Neural Network algorithm.

Table 1. Results from analysis of the Artificial Neural Network algorithm model

	True Insufficient Weight	True Normal Weight	True Obesity Type I	True Obesity Type II	True Obesity Type III	True Overweight Level I	True Overweight Level II
Pred. Insufficient Weight	82	8	0	0	0	0	0
Pred. Normal Weight	13	66	0	0	0	7	1
Pred. Obesity Type I	0	0	58	6	37	0	2
Pred. Obesity Type II	0	0	0	0	90	0	0
Pred. Obesity Type III	0	0	0	0	101	0	0
Pred. Overweight Level I	0	14	1	0	0	33	33
Pred. Overweight Level II	0	0	6	0	0	2	74

From the analysis results table above, it can be seen that the Artificial Neural Network algorithm produced is the level of obesity that occurs in the age range between 14 to 61 years and physical condition and various eating habits with an accuracy of 65.61%, and the classification carried out by this algorithm can be said to be not good and the precision is 63.93%.

3.2. Naïve Bayes

In this research, the classification process was carried out using the Naive Bayes algorithm model. The results of testing the Naive Bayes algorithm show that the classification accuracy level reaches 62.30%. This shows that the Naive Bayes algorithm is not appropriate to use for classification. The following are the results of testing the Naive Bayes algorithm.

Table 2. Results from analysis of the Naive Bayes algorithm model

	True Insufficient Weight	True Normal Weight	True Obesity Type I	True Obesity Type II	True Obesity Type III	True Overweight Level I	True Overweight Level II
Pred. Insufficient Weight	65	5	13	0	0	5	0
Pred. Normal Weight	28	32	9	0	0	8	6
Pred. Obesity Type I	0	2	64	33	0	1	5
Pred. Obesity Type II	0	0	4	87	0	0	1
Pred. Obesity Type III	0	0	0	0	104	0	1
Pred. Overweight Level I	2	3	45	0	0	24	6
Pred. Overweight Level II	1	8	42	10	0	1	19

From the analysis results table above, it can be seen that the Naive Bayes algorithm produced is the level of obesity that occurs in the age range between 14 to 61 years and physical condition and various eating habits with an accuracy of 62.30%, and the classification carried out by this algorithm can be said to be not good and the precision is 63.76%.

3.3. K-Nearest Neighbor

In this research, the classification process was carried out using the K-Nearest Neighbor algorithm model. The results of testing the K-Nearest Neighbor algorithm show that the classification accuracy level reaches 95.74%. This shows that the K-Nearest Neighbor algorithm is very appropriate to use for classification. Following are the results of testing the K-Nearest Neighbor algorithm.

Table 3. Results from analysis of the K-Nearest Neighbor algorithm model

	True Insufficient Weight	True Normal Weight	True Obesity Type I	True Obesity Type II	True Obesity Type III	True Overweight Level I	True Overweight Level II
Pred. Insufficient Weight	90	0	0	0	0	0	0
Pred. Normal Weight	6	71	0	0	0	9	1
Pred. Obesity Type I	0	0	101	1	0	0	1
Pred. Obesity Type II	0	0	0	87	3	0	0
Pred. Obesity Type III	0	0	0	0	101	0	0
Pred. Overweight Level I	0	2	0	0	0	78	1
Pred. Overweight Level II	0	0	2	0	0	1	79

From the analysis results table above, it can be seen that the K-Nearest Neighbor algorithm produced is the level of obesity that occurs in the age range between 14 to 61 years and physical condition and various eating habits with an accuracy of 95.74%, and the classification carried out by this algorithm can be said to be very good and precision of 95.71%. These results show that the K-Nearest Neighbor algorithm is the right choice when classifying data.

3.4. Comparison of Classification Artificial Neural Network, Naive Bayes & K-Nearest Neighbor algorithm

The comparison between Artificial Neural Network, Naive Bayes, and K-Nearest Neighbor algorithms can vary depending on the dataset, model configuration, and evaluation metrics used. The following is a general description of the comparison of the three algorithms in terms of accuracy, precision and recall. The results of the three algorithms show Artificial Neural Network with accuracy of 65.61%, precision of 63.93%% and recall of 64.16%. Naive Bayes algorithm with accuracy results of 62.30%, precision 63.76% and recall 60.06%. while the K-Nearest Neighbor algorithm resulted in an accuracy of 95.74%, precision of 95.71% and recall of 95.56%.

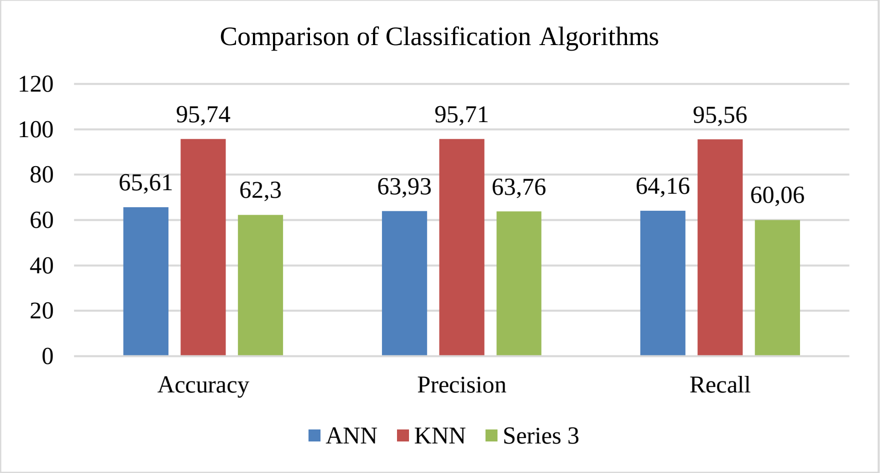

Figure 2. Comparison of Classification Algorithms

Artificial Neural Network algorithms have the ability to handle complex problems and can learn complicated patterns. However, its performance is highly dependent on large amounts of data and proper model configuration. While the Naive Bayes algorithm can be effective for datasets with features that are not very related to each other. However, it is assumed that the features are independent, which can be an unrealistic assumption in some cases.

K-Nearest neighbor can provide good results if the data has a clear spatial structure and is not too complex. In this case, obesity data shows the highest accuracy of 95.74%. The results of these three algorithms are highly dependent on the specific requirements and characteristics of the classification problem at hand. Preferably, before selecting an algorithm, it is important to understand the nature of the dataset and engage in careful experimentation to determine the most suitable model.

4. CONCLUSION

Based on the results of the research and discussion that has been carried out, it can be concluded that among the Naive Bayes, k-nearest neighbor, and artificial neural network algorithms used in this case study to classify the level of possible risk of obesity or cardiovascular disease, states that the k-nearest neighbor classification algorithm is the most suitable classification algorithm to use because it has the highest level of accuracy, precision and recall compared to the naive bayes classification algorithm and artificial neural network, which reaches 95%. In addition, the K-Neighbor algorithm is easy to implement because considering the simplicity and accuracy of the algorithm, KNN is one of the first classifiers that novice data scientists should learn. However, KNN does not work well on large datasets because it tends to reduce the performance of the algorithm.

REFERENCES

[1] Z. Shan *et al.*, “Association Between Healthy Eating Patterns and Risk of Cardiovascular Disease,” *JAMA Intern Med*, vol. 180, no. 8, pp. 1090–1100, Aug. 2020, doi: 10.1001/JAMAINTERNMED.2020.2176.

[2] J. P. Chaput, “Sleep patterns, diet quality and energy balance,” *Physiol Behav*, vol. 134, no. C, pp. 86–91, 2014, doi: 10.1016/J.PHYSBEH.2013.09.006.

[3] R. N. Haththotuwa, C. N. Wijeyaratne, and U. Senarath, “Worldwide epidemic of obesity,” *Obesity and Obstetrics*, pp. 3–8, Jan. 2020, doi: 10.1016/B978-0-12-817921-5.00001-1.

[4] C. Cercato and F. A. Fonseca, “Cardiovascular risk and obesity,” *Diabetol Metab Syndr*, vol. 11, no. 1, pp. 1–15, Aug. 2019, doi: 10.1186/S13098-019-0468-0/TABLES/1.

[5] A. Y. A. A. Baioumi, “Comparing Measures of Obesity: Waist Circumference, Waist-Hip, and Waist-Height Ratios,” *Nutrition in the Prevention and Treatment of Abdominal Obesity*, pp. 29–40, Jan. 2019, doi: 10.1016/B978-0-12-816093-0.00003-3.

[6] H. Wang, “Analysis and Prediction of CET4 Scores Based on Data Mining Algorithm,” *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/5577868.

[7] H. Wickham, M. Çetinkaya-Rundel, and G. Grolemund, “R for data science : import, tidy, transform, visualize, and model data,” p. 548, Accessed: Jan. 01, 2024. [Online]. Available: https://books.google.com/books/about/R_for_Data_Science.html?hl=id&id=xU-gzwEACAAJ

[8] E. H. A. Rady and A. S. Anwar, “Prediction of kidney disease stages using data mining algorithms,” *Inform Med Unlocked*, vol. 15, p. 100178, Jan. 2019, doi: 10.1016/J.IMU.2019.100178.

[9] A. N. Arbain and B. Y. P. Balakrishnan, “A Comparison of Data Mining Algorithms for Liver Disease Prediction on Imbalanced Data,” *International Journal of Data Science and Advanced Analytics*, vol. 1, no. 1, pp. 1–11, Feb. 2019, Accessed: Jan. 01, 2024. [Online]. Available: https://ijdsaa.com/index.php/welcome/article/view/2

[10] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, and J. Li, “A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis,” *Energy and Built Environment*, vol. 1, no. 2, pp. 149–164, Apr. 2020, doi: 10.1016/J.ENBENV.2019.11.003.

[11] T. A. Welborn and S. S. Dhaliwal, “Preferred clinical measures of central obesity for predicting mortality,” *Eur J Clin Nutr*, vol. 61, no. 12, pp. 1373–1379, Dec. 2007, doi: 10.1038/SJ.EJCN.1602656.

[12] M. M. Saritas and A. Yasar, “Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification,” *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, pp. 88–91, Jun. 2019, doi: 10.18201//ijisae.2019252786.

[13] Ü. Ağbulut, A. E. Gürel, and Y. Biçen, “Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison,” *Renewable and Sustainable Energy Reviews*, vol. 135, p. 110114, Jan. 2021, doi: 10.1016/J.RSER.2020.110114.

[14] S. Chen, G. I. Webb, L. Liu, and X. Ma, “A novel selective naïve Bayes algorithm,” *Knowl Based Syst*, vol. 192, p. 105361, Mar. 2020, doi: 10.1016/J.KNOSYS.2019.105361.

[15] M. Wongkar and A. Angdresey, “Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter,” *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*, Oct. 2019, doi: 10.1109/ICIC47613.2019.8985884.

[16] M. Canesche, L. Bragança, O. P. V. Neto, J. A. Nacif, and R. Ferreira, “Google Colab CAD4U: Hands-on cloud laboratories for digital design,” *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 2021-May, 2021, doi: 10.1109/ISCAS51556.2021.9401151.

[17] S. Sarma, S. Sockalingam, and S. Dash, “Obesity as a multisystem disease: Trends in obesity rates and obesity-related complications,” *Diabetes Obes Metab*, vol. 23 Suppl 1, no. S1, pp. 3–16, Feb. 2021, doi: 10.1111/DOM.14290.

[18] S. Khatir, S. Tiachacht, C. Le Thanh, T. Q. Bui, and M. Abdel Wahab, “Damage assessment in composite laminates using ANN-PSO-IGA and Cornwell indicator,” *Compos Struct*, vol. 230, p. 111509, Dec. 2019, doi: 10.1016/J.COMPSTRUCT.2019.111509.

[19] M. M. Saritas and A. Yasar, “Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification,” *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, pp. 88–91, Jun. 2019, doi: 10.18201//ijisae.2019252786.

[20] N. Aalimahmoody, C. Bedon, N. Hasanzadeh-Inanlou, A. Hasanzade-Inallu, and M. Nikoo, “BAT Algorithm-Based ANN to Predict the Compressive Strength of Concrete—A Comparative Study,” *Infrastructures 2021, Vol. 6, Page 80*, vol. 6, no. 6, p. 80, May 2021, doi: 10.3390/INFRASTRUCTURES6060080.