Jurnal Teknik Ilmu dan Aplikasi
ISSN: 2460-5549 | E-ISSN: 2797-0272

H2O ALGORITHM FOR JATROPHA CURCAS DISEASE IDENTIFICATION WITH FEATURE SELECTION USING GENETIC ALGORITHM

Rahmat Ramadhani[1)], Triando Hamonangan Saragih[2)] , Muhammad Haekal[3)]
Department of Computer Science, Faculty of Mathematics and Natural Science,
Lambung Mangkurat University, Indonesia
Jalan Ahmad Yani
[1)]*rahmat.ramadhani@ulm.ac.id*
[2)] *triando.saragih@ulm.ac.id*
[3)] *muhammadhaekal77@gmail.com*

Abstract

Jatropha curcas is a plant that can be used as a substitute for diesel fuel. Lack of knowledge of farmers and the limited number of experts and extension agents into the problem of dealing with the disease Jatropha curcas plant which resulted in lower quality of Jatropha curcas. H2O Algorithm can be used for Jatropha Curcas disease identification. Based on previous research, H2O Algorithm gave 96.066%. In this research, we used Genetic Algorithm to do feature selection. H2O algorithm with feature selection gave average accuracy 97.03%, that means were better than without feature selection. The parameters that we got are number of populations 600, crossover rate 0.8 and mutation rate 0.2, and number of iterations 400. However, the time spent using feature selection is so longer than without feature selection.

Keywords: Classification, Feature Selection, Genetic Algorithm, H2O algorithm, Jatropha curcas.

1. INTRODUCTION

Jatropha Curcas is a plant that has various functions that are often used as a wound medicine and the leaves are used for the treatment of malaria. These plants can also be used to save from erosion, manage soil degradation, and grow soil fertility [1]. In recent years, Jatropha Curcas has been quite well known for its use as a supplier of vegetable oil as an alternative to petroleum, and especially in the manufacture of biodiesel [2]. There has been a lot of research on this plant as an alternative to biofuels. Currently, Jatropha Curcas can be traded as a raw material for the treatment of various diseases, including cancer, skin diseases, respiratory, and infectious diseases [3].

Various kinds of diseases that attack Jatropha Curcas can reduce the quality of Jatropha Curcas produced [4]. Lack of expertise and farmer information about Jatropha Curcas gives poor results for Jatropha Curcas. Problems that are not resolved as quickly as appropriate have a negative impact on the quality of Jatropha Curcas. This problem can be solved by using an expert system. An expert system is a system that adopts expert knowledge then correctly entered into a computer and then the computer can provide solutions to problems like an expert [5]–[7].

Previous research conducted by Mazdadi et al used the H2O Algorithm to identify Jatropha disease. In this study, the results obtained an average accuracy of 96.066% [8]. The results obtained are good, but feature selection is needed to improve accuracy results and remove some features that can cause accuracy to be not optimal.

One method to perform feature selection is the Genetic Algorithm. Genetic Algorithm is an algorithm that uses the concept of searching based on the nearest neighbor solution to get the optimal solution. Genetic Algorithms are often used in optimization to feature selection. Rostami et al conducted a feature selection study using the Genetic Algorithm, PSO, Ant Colony and Bee Colony methods. They stated that the results of the Genetic Algorithm were better than other optimization methods when performing feature selection [9]. Another study conducted by Schulte et al. also performed feature selection on lower limb pattern recognition. In this study, Schulte et al performed feature selection using the Genetic Algorithm which resulted in lower errors than without feature selection [10].

Based on the explanation that has been explained, this research will use the classification of Jatropha plant disease using the H2O Algorithm method by selecting features using the Genetic Algorithm. This research is expected to provide better results without using feature selection.

The detailed methodology and experimental setup are described in Section 2. Section 3 will be discusses the results, then the conclusion is presented in Section 4.

2. MATERIALS AND METHODS

The proposed methodology aims to improve H2O Deep Leaning with genetic algorithm as feature selection method to select important feature with high correlation through the training section in classification process. the genetic algorithm method consist of 3 main process after creating population as operators: 1. Crossover, 2. Mutation, 3. Selection. In this research we used Jatropha Curcas disease as main dataset.

Dataset

Jatropha disease arises due to the presence of live pathogens [4]. The types of pathogenic fungi that attack Jatropha include Helminthosporium tetramera, Pestalotiopsis paraguarensis, P. Vesicolor, Cercospora jatrophaecurces, Phutophthora spp, Pythium spp., Fusarium spp, Dothiorella sp, Colletotrichum sp., O., Alternaria sp, Fusarium sp, Xanthomonas sp, J. Gossypiellap, and Armillaria tabescens [11]. This pathogenic variety causes various diseases, such as leaf spot, root rot, and others. Information about Jatropha disease and some of its symptoms can be seen in Table 1 [4].

TABLE 1. EFFECTS AND CAUSES OF JATROPHA DISEASE

Diseases	Pathogen	Symptoms	Causes
Antraknosa	*Colletotrichum gloeosporioides* Fungi	- Leaves or fruit become damaged. - Sprout shoots off.	- Brown round spots are restricted yellow halo. - If attacking the edge of the leaves are irregular spots. - Blackish brown spots on the fruit surface
Bacterial Blight	*Xanthomonas campestris*.		- Aqueous spots bordering leaf repeats to form angled spots. - Blackish spots on the leaves. - Under the surface, leaves look shiny
Fusarium Wilt	*Fusarium spp*. Fungi	Plants become dead	- Plant withered with yellowish leaves. - If the stem is defended will look the part of the woody brown ribbed.
Dieback	Not yet known		- The rot starts from the tip/top of the plant. - Leaves fall and stems look bare. - Side shoots cannot grow because the branches rot. - The rotten part is usually watery and the shoots dry out. - If the split part will be seen the vessels and brown pith.
Bacterial Blight	Xanthomonas campestris.		- Aqueous spots bordering leaf repeats to form angled spots. - Blackish spots on the leaves. - Under the surface, leaves look shiny

Diseases	Pathogen	Symptoms	Causes
Charcoal Roat	Rhiczoctonia bataticola Fungi	May cause sprouts to die before or after surface.	- The leaves wither in all parts of the plant suddenly. - The leaves wither yellowing at the bottom of the plant and fall out. - Root looks blackish.
Powdery Mildew	Pseudoidium jatrophae Fungi	- Leaf fall or shoot does not develop and die. - Young fruits usually change shape and fall.	The presence of white powdery mildew on the leaves, fruit, and stems when they are still young or shoot.

H2O Algorithm

H2O is a software for machine learning and data analysis, which can work on processes such as trees, linear models, unsupervised learning, etc. H2O is developed open source. The purpose of developing H2O is to make it easier for users to use, such as scaling big data, well-documented coding to support commercial processes, running on third-party systems, and having extensive programming language support [12].

One example of the use of H2O is in a study conducted by Domingos et al. namely the process to identify anomalies in IT infrastructure purchases. In this study, the method and modeling based on CRISP-DM were tested using the H2O Tool. The results of this study showed the MSE of 0.0012775 [13]. In another study entitled "Superchord: decoding EEG signals in the milisecond range" by Normand and Ferreira, (2015) used H2O.Ai as a classification algorithm. The use of H2O.Ai aims to build a model which will then be tested on a validation dataset. In this study, the average accuracy was above 80% for 109 subjects. Research by Lopes et al. to predict recovery of credit operations by implementing an integrated platform between H2O.ai and R programming, which has advantages in grid mode and parallel processing model. In this study, the results of the evaluation of the ROC graph were of high value, which had an average accuracy of above 90% [14].
The H2O algorithm is based on a multi-layer feed-forwarding artificial neural network trained with stochastic gradient descent using backpropagation. The network can contain a large number of hidden layers consisting of neurons with tanh, rectifier, and maxout activation functions. Each compute node trains a copy of the global model parameters on its local data by multi-threading (asynchronously) and periodically contributes to the global model by averaging the model across the network [15].

Genetic Algorithm

The genetic algorithm is an algorithm inspired by Darwin's theory of evolution which is a relatively new and popular software paradigm [16]. This genetic algorithm method can be applied in data mining systems to classify data in order to obtain useful information and improve the preformance in data mining [17].

The advantage of this genetic algorithm is that it is able to handle complex and parallel problems. This method handles various optimizations depending on whether the objective function is linear or not, balanced or not, continuous or not or with random noise [18].

This genetic algorithm includes coding the optimization function as an array containing bits or

characters in the form of a string to describe the chromosomes, manipulation of string operations with genetic operators, and selection according to fitness which aims to find the best solution and optimization encountered.
Representation of the chromosome were used that using binary representation. There are 30 genes in one chromosome. Each gene has a value of 0 and 1 representing every feature of jatropha curcas plant diseases. Fig. 1 shows an example of chromosome representation.

1	2	3		10	11	12	13		21	22	23		30

Figure 1. Chromosome representation.

Gi are representation of feature of Jatropha curcas plant diseases, Gi value at 0 and 1 and the value of i is between 0 and 1. The Jatropha curcas plant diseases poblem there are 30 criteria and 9 type of illness.
In the selection process using the fitness value derived from the value of the accuracy of the calculation based on the Dempster-Shafer belief contained in each chromosome. There are 50 examples of cases that are used for the calculation of fitness value using Eq. (1).

$$fitness = \frac{the\ number\ of\ cases\ that\ is\ true}{total\ number\ of\ cases} (1)$$

In this stage, to produce offspring. The method used is crossover and mutation. This process relies on the crossover rate and mutation rate are included. In this paper, crossover method used one-cut point and mutation method used random mutation [21]. A one-cut point crossover process is done by selecting two individuals and select one point to randomly take the left from the first individual or P1 and the right of the second individual or P2 to form a new individual. This process are described on Fig. 2.

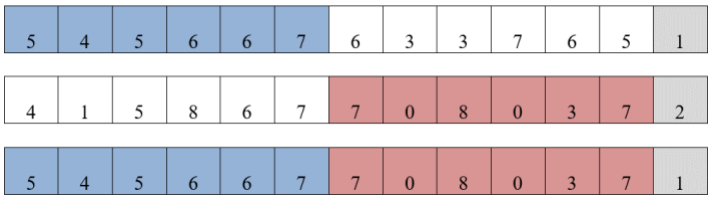

Figure 2. One-cut point crossover

While a random mutation process is done by selecting one individuals to randomly from all individuals and then select two point to randomly, exchange to form a new individual, then the illustration shown at Fig. 3.

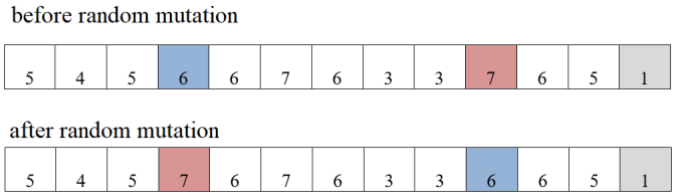

Figure 3. Random mutation

Selection is the stage at which the selection to get the best fitness value. Selection method that be used on this research is the Selection elitism which took the best individuals based on all the existing population.
In the process accuracy testing used the value of belief that has been optimized. Accuracy testing of data uses 166 test cases. If the system is issuing more than one decision and worth valued properly, the properly value were used that one divided by the number of decisions issued by the system as shown in Eq. (2).

$$accuracy = \frac{the\ number\ of\ cases\ that\ is\ true}{total\ number\ of\ cases} (2)$$

Operator
The main operators of genetic algorithms are:

1. Crossover is the process of swapping parts based on solutions (chromosomes) using other "parent" parts to form an asynchronous type of chromosome that may be a new solution to solving problems. Its main role is to put solution mixing and convergence in subspaces (forming new solutions).
2. Mutation is a change in one part of the solution chosen at random, which increases the variability based on the population and forms a procedure.
3. Selection of fitness or elitism, namely the use of solutions with high fitness values to pass to the next generation, which is often done in terms of several forms of selection of the best solutions[1][2].

Working Process

The working process of Genetic Algorithm for feature selection is as follows:

1. Initialization parameter.
2. Generate random first generation
3. Evaluate the fitness value of each chromosome in the population.
4. Generate a new population using the following process:
    a. Selection: Take two parent chromosomes from the existing population
    b. Crossover: Do crossover against two parent chromosomes to produce new offspring
    c. Mutation: Offspring formed from the existing parent mutations
5. Obtain a new population in the next generation.
6. Repeat the process again from the beginning to find the desired needs.

3. RESULT AND DISCUSSION

In this study, several tests were carried out, namely the number of populations testing, the combination of crossover and mutation rates testing, and the number of iterations testing. This test aims to determine the optimal parameters to produce the best generation in the optimization.

In testing conducted using population every multiple of 100 starting from the number 10. Rated crossover rate and mutation rate were used that 0.9 and 0.1 and the number of iterations as many as 10. The results of these tests can be seen in Figure 4.

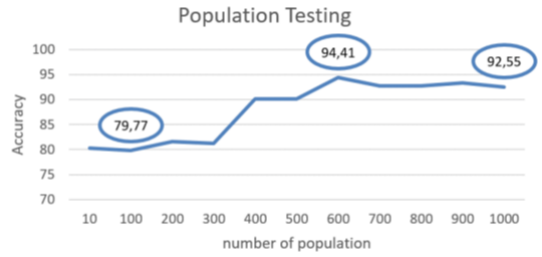

Figure 4. The results of populations testing.

The results of the population testing in Figure 4 indicates that the most optimal results possessed a population of 600 with an average value of 94.41% accuracy. The increasing number of the population are increasingly making the value of the accuracy of the system is declining.

In the test based on the value of crossover rate and mutation rate used to determine the value of crossover and mutation rate optimal as the best solution in this optimization. Population values used are 600 because it has best average accuracy. The number of iterations used as many as 10. The results are shown in Figure 5.

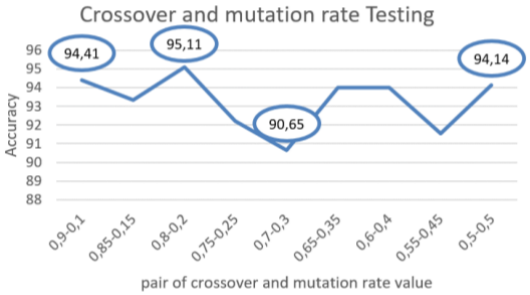

Figure 5. The result of combination crossover and mutation rate testing

In the testing based on the value of crossover rate and mutation rate for a total population of 600, said that a value of cr is 0.8 and mr is 0.2 had the highest average accuracy of 95.11%. The number of iterations testing aims to find value in the number generation has optimal results in this optimization. Iteration testing used multiple value 100 starts at a value of 10 to 2000. The results of the number of iterations testing can be seen in Figure 6.

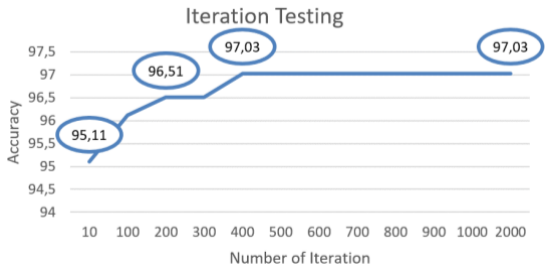

Figure 6. One-cut point crossover

Based on Figure 6 for the result test obtained iteration on the optimal value generation 400 with average accuracy 97.03%. At iteration of grades 10 to 400, an increase accuracy value, while the value of 500 to 20000 indicates the value of accuracy is stable and equal to the value of accuracy in the 400th generation. This causes an early convergent. Increasing number of iterations provides a long time in computing and does not always give better accuracy.

In the next step then compared with previous result. The results of the comparison with previous studies can be seen in Table 2.

TABLE 2. COMPARISON WITH PREVIOUS RESULTS.

Method	Average Accuracy	Time Spent
H2O Algorithm [3]	96.066%	1 sec
H2O Algorithm with Feature Selection	97.03%	4 hr 24 min 51 sec

Based on the comparison of results with previous studies in Table 2, it can be seen that there was an increase of almost 1%. However, to get

these upgrades it takes longer hours so that the improvements you get feel like a lot of sacrifices.

In the next step then compared with previous research that had been done. The results of the comparison with previous studies can be seen in Table 3.

TABLE 3. COMPARISON WITH OTHER METHODS.

Method	Average Accuracy
Extreme Learning Machine [4]	60.61%
Extreme Learning Machine with Simulated Annealing Optimization [5]	62.5%
Extreme Learning Machine with Simulated Annealing Optimization and Decision Tree [5]	90.955%
H2O Algorithm [3]	96.066%
H2O Algorithm with Feature Selection	97.03%

Based on the comparison of results with previous studies in Table 3, we can see that the H2O Algorithm method with feature selection has the best accuracy compared to the methods used in previous studies. This proves that feature selection has a good effect on increasing accuracy by removing features that are considered to give poor results when doing classification.

4. Conclusions

Based on the overall results of this study, the best average accuracy results were obtained with a value of 97.03% with the parameter population number 600, crossover rate 0.8 and mutation rate 0.2, and iteration number 400. Based on comparison with previous result, we can claim that selection feature gave better result than without result. However, the time spent of using Genetic Algorithm for feature selection took longer hours than without feature selection. For next research, we suggest to do feature selection that give less time spent like Simulated Annealing [22]–[24], Harmony Search [25], [26], etc.

5. Acknowledgements

This research is supported by the Department of Computer Science, Faculty of Mathematics and Natural Science, Lambung Mangkurat University, Indonesia.

6. References

[1] W. F. Abobatta, "Jatropa curcas: an overview," *J. Adv. Agric.*, vol. 10, pp. 1650–1656, 2019.

[2] A. J. King, W. He, J. A. Cuevas, M. Freudenberger, D. Ramiaramanana, and I. A. Graham, "Potential of Jatropha curcas as a source of renewable oil and animal feed," *J. Exp. Bot.*, vol. 60, no. 10, pp. 2897–2905, 2009, doi: 10.1093/jxb/erp025.

[3] H. A. Abdelgadir and J. Van Staden, "Ethnobotany, ethnopharmacology and toxicity of Jatropha curcas L. (Euphorbiaceae): A review," *South African J. Bot.*, vol. 88, pp. 204–218, Sep. 2013, doi: 10.1016/j.sajb.2013.07.021.

[4] T. Yulianti and N. Hidayah, *Jatropha Curcas Disease*. Malang: Balai Penelitian Tanaman Pemanis dan Serat, 2015.

[5] T. H. Saragih, D. M. N. Fajri, and A. Rakhmandasari, "Comparative Study of Decision Tree, K-Nearest Neighbor, and Modified K-Nearest Neighbor on Jatropha Curcas Plant Disease Identification," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, vol. 5, no. 1, 2020, doi: 10.22219/kinetik.v5i1.1012.

[6] L. Dymova, P. Sevastjanov, and K. Kaczmarek, "A Forex trading expert system based on a new approach to the rule-base evidential reasoning," *Expert Syst. Appl.*, vol. 51, pp. 1–13, 2016.

[7] T. Yüksel, "Intelligent visual servoing with extreme learning machine and fuzzy logic," *Expert Syst. Appl.*, vol. 72, pp. 344–356, 2017, doi: 10.1016/j.eswa.2016.10.048.

[8] M. I. Mazdadi, R. Ramadhani, T. H. Saragih, and M. Haekal, "Klasifikasi Tanaman Jarak Pagar Menggunakan Algoritme Deep Learning H2O," *J. Komputasi*, vol. 9, no. 1, 2021, doi: 10.23960/komputasi.v9i1.2774.

[9] M. Rostami, K. Berahmand, and S. Forouzandeh, "A novel community detection based genetic algorithm for feature selection," *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-020-00398-3.

[10] R. V. Schulte, E. C. Prinsen, H. J. Hermens, and J. H. Buurke, "Genetic Algorithm for Feature Selection in Lower Limb Pattern Recognition," *Front. Robot. AI*, vol. 8, no. October, pp. 1–12, 2021, doi: 10.3389/frobt.2021.710806.

[11] K. Anitha and K. S. Varaprasad, "Jatropha Pests and Diseases: An Overview," in *Jatropha, Challenges for a New Energy Crop*, New York, NY: Springer New York, 2012, pp. 175–218.

[12] D. Sarkar, R. Bali, and T. Sharma, *Practical Machine Learning with Python*. Berkeley, CA: Apress, 2018.

[13] S. L. Domingos, R. N. Carvalho, R. S. Carvalho, and G. N. Ramos, "Identifying it purchases anomalies in the Brazilian Government Procurement System using deep learning," *Proc. - 2016 15th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2016*, no. Cic, pp. 722–727, 2017, doi: 10.1109/ICMLA.2016.106.

[14] R. G. Lopes, R. N. Carvalho, M. Ladeira, and R. S. Carvalho, "Predicting Recovery of Credit Operations on a Brazilian Bank," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2016, pp. 780–784, doi: 10.1109/ICMLA.2016.0139.

[15] A. A. Candel, Arno, Viraj Parmar, Erin LeDell, "Deep learning with H2O," *H2O. ai Inc*, no. October, pp. 1–21, 2016.

[16] D. G. A. Adnyana and N. W. S. Suprapti, "PENGARUH KUALITAS PELAYANAN DAN PERSEPSI HARGA TERHADAP KEPUASAN DAN LOYALITAS PELANGGAN GOJEK DI KOTA DENPASAR," *E-Jurnal Manaj. Univ. Udayana*, vol. 7, no. 11, p. 6041, Aug. 2018, doi: 10.24843/EJMUNUD.2018.v07.i11.p09.

[17] [17] P. Guo, W. Cheng, and Y. Wang, "Hybrid evolutionary algorithm with extreme machine learning fitness function evaluation for two-stage capacitated facility location problems," *Expert Syst. Appl.*, vol. 71, pp. 57–68, 2017, doi: 10.1016/j.eswa.2016.11.025.

[18] T. H. Saragih, W. F. Mahmudy, and Y. P. Anggodo, "Genetic algorithm for optimizing FIS Tsukamoto for dental disease identification," *2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017*, vol. 2018-Janua, pp. 345–349, 2018, doi: 10.1109/ICACSIS.2017.8355057.

[19] T. H. Saragih, W. F. Mahmudy, and Y. P. Anggodo, "Optimization of Dempster-Shafer's Believe Value Using Genetic Algorithm for Identification of Plant Diseases Jatropha Curcas," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 1, 2018.

[20] Q. Kotimah, W. F. Mahmudy, and V. N. Wijayaningrum, "Optimization of Fuzzy Tsukamoto Membership Function using Genetic Algorithm to Determine the River Water," *Int. J. Electr. Comput. Eng.*, vol. 7, no. 5, pp. 2838–2846, 2017, doi: 10.11591/ijece.v7i5.pp2838-2846.

[21] W. F. Mahmudy, R. M. Marian, and L. H. S. Luong, "Real Coded Genetic Algorithms for Solving Flexible Job-Shop Scheduling Problem - Part II: Optimization," *Adv. Mater. Res.*, vol. 701, pp. 364–369, May 2013, doi: 10.4028/www.scientific.net/AMR.701.364.

[22] L. M. R. Rere, M. I. Fanany, and A. M. Arymurthy, "Simulated Annealing Algorithm for Deep Learning," *Procedia Comput. Sci.*, vol. 72, pp. 137–144, 2015, doi: 10.1016/j.procs.2015.12.114.

[23] T. H. Saragih, W. F. Mahmudy, A. L. Abadi, and Y. P. Anggodo, "Application of extreme learning machine and modified simulated annealing for jatropha curcas disease identification," *Int. J. Adv. Soft Comput. its Appl.*, vol. 10, no. 2, pp. 108–119, 2018.

[24] G. A. F. Alfarisy, A. N. Sihananto, T. N. Fatyanosa, M. S. Burhan, and W. F. Mahmudy, "Hybrid Genetic Algorithm and Simulated Annealing for Function Optimization," *J. Inf. Technol. Comput. Sci.*, vol. 1, no. 2, pp. 82–97, 2017.

[25] S. S. Sarkar, K. H. Sheikh, A. Mahanty, K. Mali, A. Ghosh, and R. Sarkar, "A Harmony Search-Based Wrapper-Filter Feature Selection Approach for Microstructural Image Classification," *Integr. Mater. Manuf. Innov.*, vol. 10, no. 1, pp. 1–19, Mar. 2021, doi: 10.1007/s40192-020-00197-x

[26] N. Yusup, A. M. Zain, and A. A. Latib, "A review of Harmony Search algorithm-based feature selection method for classification," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012038.