International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Building an Integrated Model Using Decision Trees to Improve the Quality of ECG Signals Recognition

Dinh Do Van

Submitted: 26/04/2023 **Revised**: 28/06/2023 **Accepted**: 05/07/2023

Abstract: Recognizing and improving the quality of recognition of electrocardiographic signals has many published scientific works, each with different methods. To improve the quality of ECG signal recognition, the article proposes a solution to improve the quality (accuracy) of ECG signal recognition (Electro Cardio Graphy), based on the use of binary decision trees to combine many single recognition models, which are classic neural networks MLP (Multi Layer Perceptron), neuro-fuzzy TSK network (Takaga-Sugeno-Kang), SVM (Support Vector Machines) and RF (Random Forest). The article uses Hermite basis functions (Hermite Basis Functions) to develop QRS complex and two time characteristics which are the distance between two consecutive peaks R (R-R), the average value of the last 10 R-R distances. The algorithms have been tested and tested on the classic data sets of the international classic database MIT-BIH (Massachusetts Institute of Technology, Boston's Beth Israel Hospital) and MGH database from the Web site http://physionet.org.

Keywords: *Neural network, MLP, TSK, SVM, Integrated System, Decision Tree, Hermite Basis Functions, Electrocardiogram (ECG) Signals.*

1. Introduction

The ECG signal has a very strong variation in both shape and amplitude in pathological cases. Signals are also susceptible to external interference, the patient's physical or psychological condition. Therefore, ECG identification is one of the difficult problems. In fact, there is a demand for smart electrocardiogram devices with automatic identification of pathological cases, requiring identification solutions that need high accuracy and distinguish many diseases to be applicable to many patients, etc.

Stemming from the above practical needs, the goal of this paper is to propose a solution to improve the quality of the ECG signal recognition (reduce the number of false positives).

Currently, there are many different solutions to improve the quality of ECG signal recognition, which have been researched and published by domestic and international authors, such as from collection, pre-processing, feature extraction or identification block (nonlinear). Most solutions are in the form of “single model”, a few solutions are in the form of “combined model”. As in the study [1], the author combined two single models, SVM and PSO (Particles Swarm Optimization), the test results on the MIT-BIH database had an accuracy increase of about 4% compared to when using a single SVM, or as in the work [4] the author combined Fuzzy K Nearest Neighbors (Fuzzy K Nearest Neighbors) and MLP networks, which improved from 97,3% to 98,0% accuracy.

Dinh Do Van, Sao Do University, Hai Duong-03500, Viet Nam
ORCID ID : 0000-0003-4425-2421
** Corresponding Author Email: dinh.dv@saodo.edu.vn*

The current growing trend is to use combined models for identification, especially for problems requiring high recognition accuracy such as ECG signal recognition, the model uses many single recognition models to make conclusions (possibly different), then one more processing step to synthesize the results from single recognition models to arrive at the final conclusion of the solution, some advantages of the "integrat model" solution:

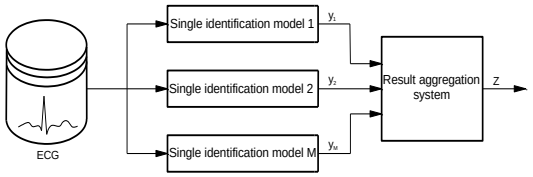

Fig. 1. General diagram of the coordination model

- Each “single model” is considered as an independent expert, the combination of many experts will give more reliable results;

- Using assessments from many angles, from many different methods, so the information can be richer leading to higher decision quality.

However, this model also has some disadvantages such as:

- The system will be cumbersome, more complex;

- It is necessary to develop a suitable synthesis method, if the coordination is not good, it will make the results worse.

The general block diagram of the combined solution is presented as shown in Figure 1, where the single recognition systems will process the same input signal from the object (but in different ways) and the output of the single recognition systems will form the aggregate block input, the

composite block result will also be the final recognition result.

From the purpose set out above, through studying and analyzing the advantages of solutions to improve the quality of ECG signal recognition, the article is selected in the direction of the second research, that is, using a parallel model of many single recognition models.

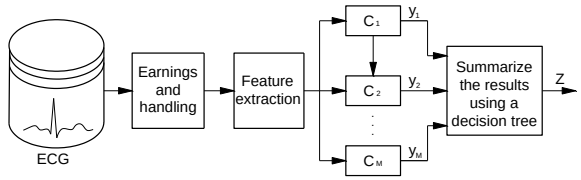

Fig. 2. Association model using decision tree

Some methods of synthesizing results have been applied by other authors such as [5, 6]:

- Voting by majority (Majority voting);

- Weighted voting;

- Synthesized according to Bayesian probability,...

This paper proposes to use Binary Decision Tree as the result aggregation block, and the single recognition models used are: classical neural network MLP, fuzzy logic neural network TSK, support vector machine SVM and RF random forest. These single models were chosen because these are results that have been published in international journals and conferences, so they ensure objectivity and accuracy [7], and at the same time, the results are made on the same input data set, so the comparison will be convenient and convincing, in which the RF random forest method is further elaborated in this paper.

As shown in Figure 2, it is assumed that each single identifier block C_i will produce the corresponding output y_i ($i = 1,2,\ldots,M$) which is a value containing the identifier. Then the input of the decision tree will be the composite vector $\mathbf{x} = [y_1, y_2, \ldots, y_M]$. The output of the Decision Tree z will be the final conclusion about the processing heartbeat.

2. Decision Tree

Decision Tree (DT) is a classical data classification model that has been widely used in many practical applications. Trees are usually depicted in the form of a hierarchical structure as shown in Figure 3. A tree consists of a set of nodes and branches with the following conditions:

- There exists a node called the root node;

- The node contains the branching condition.

At a leaf node (with no child branches), the identification result will be the overall result of the leaf.

The decision tree has simple conditional nodes, but because of the combination of many nodes, we get a division function with high nonlinearity suitable for complex classification problems, but the construction of node conditions is still relatively simple (the article uses the ID3 algorithm to build the tree).

In the paper, a binary tree (order 2) is used to simplify the description of algorithms, which does not reduce the generality of the tree because any tree of any degree can be converted to an equivalent binary tree as shown in Figure 4.

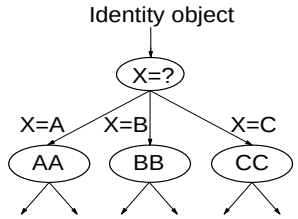

Fig. 3. Example of a decision tree structure

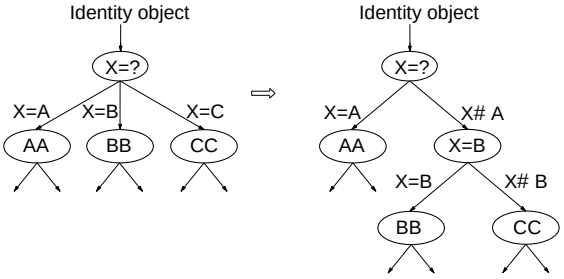

Fig. 4. Method to convert a higher order node (left figure) into a binary node (second order) (right figure)

3. Single Identification Models

3.1. Neural network MLP

MLP network is the most popular neural network, it is a feedforward network with basic elements called neurons. In the article, we use MLP network with a hidden layer with structure like Figure 5. The task is to define a fixed structure for the MLP network: The number of hidden layers, the transfer function of each layer, the number of neurons per layer, the weight of the coupling between neurons in the MLP model can be adjusted accordingly during the learning process to output the desired output signals. The learning algorithm used for the MLP model in this paper was proposed by Levenberg and Marquardt [9].

3.2. TSK fuzzy neural network

The second single identity model used in this paper is the Takagi–Sugeno–Kang (TSK) network. TSK network has been presented in quite detail in [9], so in this paper, it will not be presented again, the paper uses the structured TSK model as in the study [7, 8].

3.3. SVM classification model

The third single recognition model used in this paper is

SVM (Support Vector Machine), also known as support vector machine. Although the SVM model only divides the data into 2 classes, the classification for more classes is not complicated, it can be applied one-on-one, or one-on-all as in the study [5]. The more efficient method is the one-on-one method, where multiple SVM networks are built to classify in all combinations of two data classes. With N layers, we have to build a single SVM network.

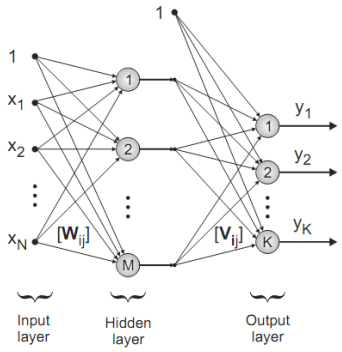

Fig. 5. MLP network with a hidden layer

3.4. Random Forest model

The final single recognition model used in this paper is the Random Forest model (RF) developed from L.Breiman (2001) [3], the basic structure of the RF model is a large set of N unpruned decision trees, the structure of each decision tree is randomly trained from a known sample data set. Steps to create the structure of RF:

- The input is the data set used for training;

- Each RF model is a set of N decision trees with N preselected.

- The structure of each decision tree is trained by a set of samples randomly taken in the common sample set.

- Single decision trees do not need to use pruning algorithms to reduce leaf nodes or to reduce the number of tiers of the tree.

- The stage of synthesizing the identification results from N popular decision trees using the majority voting method to give the final result to the RF.

Thus, for a new data sample to be tested, it is first passed through N decision trees for classification, each tree will have its own result (may be the same or different) and these N results will be put into the synthesis stage to process and give the final result.

4. Characteristics And Ekg Signal Data Extraction

4.1. Characterization of the electrocardiogram signal

The article uses Hermite basis functions (Hermite Basis Functions) to develop QRS complexes to take the expansion coefficients as feature vectors of the signal. In addition, we also use two more time characteristics: the distance between two consecutive R-vertices, the average value of the last 10 R-R distances.

The Hermite function has the following formula:

$$y_n(x) = \left(\sqrt{p} \times 2^n \times n!\right)^{-\frac{1}{2}} e^{-\frac{x^2}{2}} H_n(x) \quad (1)$$

where $H_n(x)$ is the Hermite polynomial defined recursively:

$$H_{n+1}(x) = 2x \times H_n(x) - 2n \times H_{n-1}(x) \quad (2)$$

Give $n \geq 1$, with $H_0(x) = 1; H_1(x) = 2x$.

Observed in Figure 6, we can see that the higher the order of the Hermite function, the larger the rate of change of the function, or in other words the function will contain more components of higher order. At the same time, the posture of the functions is also quite similar to the shape of the basic components in the ECG signal. This is the basis of using the Hermite function to analyze ECG signals.

To represent the ECG signal s(t) according to the first N Hermite functions as in formula (3), we use the analysis according to the singular values SVD (Singular Value Decomposition) to find the optimal solution of the system of first-order equations with more equations than the unknowns, details can be found in [7, 8].

$$s(t) \approx \sum_{i=0}^{N-1} c_i \times y_i(t) \quad (3)$$

From Figure 7 we can see that the ECG signal and especially the QRS complex segment has been approximated very well when using the first 16 Hermite basic functions, the error at the Q, R and S peaks is small, in Figure 8 we see that even with pathological cases, the signal varies strongly, the expansion to the first 16 basic Hermite functions is still quite good.

This set of 16 c_i values is used to form the characteristic vector of the ECG signal. In addition, two more time characteristics are used: the distance between two consecutive R-R peaks, the average value of the last 10 R-R distances. Thus, the feature vector has 18 values.

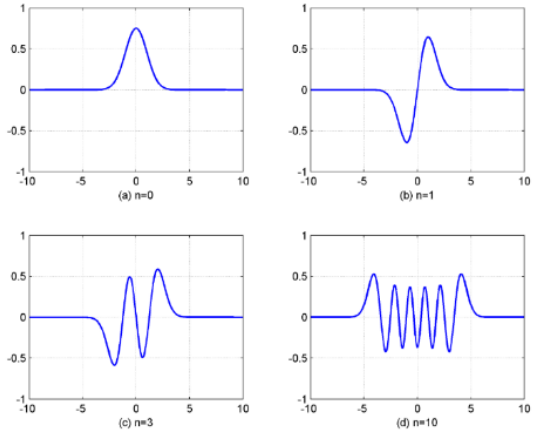

Fig. 6. Graph of the Hermite function of order n: (a) n=0; (b) n=1; (c) n=3; (d) n=10

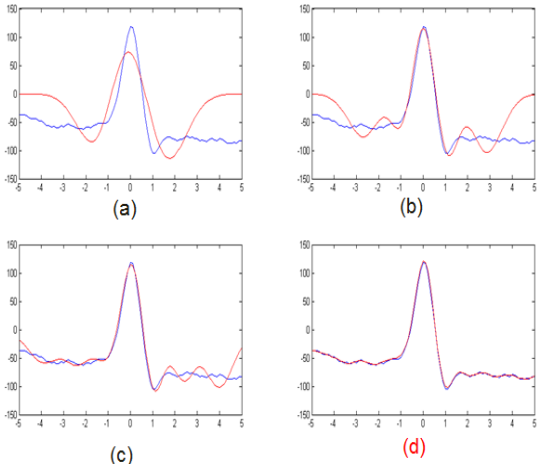

Fig. 7. Approximate ECG signal by first N Hermite functions: (a) N=5; (b) N=10; (c) N=12; (d) N=16.

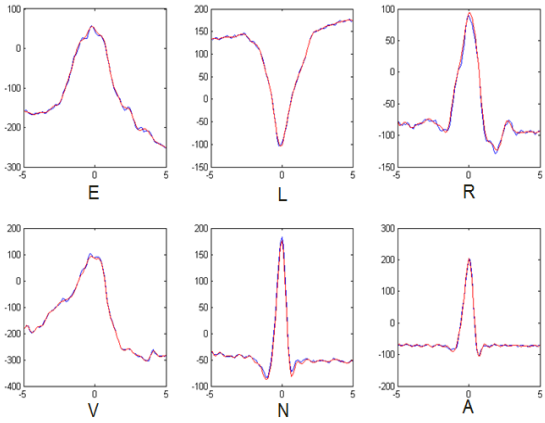

Fig. 8. An approximation of the ECG signal by the first 16 Hermite functions for several other heart

4.2. Database of ECG signals

4.2.1. MIT-BIH database

The first ECG database used in the article is the famous arrhythmia database MITBIH [2], which selected the records of 19 patients (the codes in the database are 100, 105, 106, 109, 111, 114, 116, 118, 119, 124, 200, 202, 207, 208, 209, 212, 214, 221 and 222), and the six types of arrhythmias considered are: Left bundle branch block beat (L), Right bundle branch block (R), Fremature Atrial contractions (A), Premature Ventricular contractinons (V), Ventricular flutter wave (I), and Ventricular Escape beat (E) and 1 Normal rhythm pattern (N). The number of sample details taken from the records of 19 patients is detailed in Tables 1 and 2 below:

Table 1. Table of distribution of study and test samples of 7 types of arrhythmias from the MIT-BIH database

Type of span	Number of samples	Learn samples	Test samples
N	2000	1065	935
L	1200	639	561
R	1000	515	485
A	902	504	398
V	964	549	451
I	472	271	201
E	105	68	37
Sum	*6643*	*3611*	*3068*

Table 2. Table of division of the number of learning and test samples of 2 types of span

Type of span	Number of sample	Learn sample	Test sample
Normal	2000	1065	935
Abnormal	4643	2546	2133
Sum	*6643*	*3611*	*3068*

4.2.2. MGH database

The second database is MGH [11], which includes 250 records of ECG signals, collected from 250 cardiovascular patients in the intensive care unit, operating room, cardiac catheterization laboratory,... at Massachusetts General Hospital. The article selects using ECG signal samples of 20 records with codes: 029, 030, 058, 105, 106, 107, 108, 110, 111, 114, 117, 119, 121, 123, 124, 125, 128, 131, 137, 142, taking out a total of 4500 samples with 3 types of rhythms: Normal sinus rhythm (N), Supraventricular premature bear (S), Ventricular premature contraction (V) detailed statistics in Table 3 and Table 4 below.

Table 3. Table of division of the number of learning and test samples of 3 types of span

Type of span	Number of sample	Learn sample	Test sample
N	3000	1997	1003
S	750	502	248
V	750	501	249
Sum	*4500*	*3000*	*1500*

Table 4. Table of division of the number of learning and test samples of 2 types of span

Type of span	Number of sample	Learn sample	Test sample
Normal	3000	1997	1003
Abnormal	1500	1003	497
Sum	*4500*	*3000*	*1500*

5. Calculation Results

5.1. Test results on the MIT-BIH database

5.1.1. Types of heart rate recognition

With 4 single recognition models MLP, SVM, TSK, RF in the paper, the parameters of these models are trained independently on the same learning dataset, with the following results:

- First, with the structure of the MLP model with 1 hidden layer, with 20 neurons, including 7 output neurons (corresponding to 7 types of arrhythmias),

- As for the parameters of the SVM model: with 7 classes and by the 1-on-1 method to find the winning class in the SVM model. Given the sample set has 7 classes, the authors have to build 21 single SVM networks for each pair of 2 signal types at the same time.

- The structure of the TSK network has 21 inference rules and 7 outputs.

- Finally, the RF model has 100 decision trees, each tree has a maximum of 9 floors, summarizing the results according to the method of crowd voting. All the outputs from the above single recognition models will be pushed into the input for the Decision Tree (DT), and there will be one more learning process to build the parameters for the DT, the end result of the ECG signal recognition is the output of the DT. For the above 4 single recognition models, we will test cases that combine results from 3 models (there are 4 possible combinations: MLP-TSK-SVM; MLP-TSK-RF; MLP-RF-SVM and RF-TSK-SVM) and there is 1 model that integrated all 4 single models MLP-SVM-TSK-RF. Use a common set of sample data to test the identity model. The results of this test will be used to compare with the results of previous studies. Table 5 and Figure 9 show the test error results of 4 basic identification models MLP, TSK, SVM, RF and 4 integrated models. All classification model networks will first be trained on the same learning dataset and then tested on another test dataset.

Table 5. Results of recognizing 7 types of spans (MIT-BIH database) by single models and integrated models

Classification system	Number of errors	Percent error (%)
MLP	110	3,59
TSK	100	3,26
SVM	60	1,96
RF	70	2,28
MLP-TSK-SVM	38	1,24
MLP-TSK-RF	43	1,40
MLP-RF-SVM	40	1,30
RF-TSK-SVM	39	1,27
MLP-TSK-SVM-RF	37	1,21

From Table 5, we can see that the results of the integrated models using the DT have been improved compared to the results of the single recognition models. The quality of the integrated model depends on the quality of each single recognition model and the number of single models, usually the larger the number of single recognition models, the more reliable the combined results.

5.1.2. Identify 2 types of normal and abnormal heart rhythm

Doing the same as for when the models recognize 7 types of heart rate in item (a), we build 4 single recognition models MLP, SVM, TSK, RF and have the following results:

- First, with the structure of the MLP model with 1 hidden layer, with 20 neurons, including 2 output neurons (corresponding to 2 types of normal and abnormal heartbeat);

- As for the parameters of the SVM model, just build with 1 layer and follow the 1-on-1 method to classify 2 types of normal and abnormal heart rhythms;

- The structure of the TSK network has 18 inference rules and 2 outputs;

- Finally, the RF model has 100 decision trees, each tree has 9 floors, summarizing the results according to the method of crowd voting.

From Table 6, we can see that for the case of classifying 2

types of normal and abnormal heart rhythms, the results of the integrated models using the decision tree are significantly higher than the results of the single recognition models.

Table 6. Results of identifying 2 types of spans (MIT-BIH database) by single models and integrated models

Classification system	Number of errors	Percent error (%)
MLP	39	1,27
TSK	41	1,34
SVM	26	0,85
RF	37	1,21
MLP-TSK-SVM	21	0,68
MLP-TSK-RF	22	0,72
MLP-RF-SVM	23	0,75
RF-TSK-SVM	16	0,52
MLP-TSK-SVM-RF	15	0,49

5.2. Test results on the MGH database

In order to further evaluate the accuracy and reliability of the model of receiving coordination by Decision Tree (DT), the article further tested with the MGH database, and obtained the results as shown in Tables 7 and 8.

Table 7. Results of identifying 3 types of rhythm (MGH/MF database) by single models and integrated models

Classification system	Number of errors	Percent error (%)
MLP	66	4,40
TSK	73	4,87
SVM	32	2,13
RF	96	6,40
MLP-TSK-SVM	25	1,67
MLP-TSK-RF	30	2,00
MLP-RF-SVM	25	1,67
RF-TSK-SVM	25	1,67
MLP-TSK-SVM-RF	21	1,40

Table 8. Results of identifying 2 types of rhythm (MGH/MF database)

Classification system	Number of errors	Percent error (%)
MLP	37	2,47
TSK	62	4,13
SVM	20	1,33
RF	78	5,20
MLP-TSK-SVM	17	1,13
MLP-TSK-RF	20	1,33
MLP-RF-SVM	19	1,27
RF-TSK-SVM	18	1,20
MLP-TSK-SVM-RF	15	1,00

5.3. Result evaluation

From the above test results, we have some conclusions as follows:

- Through the test results on the MIT-BIH and MGH/MF databases (which are often used by international research groups for reference), the article has demonstrated that the solution of parallel combination of many basic recognition models by DT has continued to improve the quality of ECG signal recognition results. The test error (number of falsely identified samples) of the integrated models is lower than that of the basic recognition models;

- Only one case is equal – in Table 8, the error of the SVM

model and the MLP-TSK-RF integrated model has 20 false positives.

6. Conclude

The article has proposed a solution to improve the quality of ECG signal recognition based on the use of Decision Trees (DT) to combine multiple single recognition models. Algorithms have been partly tested on the classic international data sets MIT-BIH and MGH/MF used for reference by international research groups.

References

[1] Bazi F. and Melgani Y., "Classification of electrocardiogram signals with support vector machines and particle swarm optimization", *IEEE Transactions on Information Technology in Biomedicine*, vol. 12(5), 2008, pp. 667–677.

[2] G. và R. Mark Moody, "The impact of the MIT-BIH Arrhythmia Database", *IEEE Eng. in Medicine and Biology* 20(3)2001, pp. 45–50.

[3] L. Breiman, “Random forests”, *Machine Learning*, Vol. 45,2001, pp. 5–32.

[4] O. Castillo, E. Ramírez, J. Soria, "Hybrid System for Cardiac Arrhythmia Classification with Fuzzy K-Nearest Neighbors and Multi-Layer Perceptrons combined by a Fuzzy Inference System", 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1-6.

[5] S.Osowski, L.Tran Hoai, T.Markiewicz, "Ensemble of neural networks for improved recognition and classification of arrhythmia", *Metrology for a Sustainable Development September*, Rio de Janeiro, Brazil, 2006, pp. 17 – 22.

[6] S.Osowski, T. Markiewicz, L. Tran Hoai, "Recognition and classification system of arrhythmia using ensemble of neural networks", *Article in Measurement*, Vol. 41, 2008, pp. 610–617.

[7] Tran Hoai Linh, Pham Van Nam, Vuong Hoang Nam, “Multiple neural network integration using a binary decision tree to improve the ECG signal recognition accuracy”, *International Journal of Applied Mathematics and Computer Science*. Volume 24, Issue 3, 2014, pp. 647–655.

[8] Tran Hoai Linh, Pham Van Nam, Nguyen Duc Thao, "A hardware implementation of intelligent ECG classifier", *COMPEL: The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, vol. 34, Iss: 3, 2015, pp. 905 – 919.

[9] Tran Hoai Linh, "Neural networks and their applications in signal processing", *Hanoi Polytechnic Publishing House*, 2014.

[10] S.Osowski, T. Markiewicz, L. Tran Hoai, "Recognition and classification system of arrhythmia using ensemble of neural networks", *Article in Measurement*, Vol. 41, 2008, pp. 610–617.

[11] http://www.physionet.org, Accessed June 01, 2023

[12] Steven Martin, Thomas Wood, María Fernández, Maria Hernandez, .María García. Machine Learning for Educational Robotics and Programming. Kuwait Journal of Machine Learning, 2(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/179

[13] Sharma, R., & Dhabliya, D. (2019). A review of automatic irrigation system through IoT. International Journal of Control and Automation, 12(6 Special Issue), 24-29. Retrieved from www.scopus.com

[14] Anupong, W., Yi-Chia, L., Jagdish, M., Kumar, R., Selvam, P. D., Saravanakumar, R., & Dhabliya, D. (2022). Hybrid distributed energy sources providing climate security to the agriculture environment and enhancing the yield. Sustainable Energy Technologies and Assessments, 52 doi:10.1016/j.seta.2022.102142