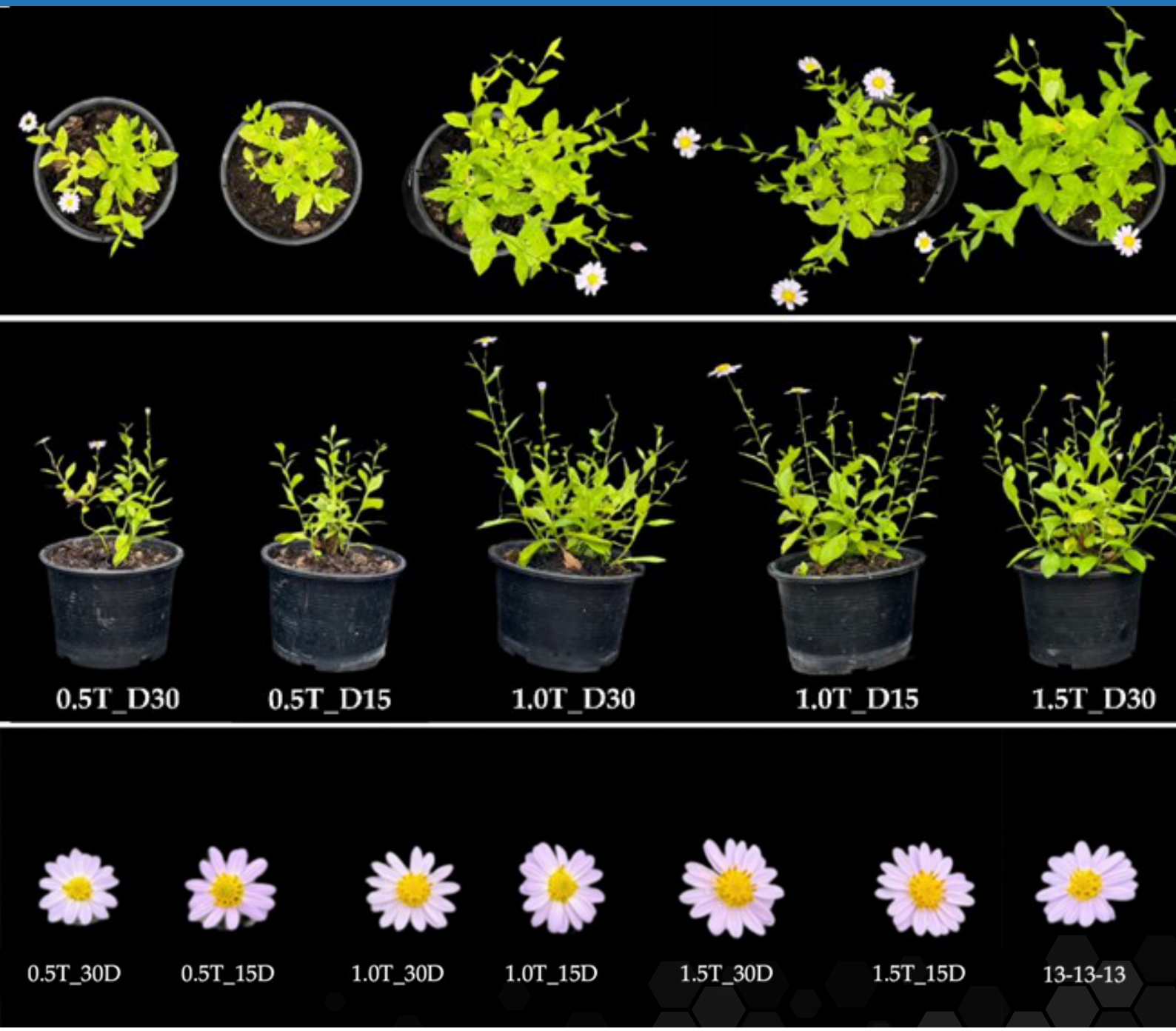
ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Vol. 27 No. 5, September - October 2024



ISSN 2773-8752 (online)
https://ph02.tci-thaijo.org/index.php/tsujournal/issue/view/17182

ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

ASEAN Journal of Scientific and Technological Reports (AJSTR)

Contact Us Institute of Research and Innovation, Thaksin University
222 M. 2 Ban-Prao sub-district, Pa-Pra-Yom district, Phatthalung province, Thailand
Tel. 0 7460 9600 # 7242 , E-mail: aseanjstr@tsu.ac.th

List of Contents

ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

Isolation of Fiber-Cellulose and Characterization from Oil Palm Frond for 3D and 4D Printing Materials Application

Warunee Manosong [1], Methawee Nuanla-ong[2], and Anurak Udomvech[1, 3*]

[1] Department of Physical Science, Faculty of Science and Digital Innovation, Thaksin University, Songkhla Campus, 90000, Thailand
[2] Farmer, Thung Rang Subdistrict, Kanchanadit District, Surat Thani Province, 84290, Thailand
[3] Research Laboratory for Intelligent Materials Design, Discovery, and Developments (RLIMD[3]), Faculty of Science and Digital Innovation, Thaksin University, Songkhla Campus, 90000, Thailand
* Correspondence: anurak@tsu.ac.th

Citation:
Manosong, W.; Nuanla-ong, M.; Udomvech, A. Isolation of fiber-cellulose and characterization from oil palm frond for 3D and 4D printing materials application. *ASEAN J. Sci. Tech. Report.* **2024**, *27*(5), e253216. https://doi.org/10.55164/ajstr.v27i5.253216.

Article history:
Received: March 15, 2024
Revised: July 15, 2024
Accepted: August 8, 2024
Available online: August 27, 2024

Abstract: The oil palm frond (OPF), a primary biomass derived from agricultural waste in oil palm production, holds significant potential for evolving into cellulose fiber. Leveraging natural materials for cellulose extraction aligns with environmental sustainability. In this study, we focus on separating cellulose fibers from OPF biomass. The OPF is divided into its leaflet and frond-axial parts, undergoing a sequence of steps: dewaxing, alkaline treatment, delignification, hydrogen peroxide treatment, and final filtration. These steps assess both efficacy and cellulose fiber outputs. Notably, the axial and leaflet portions yield cellulose fibers, effective as an initial extraction step before progressing to nanoscale slicing. Quantitatively, extraction yields indicate that 14.13% and 19.52% of the 100 g leaflet and frond-axial portions can be harvested. SEM images confirm well-dispersed individual fibers. Additionally, GCMS results reveal slightly higher carbon dioxide gas in the palm leaflet sample than in the palm frond-axial sample after five days of water soaking. This procedure efficiently initializes OPF for cellulose extraction, positioning it as an initial ingredient for subsequent 3D/4D printing material production.

Keywords: Oil Palm Frond (OPF); Cellulose Fiber; Morphology

1. Introduction

Over 100 million palm trees thrive globally, with oil palm fronds (OPF) constituting a significant biomass—accounting for 47% of total oil palm waste [1,2]. Despite being considered waste, OPF is a valuable raw material due to its affordability, accessibility, and abundance [3,4]. Notably, in the realm of additive manufacturing (AM) and 3D/4D printing technologies [5-9], a transformative landscape has emerged for rapid production and plastic recycling in personalized consumer goods, automotive components, medical devices, aerospace applications, and energy-related products [10,11]. Leveraging its mechanical strength akin to steel, OPF cellulose stands out as an excellent reinforcement option for 3D/4D printing materials. Moreover, its biodegradability and renewable nature position natural fiber as a sustainable and ecologically acceptable bio-composite material for biomedical and industrial purposes [12].

Oil Palm Fronds (OPF) constitute agricultural waste generated during the regular pruning of palm trees, which occurs approximately every 20 days when fresh fruit bunches (FFB) are harvested. Only a small portion of OPF is repurposed for composting, while the majority is incinerated on-site [13].

Unfortunately, this practice contributes to air pollution and poses risks to human safety. Efficient waste management strategies are essential to address these challenges and mitigate environmental concerns. Oil Palm Fronds (OPF) possess a high fiber content due to their inherent spongy and fibrous nature. Consequently, they are well-suited for cellulose fiber synthesis. Within OPF, vascular bundles and parenchyma contain a mixture of extractives, cellulose, hemicellulose, and lignin. These other components must be selectively removed through enzymatic, chemical, mechanical, or chemo-mechanical treatments to obtain cellulose fibers. In essence, cellulose comprises a homopolymer of β-D-glucopyranosyl repeating units connected by 1,4-glycosidic bonds [14].

This work aims to extract cellulose fibers from oil palm frond (OPF) biomass. Using cost-effective and straightforward chemical extraction methods, we have divided the OPF into leaflet and frond-axial parts. Our investigation assesses the efficiency and cellulose fiber yields for leaflet and frond-axial portions.

2. Materials and Methodologies

2.1 Raw Materials and Sample Preparation

The oil palm frond (OPF) waste biomass has been obtained from a palm plantation in Suratthani Province, Thailand. Initially, we sorted out the leaflets from the frond. Subsequently, both components were shredded and dried under ambient conditions for 72 hours. Afterward, the raw materials were dried at room temperature for 24 hours and then cut into smaller pieces with an average length of 0.2 mm. Finally, the samples were oven-dried at 100±2°C.

2.2 Extraction of Cellulose Fiber

Raw cellulose fibers underwent dewaxed in Soxhlet equipment for 6 hours with 200 mL of 70% (v/v) ethyl alcohol, with a fiber-to-solvent ratio of 1:10 (g/L). After proper rinsing to eliminate alcohol residues, the dewaxed fibers are suspended in a 10% sodium hydroxide solution. The beaker, wrapped in aluminum foil, is autoclaved at 121°C and 1.5 bar pressure for an hour. These steps prepare the cellulose fibers for further use. Finally, the fibers retrieved from the supernatant were thoroughly rinsed with double-distilled water.

After autoclaving, the fibers underwent delignification by immersing them in a 1:1 (v/v) combination of 20% formic acid and 10% hydrogen peroxide. This mixture was preserved in a water bath at 85°C for 2 hours before being filtered to collect purified fibers. Initially, the fibers were cleaned with 10% formic acid and successive washings using double-distilled water. The resulting extracted cellulose fibers exhibited a pale yellow hue. Subsequently, a 10% hydrogen peroxide solution was applied to the cellulose fibers for 90 minutes at 60°C. The pH of the cellulose fiber suspension was adjusted using a 10% solution of sodium hydroxide. After filtration and repeated cleaning, the insoluble portion of cellulose was gathered, and the yield (w/w based on dry weight) was computed.

2.3 Characterizations

We examined the morphology of the extracted fibers and assessed the impact of sample treatments using scanning electron microscopy (SEM). Specifically, we employed the FEI Quanta 450 FEG SEM, operating at 5.0 kV high voltage under coverage pressures ranging from 9.4×10^{-4} to 1.1×10^{-3} Pa. Additionally, we verified carbon levels released during our procedures and from the samples using gas chromatography (GC) and mass spectrometry (MS). For this purpose, we utilized the GCMS-HSS instrument in our study.

3. Results and Discussion

3.1 Cellulose Fiber from Extraction Processes

To separate and remove dissolved polysaccharides and other constituents from water, we finely chopped and dried oil palm leaflets and frond samples in 2-L bottle containers. Each bottle was filled with water and allowed to soak for a week, softening the samples and facilitating the separation of polysaccharides and soluble substances. Subsequently, we filtered the samples using a sieve tool and rinsing with clean water and additional filtration 2–3 times. Figure 1 illustrates the two types of fibers obtained: those from the oil palm

frond sample exhibit a brunette color (Figure 1(a)), while the oil palm leaflet sample fibers appear green (Figure 1(b)). Finally, the dewaxing procedure yields the fibers shown in Figure 1(c).

Next, we examine the delignification results using two fiber samples from the previous procedure. These samples are placed in two-liter round-bottom bottles, each containing a 10% sodium hydroxide solution of 1,000 ml and a 30% hydrogen peroxide solution of 250 ml. We assemble the reflux apparatus kit and allow the solutions to boil for 3 hours, during which a significant amount of air bubbles may form, and vigorous boiling can occur. After boiling, we let the mixture cool down, then strain and rinse it with clean water multiple times. Finally, we squeeze the fibers until they are completely dry. The cellulose fibers produced through this delignification process exhibit a white color for the leaflet's fiber (see Figure 1(d)). However, the OPF fiber products from the frond part (Figure 1(c)) appear darker. This difference is attributed to the higher lignin content in the OPF frond, making alkali treatment more challenging than the leaflet portion. Consequently, the lignin in OPF frond fibers is more densely packed than in the leaflet fibers.

Figure 1. Samples of two types of fibers: (a) from palm fronds and (b) from palm leaflets after soaking for one week. The fibers resulting from the dewaxing procedure are shown in (c) and sourced from (a). Additionally, (d) illustrates the characteristics of palm fibers after bleaching and lignin removal from the source (b).

Figure 2. Characteristics of OPF fibers: (a) after undergoing the digestion process, and (b) and (c) after being squeezed until the leaflet and frond fibers are completely dry, respectively.

3.2 The Process of Digestion to be Small-Scaled Cellulose

Each sample type from the previous step was combined with 500 mL of a 20% acetic acid solution and 150 mL of a 30% hydrogen peroxide solution in a 2-L Erlenmeyer flask. The flask's mouth was sealed with a glass funnel. Both samples were immersed in a hot water bath at 85°C for 90 minutes, followed by cooling, filtration, and multiple rinses with clean water. Figure 2(a) illustrates the characteristics of OPF fibers after passing through this digestion process. Finally, the material was squeezed until the fibers were completely dry, as depicted in Figures 2(b) and 2(c) for the leaflet and frond. Subsequently, the cellulose fiber products were oven-dried at 110°C for 2–3 hours. The extraction results indicate that the yield percentage of cellulose from leaflet and frond oil palm biomasses (100 g each) is 14.13% (14.13 g) and 19.52% (19.52 g), respectively.

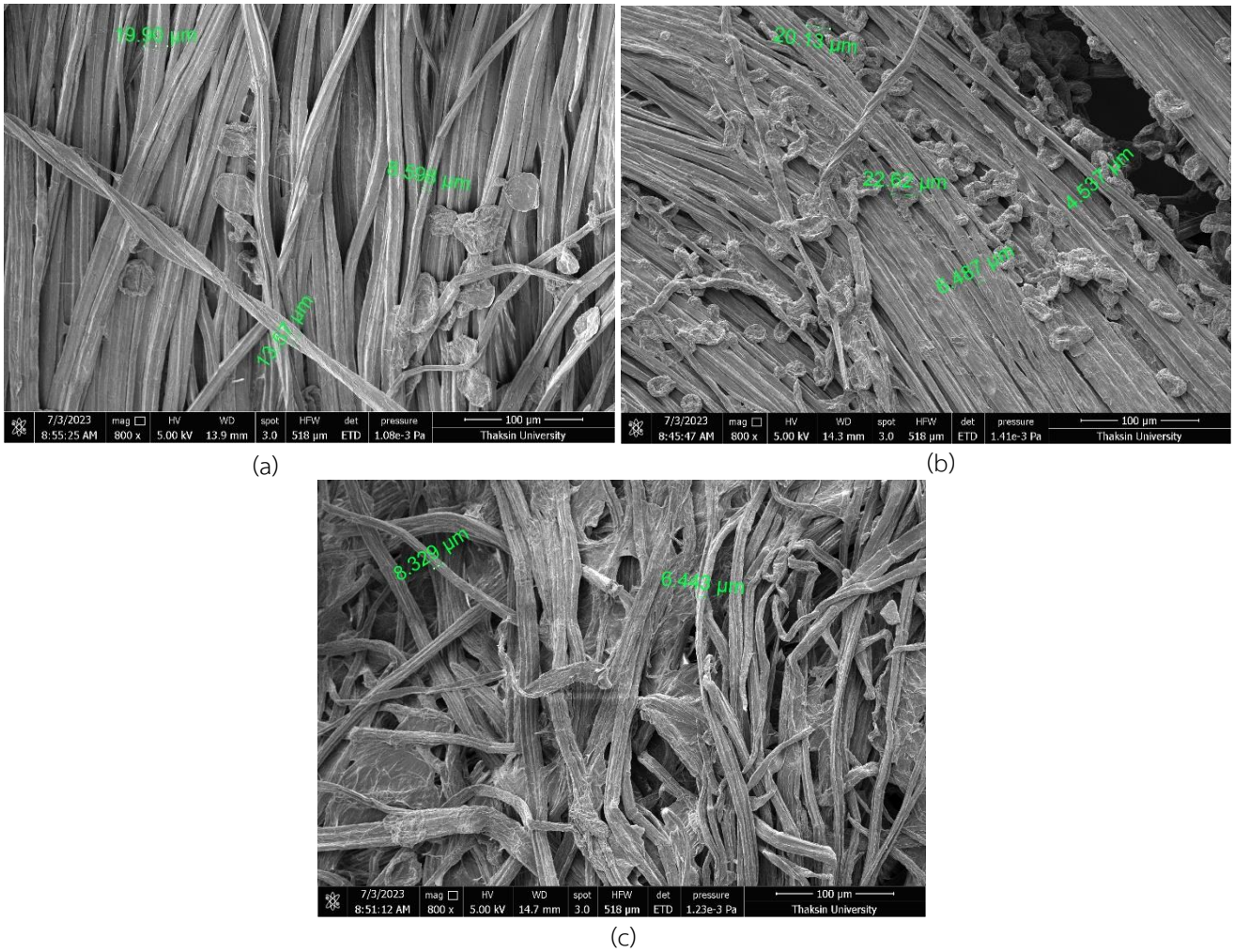
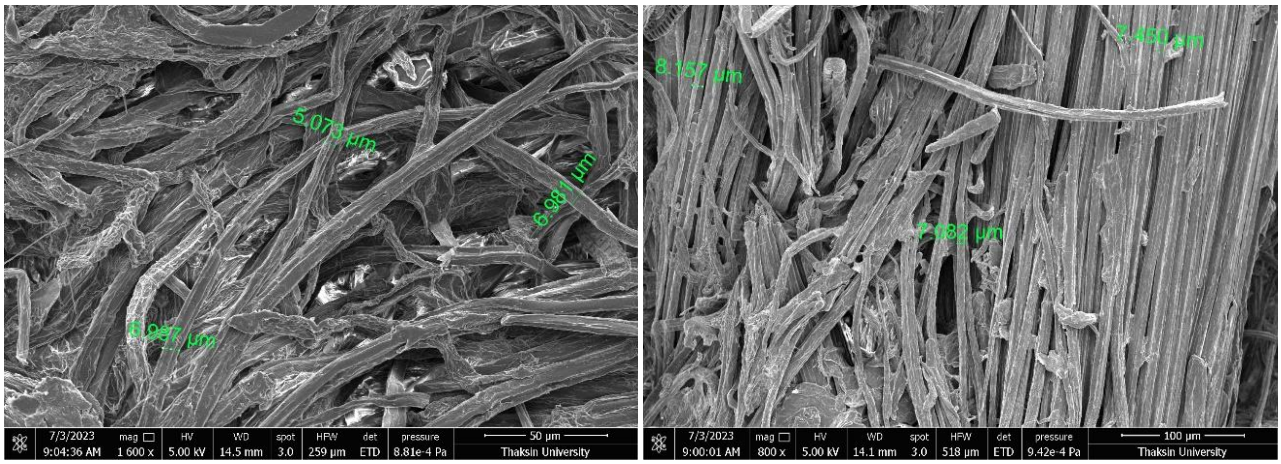


Figure 3. SEM micrographs of the raw frond-axial part of OPF reveal the following fiber thicknesses: (a) approximately ~8.6 – 19.9 µm, (b) approximately ~4.5 – 6.5 µm, and (c) approximately ~6.4 – 8.3 µm.

This result can be compared with the fronds and leaves, which yielded 2.90 g and 3.01 g, respectively, using the familiar extraction technique described by Mehanny et al. [2]. In contrast, the extraction yields from different methods applied to rice straws, corncobs, pineapple leaves, and pineapple peel biomass materials were 32.26%, 38.18%, 16.60%, and 9.05% (w/w), respectively [15]. Notably, our method demonstrates the effectiveness in extracting sufficient yields compared to other studies involving similar biomasses for cellulose production. The chosen procedures stand out due to their simplicity, reproducibility, and low operational costs. While it may perform well in cellulose yield, it exhibits attractive characteristics. Sufficient amounts of these cellulose products could potentially support material experiments for 3D/4D printing in subsequent stages.

3.3 Scanning Electron Microscopy (SEM)

The morphology of OPF fibers in leaflet and frond-axial samples has been examined using SEM. The treatment applied to the fibers significantly impacts their morphology. When extracting frond-axial fibers, various non-cellulosic and macromolecular substances—such as hemicelluloses, lignin, pectin, and wax—can be removed by exposing the fiber surface. SEM micrographs (see Figure 3) depict the extracted fibers from the frond-axial portion, which still exhibit an average size in the large micro-scale range (4.5–20 µm). In contrast, SEM images of the extracted cellulose fibers from the leaflet portion (see Figure 4) show an average size in the range of 5.1–8.2 µm.



(a) (b)

Figure 4. SEM micrographs of the raw leaflet part of OPF reveal the following fiber thicknesses: (a) approximately ~5.1 – 7.0 µm and (b) approximately ~7.1 – 8.2 µm.

Our extraction method demonstrates sufficient efficiency to serve as an initial step before further processing into nanoscale particles. These nanoscale particles can enhance reinforcement in mechanical strength, resulting in lighter, biodegradable, and more resilient materials suitable for 3D/4D printing. Such materials find applications in novel products and surface-modified assembly, enabling rapid tooling and scalable production. However, it's important to note that OPF fiber surfaces are coated with fatty materials, wax, impurities, and globular projections called tyloses. Despite this, the alkaline treatment effectively reduces roughness by eliminating contaminants from the fiber surface. Specifically, the interaction with sodium during treatment removes contaminants like wax and cuticles [16].

3.4 Gas Chromatography—Mass Spectrograph (GC-MS)

We employed this technique to verify whether our extraction process releases low levels of carbon. Since the waste generated during our procedure consists of organic compounds acting as non-toxic solvents, it is environmentally friendly. Simultaneously, this approach minimizes water contamination and allows for

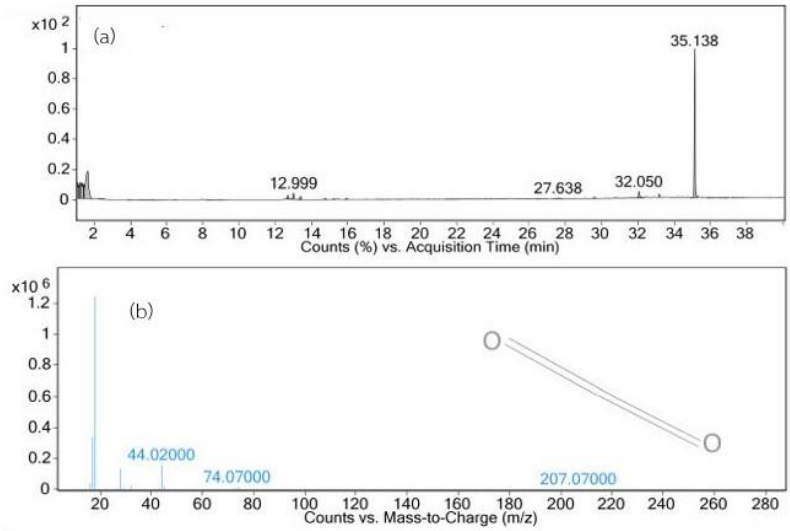


Figure 5. The chromatogram (a) and the spectrum (b) display the results of analyzing carbon dioxide (CO_2) released from frond-axial palm samples soaked in water for five days. Gas chromatography and mass spectrometry techniques were employed for this analysis.

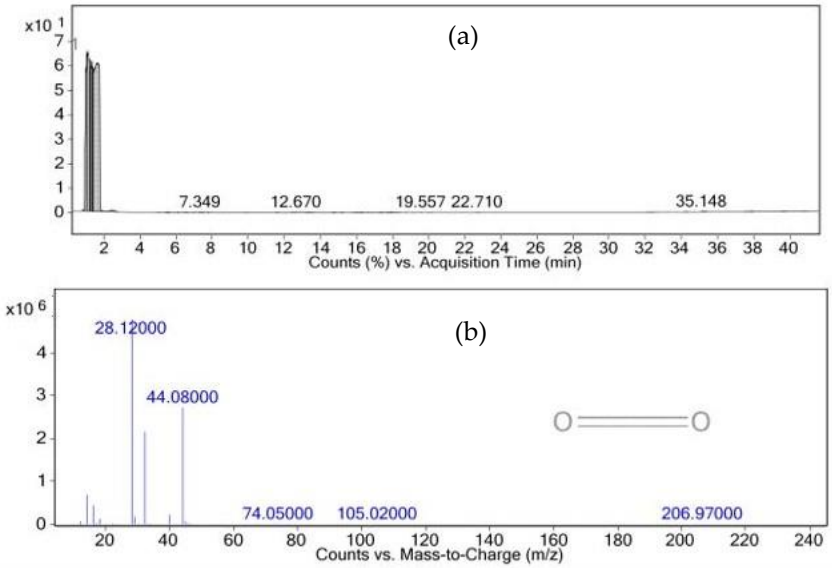


Figure 6. The chromatogram (a) and the spectrum (b) display the results of analyzing carbon dioxide (CO_2) released from leaflet palm samples soaked in water for five days. Gas chromatography and mass spectrometry techniques were employed for this analysis.

recycling. The analysis revealed a carbon dioxide peak at retention time (RT) equal to 35.138 minutes for the palm frond-axial crushed sample (see Figure 5(a)) and 35.148 minutes for the palm leaflet crushed sample (see Figure 6(a)). Additionally, we observed an ionized molecular species with a mass of 44.02000 m/z for the palm frond-axial crushed sample (see Figure 5(b)) and 44.08000 m/z for the palm leaflet crushed sample (see Figure 6(b)). Notably, this mass corresponds to the molecular weight of carbon dioxide gas (44.0100 m/z).

These findings indicate that soaking palm samples in water for five days produces trace carbon dioxide levels. Interestingly, the palm leaflet crushed sample exhibited slightly higher carbon dioxide gas production than the palm frond-axial crushed sample. Overall, both samples exhibit minimal carbon dioxide release.

4. Conclusions

The study revealed that using our procedure, cellulose fibers extracted from the frond-axial and leaflet parts of oil palm fronds (OPF) are efficient as an initial extraction step before further processing into nanoscale particles. Our methodology significantly improves the separated fibers compared to their original raw material. Specifically, the extraction yields for leaflet and frond-axial parts per 100 g of raw material are 14.13% and 19.52%, respectively. SEM images confirm that individual fibers are well-dispersed. Gas chromatography-mass spectrometry (GCMS) results indicate that the palm leaflet sample produces slightly more carbon dioxide than the palm frond-axial sample after soaking in water for five days. However, both samples exhibit minimal carbon dioxide release. Consequently, our method effectively enables OPF extraction for subsequent cellulose material synthesis.

5. Acknowledgments

W.M. and A.U. express gratitude to Dr. Nilubon Nuanchankhong for all conveniently support of chemical compounds and facilities.

Author Contributions: Conceptualization, A.U.; methodology, W.M.; formal analysis, W.M. and A.U.; investigation, A.U.; resources and raw materials, W.M. and M.N. ; data curation, A.U.; writing—original draft preparation, A.U.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest

References

[1] Roslan, A. M.; Zahari, M. A. K.; Hassan, M. A.; Shirai, Y. Investigation of Oil Palm Frond Properties for Use as Biomaterials and Biofuels. *Trop. Agr. Develop.* 2014, *58(1)*, 26–29. https://doi.org/10.11248/jsta.58.26

[2] Wang, Q.; Sun, J.; Yao, Q.; Ji, C.; Liu, J.; Zhu, Q. 3D printing with cellulose materials. *Cellulose* **2018**, 25, 4275–4301. https://doi.org/10.1007/s10570-018-1888-y

[3] Mehanny, S.; Magd, E. E. A-E.; Ibrahim, M.; Farag, M.; Gil-San-Millan, R.; Navarro, J.; Habbak, A. E. H. E.; El Kashif, E. Extraction and Characterization of Nanocellulose from three types of Palm Residues, *JMR&T* **2021**, *10*, 526-537. https://doi.org/10.1016/j.jmrt.2020.12.027

[4] Finny, A.S.; Popoola, O.; Andreescu, S. 3D-Printable Nanocellulose-Based Functional Materials: Fundamentals and Applications. *Nanomaterials* **2021**, *11*, 2358. https://doi.org/10.3390/nano11092358

[5] Pal, A. K.; Mohanty, A. K.; Misra, M. Additive Manufacturing Technology of Polymeric Materials for Customized Products: Recent Developments and Future Prospective. *RSC Advances* **2021**, *11*, 36398 – 36438. https://doi.org/10.1039/D1RA04060J

[6] Guvendiren, M.; Molde, J.; Soares, R. M.D.; Kohn, J. Designing Biomaterials for 3D Printing. *ACS Biomater. Sci. Eng.* **2016**, *2(10)*, 1679-1693. https://doi.org/10.1021/acsbiomaterials.6b00121

[7] Ahmed, A.; Arya, S.; Gupta, V.; Furukawa, H.; Khosla, A. 4D Printing: Fundamentals, Materials, Applications and Challenges. *Polymer* **2021**, *228*, 123926. https://doi.org/10.1016/j.polymer.2021.123926

[8] Kumar, S. B.; Jeevamalar, J.; Ramu, P.; Suresh, G.; Senthilnathan, K. Evaluation in 4D printing – A Review. *Materials Today: Proceedings* **2021**, *45*, Part 2. 1433-1437. https://doi.org/10.1016/j.matpr.2020.07.335

[9] Alsaadi, M.; Hinchy, E.P.; McCarthy, C.T.; Moritz, V.F.; Zhuo, S.; Fuenmayor, E.; Devine, D.M. Liquid-Based 4D Printing of Shape Memory Nanocomposites: A Review. *J. Manuf. Mater. Process.* **2023**, *7*, 35. https://doi.org/10.3390/jmmp7010035

[10] Dizon, J.R.C.; Gache, C.C.L.; Cascolan, H.M.S.; Cancino, L.T.; Advincula, R.C. Post-Processing of 3D-Printed Polymers. *Technologies* **2021**, *9*, 61. https://doi.org/10.3390/technologies9030061

[11] Petousis, M.; Vidakis, N.; Mountakis, N.; Papadakis, V.; Kanellopoulou, S.; Gaganatsiou, A.; Stefanoudakis, N.; Kechagias, J. Multifunctional Material Extrusion 3D-Printed Antibacterial Polylactic Acid (PLA) with Binary Inclusions: The Effect of Cuprous Oxide and Cellulose Nanofibers. *Fibers* **2022**, *10*, 52. https://doi.org/10.3390/fib10060052

[12] Khalil, H.P.S.A.; Mohamed, S.A.; Ridzuan, R.; Kamarudin, H.; Khairul, A. Chemical Composition, Morphological Characteristics, and Cell Wall Structure of Malaysian Oil Palm Fibers. *Polym. Plast. Technol. Eng* 2**008**, *47*, 273–280. https://doi.org/10.1080/03602550701866840

[13] Rosli, W.D.W.; Zainuddin, Z.; Law, K.N.; Asro, R. Pulp from Oil Palm Fronds by Chemical Processes. *Ind. Crops. Prod.* **2007**, *25*, 89–94. https://doi.org/10.1016/j.indcrop.2006.07.005

[14] Chen, H.; 4 - Lignocellulose biorefinery conversion engineering. In *Lignocellulose Biorefinery Engineering*, Chen, H., Eds.; Woodhead Publishing: Amsterdam. **2015**, pp. 87-124.

[15] Romruen O.; Karbowiak T.; Tongdeesoontorn W.; Shiekh K.A., Rawdkuen S. Extraction and Characterization of Cellulose from Agricultural By-Products of Chiang Rai Province, Thailand. *Polymers* **2022**, *14*(9), 1830. https://doi.org/10.3390/polym14091830

[16] Izani, M.A.N.; Paridah, M.T.; Anwar, U.M.K.; Mohd Nor, M.Y.; H'ng, P.S. Effects of fiber treatment on morphology, tensile and thermogravimetric analysis of oil palm empty fruit bunches fibers. *Compos. Part B* **2013**, *45*, 1251–1257. https://doi.org/10.1016/j.compositesb.2012.07.027

ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

Risk Assessment of Rubber Tapping A Case Study: Pa Phayom District, Phatthalung Province, Thailand

Sutee Inraksa[1]*, Angoon Sungkhapong[2] and Klangduan Pochana[3]

[1] Department of Industrial and Manufacturing Engineering, Faculty of Engineering, Prince of Songkla University, Songkhla, 90110, Thailand
[2] Department of Industrial and Manufacturing Engineering, Faculty of Engineering, Prince of Songkla University, Songkhla, 90110, Thailand
[3] Smart Industry Research Center, Department of Industrial and Manufacturing Engineering, Faculty of Engineering, Prince of Songkla University, Songkhla, 90110, Thailand
* Correspondence: juk007@hotmail.com

Citation:
Inraksa, S.; Sungkhapong, A.; Pochana, K. Risk assessment of rubber tapping a case study: Pa Phayom district, Phatthalung province, Thailand. *ASEAN J. Sci. Tech. Report.* **2024**, *27*(5), e254127. https://doi.org/10.55164/ajstr.v27i5.254127

Article history:
Received: May 16, 2024
Revised: July 16, 2024
Accepted: July 23, 2024
Available online: August 27, 2024

Abstract: This cross-sectional study investigated the risks of rubber tapping among rubber farmers in Pa Phayom District, Phatthalung Province, Thailand, utilizing the EART (Ergonomic Risk Assessment Tool for Rubber Tappers) and RULA (Rapid Upper Limb Assessment) approaches. Purposive sampling was utilized to randomly choose 154 rubber farmers, aged 20 to 60 to serve as the sample group. They met the inclusion requirements of being in excellent health, being able to read and write Thai fluently, not having a history of back discomfort or injuries, and having at least a year of experience tapping rubber. The results showed that the mean RTL value was 415, SD = 182, while the mean RTI value was 1.77, SD = 1.35, with Min = 0.31 and Max = 8.64. The RULA analysis gave scores ranging from 2 to 7 as acceptable risk to high risk, suggesting that rubber farmers should improve their working posture or reduce factors that affect risk to ensure better health outcomes. It was discovered that the EART and RULA can be used as an assessment method to identify ergonomic risk problems in rubber tapping operation. It was suggested that the rubber tapping operation needs to be improved, re-evaluated, and implemented immediately.

Keywords: Risk assessment; EART; Rubber tappers; Ergonomics

1. Introduction

Rubber is an economic crop grown throughout Thailand, especially in the southern region, where arable land comprises 60% [1]. Rubber tapping in Thailand is performed manually, and rubber farmers suffer from aches, pains, and musculoskeletal injuries [2] because the work involves repetitive body movements and ergonomics. A survey of rubber farmers recorded pain in the lower back (77.6%), hands and wrists (37.2%), upper back pain (34.4%), knee pain (31.6%), and also pain in other parts of the body [3]. A model was developed for risk assessment to prevent and reduce injuries or muscle pain from work. The working posture consisted of two main factors: 1) risk factors for abnormalities in the skeletal and muscular system, such as repetitive work, posture, exertion or weight bearing and resting time, and 2) personal factors involving various parts of the body such as hands, torso, neck, arms, legs, and knees. Commonly used posture assessments such as the RULA [4], REBA [5], and OWAS [6] were used to assess symptom risk factors. Abnormalities were found in the skeletal and muscular systems when the RULA assessment and the MSD survey were applied to rubber tapping [7]. Injuries occurred in almost

every part of the body, with the highest number recorded in the lower back. Ergonomic risks are used explicitly for rubber tapping work. This research focused on the ergonomic risk assessment and work posture of rubber tappers. Ergonomic principles were used to reduce work risks such as fatigue and work-related injuries and improve the health of rubber tappers.

2. Materials and Methods

An Ergonomic Risk Assessment Tool for Rubber Tappers (EART) was used together with a work posture assessment model for Rapid Upper Limb Assessment (RULA). The sample group of 154 rubber farmers in Pa Phayom District, Phatthalung Province, aged 20 to 60, was randomly selected by purposive sampling. They had 1 year or more experience in tapping rubber with the following inclusion criteria: perfect health, ability to read and write Thai well, and no history of injury or back pain. Exclusion criteria included people who were unable to participate throughout the whole length of the project, had a history of surgery on the shoulder, arm, hand, torso, abdomen, back, hip, or thigh, and suffered from congenital diseases related to bones and muscles. In general, the gesture of rubber tapping is shown in Figure 1.

2.1 Ethics approval

This study was approved by the Ethics Committee of Thaksin University (COA No. TSU 2024_070 REC No.0174). The purpose of the study was explained to all the participants, who signed informed consent forms before data collection.

2.2 Sample size

A survey was conducted among rubber farmers in Pa Phayom District, Phatthalung Province. The sample size was calculated using the Krejecie and Morgan equation (Krejecie & Morgan, 1970)[8]. The study population numbered 256 people, and the sample size was 154, calculated at a significance level of 0.05.

2.3 Statistical analysis

This research used descriptive statistics as percentage and mean values. Standard significance was evaluated at <0.05 according to the RULA and EART methods.

Figure 1. Rubber tapping gestures of a rubber farmer (Shoulder level)

2.4 RULA Method

RULA is widely used for work posture risk assessment using different body parts, namely the neck, arms, legs, and trunk. The RULA assessment provides total points on a scale from 1 to 7, with risk working postures shown in Table 1.

Table 1. Risk level of RULA

Risk categories	RULA score	Risk level	Action
1	1-2	Negligible	Acceptable posture if it is not repeated for a more extended period
2	3-4	Low	Further investigation and change may be needed in the future
3	5-6	Medium	Investigation posture and chafe needed soon
4	7	High	Investigation posture and chafe needed immediate

The RULA technique evaluates working posture according to ergonomic principles and assigns scores to different body parts. The upper arm, lower arm, wrist, and wrist twist scores are given in Table A, while neck, trunk, and leg scores are included in Table B, with the sum of muscle and force scores shown in Table C. The final RULA score assesses the posture risks of rubber tappers, as shown in Figure 2.

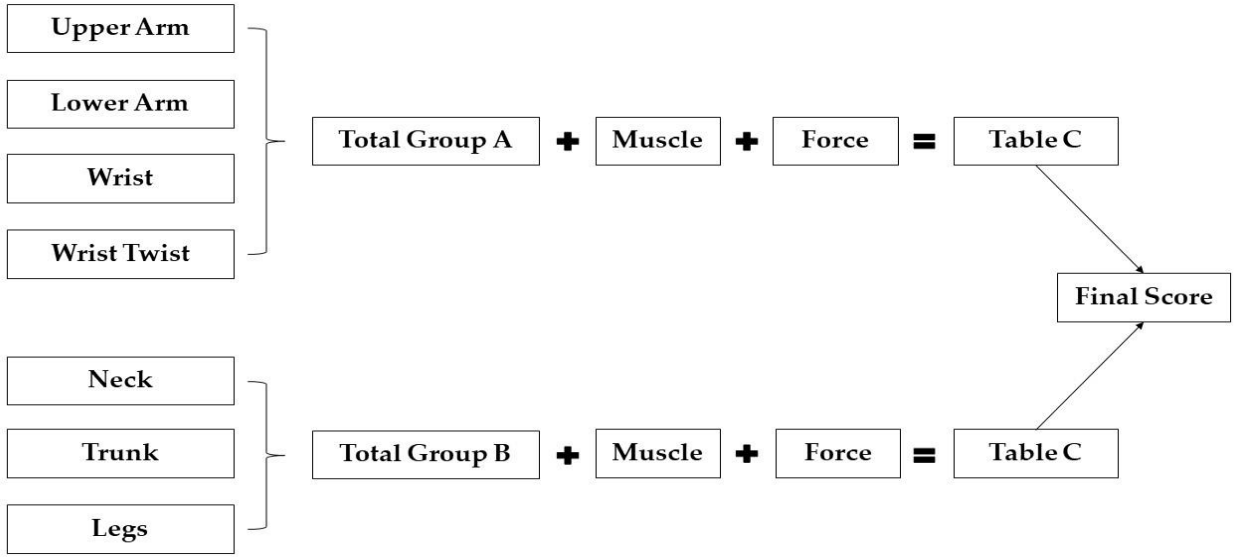


Figure 2. Body parts analyzed by RULA (Adapted from McAtamny and Corlett,1993)

2.5 EART Method

The ergonomic risk assessment tool for rubber tappers (EART)[9] was determined by the rubber tapping risk index as the ratio of the number of rubber trees tapped per day (RTD) divided by the recommended number of rubber trees for tapping per day (RTL), as shown Equation 1:

$$EART = RTD/RTL \quad (1)$$

The recommended number of rubber trees for tapping per day (RTL) was calculated from Equation 2, and the RTL value was substituted in Equation 1 to find the risk index for tapping as:

$$RTL = LC \times HM \times WM \times SM \quad (2)$$

when
HM = Tapping height multiplier factor
WM = Working area multiplier factor
SM = Stroke multiplier factor
LC = Load constant (Set at 700 trees per day for tapping system 1/3)

3. Criteria and procedures

The criteria and processes used to decide and handle the numerous variables involved in computing the rubber-tapping risk index are briefly illustrated and discussed. The styles used in this paper are:

2.5.1 Tapping height multiplier factor (HM)

Analysis of the height level of rubber tapping compared to the body was divided into four levels: 1) above the shoulder, 2) waist-shoulder, 3) knee-waist, and 4) below the knee. Para rubber tapping above shoulder level was assigned a factor of 0.93, rubber tapping at the waist to shoulder level was assigned a factor of 1, rubber tapping at the knee to waist level was assigned a factor of 0.44, and rubber tapping below the knee was assigned a factor of 0.60, as shown in Table 2.

Table 2. Tapping height multiplier factor (HM)

No.	tapping level	Multiplier factor
1	Above the shoulder	0.93
2	Waist –shoulder	1
3	Knee-waist	0.44
4	Below the knee*	0.60

* At a level below the knees, use a kneeling position.

2.5.2 *Working area multiplier factor (WM)*

A survey of rubber plantations and results of rubber tapping experiments suggested dividing the inclination into four levels: 1) 0–10°, 2) 11–20°, 3) 21–30°, and 4) >30° to determine the inclination factor of the work area. Rubber tapping in the 0–10° area used a factor equal to 1. Rubber tapping in the 11–20° area used a factor equal to 0.71, rubber tapping in the area 21–30° used a factor equal to 0.52, and rubber tapping in the inclination area >30° used a factor of 0.43, as shown in Table 3.

Table 3. Working area multiplier factor (WM)

No.	Degree	Multiplier factor
1	0-10°	1
2	11-20°	0.71
3	21-30°	0.52
4	>30°	0.43

2.5.3 Stroke multiplier factor (SM)

A survey of the rubber tapping stroke made by rubber tappers and the experimental results of rubber tapping defined the stroke interval into four ranges: 1) 10–20 times per tap, the rubber tap has a multiplier

value of 1, 2) between 21 and 30 times per tap, a multiplier of 0.80, 3) between 31 and 40 times per tap, a multiplier of 0.45, and 4) more than 40 times per tap, a multiplier value equal to 0.25, as shown in Table 4.

Table 4. Stroke multiplier factor (SM)

No.	Stroke	Multiplier factor
1	10-20	1
2	21-30	0.80
3	31-40	0.45
4	>40	0.25

2.5.4 Interpretation

The analysis results were divided into four levels to determine the rubber tapping evaluation criteria. An index value of less than or equal to 1, meaning no risk, indicated that rubber tapping was acceptable. An index value between 1.1 and 2.5, meaning low risk, suggested that rubber tapping was acceptable but may require additional monitoring assessment. An index between 2.6 and 3.5, meaning moderate risk, indicated that the rubber tapping operation must be improved and re-evaluated. An index value greater than or equal to 3.6, meaning high risk, that rubber tapping was unacceptable and must be improved immediately, as shown in Table 5.

Table 5. Risk levels in the assessment of rubber tappers

Risk categories	Exposure index	Risk level	Action
1	≤ 1	No risk	The tapping of rubber was acceptable.
2	1.1-2.5	Low risk	The tapping of rubber was acceptable but may require additional monitoring assessment.
3	2.6-3.5	Moderate risk	The rubber tapping operation must be improved and re-evaluated.
4	≥ 3.6	High risk	The tapping of rubber was unacceptable and must be improved immediately.

3. Results and Discussion

3.1 Demographic data

After collecting data from a sample of 154 people, it was found that the participants were 80 males and 74 females.. The mean age was 43.95 years. In addition, the mean of rubber tapping per day was 3.75 hr., as shown in Table 6.

Table 6. Demographic characteristics of the Rubber Tappers (n = 154)

Demographics	Frequency	Percentage
Sex		
Female	74	48.10
Male	80	51.90
Age (mean ± SD) yrs. = 43.95 ± 10.77		
Rubber tapping experience (mean ± SD) (yrs.) = 13.68 ± 8.87		
Rubber tapping per day (mean ± SD) (hr.) = 3.75 ± 1.52		

3.2 EART Method analysis

The results of the EART analysis found that tapping height multiplier factor (HM) had a mean of 0.68, SD = 0.24. The working area multiplier factor (WM) had a mean of 0.94, SD = 0.15. Stroke multiplier factor (SM) had a mean of 0.91, SD = 0.14, and when considering RTL values, it was found that there was a mean of 415, SD = 182. For the analysis of RTI values, it was found that there was a mean of 1.77 and an SD of 1.35. It was found that the value of Min. = 0.31 and the value of Max. = 8.64, which shows that farmers' rubber tapping has low-to-high risk, as shown in Table 7.

Table 7. The distribution of the multiplier factor of EART

	WM	HM	SM	RTL	RTI
Mean	.94	.68	.91	415	1.77
Std. Deviation	.15	.24	.14	182	1.35
Minimum	.52	.44	.25	98	.31
Maximum	2.00	1.00	1.00	1120	8.64

3.3 RULA Method analysis

The results of the RULA analysis gave score levels 2 (28.6%), 6 (24.0%), 3 (17.5%), 4 (16.2%), 5 (8.4%), and 7 (5.2%). Rubber farmers were at low risk, accounting for 44%, as shown in Table 8. The majority of the time, the RULA score was found to be 2, indicating that rubber tapping poses no risks. On the other hand, it was discovered that 37 individuals in the sample obtained a RULA score of 6, suggesting that they were at moderate risk and required quick changes to their work posture.

Table 8. The distribution of the RULA score

RULA score	Frequency	Percent
2	44	28.6
3	27	17.5
4	25	16.2
5	13	8.4
6	37	24.0
7	8	5.2

3.4 Risk categories by the EART and RULA methods

The risk categories for EART and RULA are shown in Figure 3. The EART assessed were 50(32.5%), 78(50.6%), 12(7.8%), and 14(9.1%) respectively, whereas 1, 2, 3 and 4. The RULA assessed were 44(28.6%), 52(33.8%), 50(32.5%), and 8(5.2%), respectively, which indicate that need to change posture immediately.

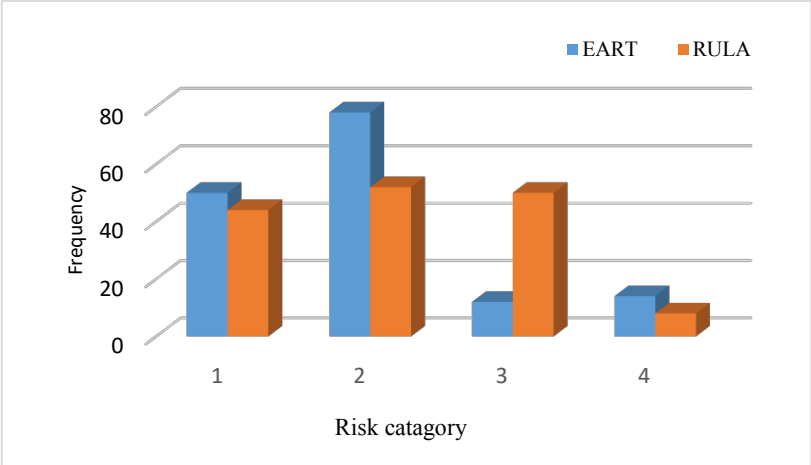


Figure 3. Distribution of risk categories by EART and RULA method

3.5 Mann-Whitney Test

Data on the estimated determined statistical significance of the differences between the crossing variables and the z-values are shown in Figure 4. The Z-value in this study was -2.371. The statistical significance of the differences, approximately calculated (p=0.018), was demonstrated to be the same. Therefore, statistically significant differences were found between the risk evaluation of the EART and RULA approaches in the research results. Comparison of risk assessment by EART and RULA methods

	Rubber tapping assessment
Mann-Whitney U	10112.000
Wilcoxon W	22047.000
Z	-2.371
Asymp. Sig. (2-tailed)	.018

3a. Grouping Variable: Risk assessment

Figure 4. Mann-Whitney Test analysis

4. Discussion

Risk assessment for rubber farmers was performed using the EART and RULA methods. The RULA analysis gave scores ranging from 2 to 7, meaning acceptable risk to high risk [10-11], and offered suggestions to change working posture[12]. However, rubber tapping is a specific gesture that depends on the rubber-tapping knife and the characteristics of the rubber tree, making it quite difficult to change the rubber-tapping posture. Consequently, using an ergonomic rubber tapping knife will be advantageous and lower the risk of tapping rubber [13]. The EART analysis gave scores ranging from 0.38 to 8.64, meaning that rubber farmers were exposed to no risk to high risk. Following the EART method, rubber trees were recommended for daily tapping amounts. To prevent the risk of working beyond the limits of the body from all three factors, if the results of the analysis reveal that the high level of rubber tapping is a risk, it is recommended that the incision be rubber-tapped using a double tapping system (Double Cut Alternative: DCA) [14-15] by alternating cuts on the rubber at the high and low tapping levels. As for the risk from the inclination of the area or rubber plantation, this research recommends that rubber farmers adjust the inclination of the area to be reduced or as close to 0° as possible. Furthermore, safety risks and working space must also be considered [16-17].

5. Conclusions

The study found that risk assessment for rubber farmers was done using the EART and RULA methods. Comparing the results, it was found that the RULA method focused on evaluating the hands, arms, legs, and body in the rubber-tapping posture. Most rubber farmers are at medium to high risk, and adjusting the posture for tapping rubber according to ergonomic principles will help reduce the risk of tapping rubber. In the EART method, the emphasis was on evaluating the multiplier factor from the rubber-tapping process. Most farmers are at risk, from low risk to high risk. However, due to the EART method, appropriate rubber trees will be recommended for tapping per day. Therefore, rubber farmers can improve the tapping factors, such as tapping height, working area, and stroke, when tapping rubber trees.

Suggestions

Future studies can evaluate the degree of complementarity of the EART and RULA methods based on inferential statistics. This research can be extended using Nordic Musculoskeletal. A questionnaire will link musculoskeletal illnesses with EART and RULA results.

Limitations

This research used a jabong knife to cut the rubber tree. The tapping system is 1/3 of the trunk. Therefore, you should also study other types of rubber-tapping knives and tapping systems. There should be research in various rubber plantation areas in the future.

6. Acknowledgements

The researcher would like to thank you. All advisors sacrifice their valuable time to give advice, advise, and edit with care and kindness. They also encourage the researcher to complete his research. Finally, the researcher would like to thank his beloved family, who have always encouraged researchers.

Author Contributions: Conceptualization, S.I.; methodology, S.I., A.S.,K.P.; software, S.I.; validation, A.S., K.P.;data curation, S.I.; writing—original draft preparation, S.I.; writing—review and editing, A.S.; visualization, A.S.,K.P.; supervision, K.P.. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest

References

[1] Ministry of Agriculture and Cooperatives. Agricultural statistics of Thailand in 2021. **2022**, Bangkok. (in Thai)

[2] Bensa-Ard, N.; Tuntiseranee, P.; Anuntaseree, S. Work conditions and prevalence of musculoskeletal pain among para-rubber planters: a case study in Tambon Nakleua, Kantang District, Trang Province. *Songklanagarind Medical Journal*, **2004**, *22*(2),101-110.

[3] Inraksa, S.; Sungkhapong, A.; Pochana, K. Prevalence and risk factors of musculoskeletal disorders in rubber tappers: A case study in Phatthalung Province, Thailand. *International Journal of Health and Medical Sciences*, **2017**, *3*(1),23-28.

[4] McAtamney, L.; Corlett, E.N.; RULA: a survey method for the investigation of work-related upper limb disorders. *Applied ergonomics*, **1993**, *24*(2),91-99.

[5] Hignett, S.; McAtamney, L. Rapid entire body assessment (REBA). *Applied ergonomics*, **2000**, *31*(2), 201-205.

[6] Karhu, O.; Kansi, P.; Kuorinka, I. Correcting working postures in industry: a practical method for analysis. *Applied ergonomics*, **1977**, *8*(4),199-201.

[7] Meksawi, S.; Tangtrakulwanich, B.; Chongsuvivatwong, V. Musculoskeletal problems and ergonomic risk assessment in rubber tappers: A community-based study in southern Thailand. *International Journal of Industrial Ergonomics*, **2012**, *42*(1),129-135.

[8] Krejcie, R.V.; Daryle W.M. Determining sample size for research activities. *Educational and psychological measurement*, **1970**, *30*(3), 607-610.

[9] Inraksa, S. Development of an Ergonomics Risk Assessment Model for Rubber Tappers (PhD's Thesis, Faculty of Engineering, Prince of Songkla University, **2022**.

[10] Kee, D.; Comparison of OWAS, RULA and REBA for assessing potential work-related musculoskeletal disorders. *International Journal of Industrial Ergonomics*, **2021**, 83, 103140.

[11] Brazil, Cristiane K., et al. Using the rapid upper limb assessment to examine the effect of the new hotel housekeeping California standard. *Applied Ergonomics*, **2023**, 106, 103868.

[12] Meksawi, S.; Tangtrakulwanich, B.; Chongsuvivatwong, V. Musculoskeletal disorder and rapid upper limb assessment scoring among rubber tappers. *Ergonomics Internationl Journal*, **2018**, *2*(6), 1-7.

[13] Pramchoo, W.; Geater, A. F.; Harris-Adamson, C.; Tangtrakulwanich, B. Ergonomic rubber tapping knife relieves symptoms of carpal tunnel syndrome among rubber tappers. *International Journal of Industrial Ergonomics*, **2018,** *68*, 65-72.

[14] Nhean, S.; Ayutthaya, S. I. N.; Songsri, P.; Gonkhamdee, S.; Sdoodee, S. First testing of the double cut alternative tapping system on rubber tree clone RRIM 600 in marginal area, Northeast Thailand. *KKU Research Journal,* **2016**, *21*(3), 28-35.

[15] Chantuma, P.; Lacote, R.; Leconte, A.; Gohet, E. The "double cut alternative" (DCA) tapping system: an innovative tapping system designed for Thai rubber smallholdings using high tapping frequency. *Paper presented at the IRRDB International Rubber Conference*, Chaing Mai, Thailand, **2011**, pp 14–17

[16] Ali, M.; Ramazan, M.; Hossein, A. A Survey of Health, Safety and Environment (HSE)Management and Safety Climate in Construction Sites. *Engineering, Technology & Applied Science Research*, **2017**, 7(1), 1334-1337.

[17] Mohammad, R.; Gholamali, S.; Abdolhosein, H.; Ehsan, A. Sensitivity Analysis of Workspace Conflicts According to Changing Geometric Conditions. *Engineering, Technology & Applied Science Research*, **2017**, *7*(1), 1429-1435.

ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

The Potential of Near-infrared Spectroscopy to Predict Soil Nutrient Contents Based on Soil Color

Piyamas Khammao[1], Wutthida Rattanapichai[2], Roongroj Pitakdantham[3], Poonpipope Kasemsap[4], and Kannika Sajjaphan[5*]

[1] Department of Soil Science Faculty of Agriculture, Kasetsart University, Bangkok, 10900, Thailand
[2] Department of Soil Science, Faculty of Agriculture, Kasetsart University, Bangkok, 10900, Thailand
[3] Department of Soil Science, Faculty of Agriculture, Kasetsart University, Bangkok, 10900, Thailand
[4] Department of Horticulture, Tropical Agriculture, Kasetsart University, Bangkok, 10900, Thailand
[5] Department of Soil Science and Center for Advanced Studies in Agriculture and Food, Kasetsart University, Bangkok, 10900, Thailand
* Correspondence: agrkks@ku.ac.th

Citation:
Khammao, P.; Rattanapichai, W.; Pitakdantham, R.; Kasemsap, P; Sajjaphan, K. The potential of near-infrared spectroscopy to predict soil nutrient content based on soil color. *ASEAN J. Sci. Tech. Report.* **2024**, *27*(5), e252637. https://doi.org/10.55164/ajstr.v27i5.252637.

Article history:
Received: February 1, 2024
Revised: July 23, 2024
Accepted: July 26, 2024
Available online: August 29, 2024

Abstract: Near-infrared spectroscopy (NIR) analysis in laboratory-based settings has the potential to predict soil elements. The aim was to explore the effects of soil color on the prediction of total nitrogen (N), available phosphorus (P), and extractable potassium (K) contents using near-infrared spectroscopy in the range of 1000–2500 nm. Two hundred forty soil samples were collected from a paddy field in northeast Thailand. We divided the soil samples based on soil color using the Munsell color chart to construct a model to predict nutrient contents based on soil color. Regression models for soil nutrient contents were developed using partial least squares regression (PLSR) models. The best predictions were obtained for N (R^2 = 0.87, RMSE = 0.131), P (R^2 = 0.87, RMSE = 7.713) and K (R^2 = 0.77, RMSE = 14.944). This research demonstrates the viability of employing Near-Infrared spectroscopy (NIRs) as a reasonable method for predicting soil nutrient contents.

Keywords: Soil color; Paddy soil; Soil nutrient contents; Near infrared; Partial square regression

1. Introduction

In recent years, there has been a significant demand for soil analysis methodologies that are precise, rapid, and pollution-free. This is because soil data information can be utilized for environmental monitoring, soil quality assessment, and precision agriculture [1, 2]. For this reason, near-infrared spectroscopy (NIRs) is considered an alternative to improve or complement traditional methods of soil analysis. Near-infrared spectroscopy has emerged over the past few decades as a rapid and robust analytical method for various agricultural applications [3]. In particular, this technique can assess various soil fertility properties simultaneously with a single spectrum, making reflectance infrared spectroscopy fast, time-saving, cost-effective, and efficient. In the NIR region, radiation is absorbed by different chemical bonds present in the sample, such as C–H, N–H, S–H, C=O, and O–H. Furthermore, the radiation is absorbed in a manner conforming with the concentration of these compounds. As a result, NIR reflectance spectra provide information about the organic composition of a soil sample. Nevertheless, NIR information cannot be directly inferred from the obtained spectra. NIR reflectance spectroscopy depends on calibrations and chemometrics techniques that employ absorbances at multiple wavelengths to

predict particular characteristics of a sample [4]. Thus, research on using near-infrared (NIR) spectroscopy in soil science has rapidly increased to determine soil properties. Several authors have demonstrated the efficacy of NIR reflectance spectroscopy in predicting macro- and micronutrients in soils [2, 5–11].

Soil color is a crucial indicator of soil properties and processes that reflect chemical, physical, and biological characteristics [12–14]. The three principal constituents of soil color are humus (black), calcium carbonates (white), and iron oxide (red or yellow). However, other soil components, such as manganese oxides, nitrogen oxides, and phosphorus oxides, can also contribute to soil color. These important nutrients can be identified by color variables such as lightness [15–18]. Moreover, soil texture, organic matter content, moisture level, and erosion influence soil color [19–21]. Soil color is commonly evaluated by a human observer visually comparing the color of a soil sample to the color chips specified by the Munsell Color System [22]. The Munsell color chips are organized based on the hue, value, and chroma color components, and the method for measuring soil color is elaborated in detail in Soil Science Division Staff [23]. Previous studies have demonstrated strong relationships between soil properties and the spectral reflectance of soils in the visible and near-infrared regions [24–31]. Thus, the goal of this paper is to investigate the potential of near-infrared spectroscopy (NIRs) to predict total nitrogen (N), available phosphorus (P), and extractable potassium (K) in paddy soil, using soil color as a criterion to divide the soil samples into a calibration data set and a validation data set.

2. Materials and Methods

2.1 Study area, soil sampling collection, and chemical analyses

The soil sample in this study is paddy soil from northeast Thailand. It covers eight provinces, including Sakon Nakhon, Phanom, Amnat Charoen, Ubon Rachathani, Sisaket, Surin, Buriram, and Roi Et, with different soil groups. The samples were collected at depths of 0 –15 cm. In total, 240 samples were used for this experiment. The soil samples were air-dried and sieved using a 2-mm sieve. The laboratory chemicals were analyzed for the total nitrogen (N) content, which was determined by the Kjeldahl method [32]. The available phosphorus (P) content was determined by the Bray II method [33]. The method described by Jackson and Chen [32] measured the extractable potassium (K) content. Table 1. shows the summary statistics of the chemical analysis of soil N, P, and K.

Table 1. Descriptive statistics data of soil fertility used in this study.

Soil color group	Soil nutrients	Min	Max	Mean	SD
Group one	N	0.21	1.40	0.68	0.37
(10YR Value 3-5)	P	3.16	126.25	39.49	34.01
	K	0.04	442.25	71.37	64.14
Group two	N	0.07	1.68	0.68	0.37
(10YR Value 6-7)	P	1.42	80.68	17.81	13.77
	K	18.41	158.69	46.65	26.50
Group three	N	0.07	6.30	0.68	0.94
(7.5YR Value 5-7)	P	5.23	63.47	17.81	9.95
	K	15.51	175.31	53.25	38.45
Group four	N	0.07	1.09	0.46	0.25
(5YR Value 6-7)	P	1.66	106.95	15.57	21.12
	K	0.01	117.35	26.23	30.86

N = total nitrogen (g kg^{-1})
P = available phosphorus (mg kg^{-1})
K = extractable potassium (mg kg^{-1})

2.2 Spectrum determination and soil color

Spectral measurements in the 1000 – 2500 nm range were made with a Fourier-transform near-infrared (FT-NIR) spectrophotometer (Buchi N-500 NIRFlex; Switzerland). The measurements were conducted with a spectral resolution of 4 nm. The soil samples were placed in petri dishes and smoothed surfaces before the spectral measurement. We obtained the soil spectra in reflectance by averaging three scans for each sample. Subsequently, we converted the reflectance spectra to an absorbance spectrum (A) using $A = \log_{10}(1/R)$. Before modeling, we divided the soil samples based on the Munsell color chart to create a model that predicts nutrient contents depending on soil color. Color can be represented using three-dimensional color space models. Soil color is typically described in both dry and moist conditions using the Munsell color system, which is based on three parameters: hue (dominant spectral color), value (lightness), and chroma (color purity) [22]. For Munsell, soil color was determined in the laboratory for moist soil samples, and the Munsell soil color was chosen for the closest chip. As a result, we can divide soil color into four groups: group one, which has a hue of 10YR and a value of 3 to 5 (~10YR Value 3-5) totaling 60 samples; group two, which has a hue of 10YR and a value of 6 to 7 (~10YR Value 6-7) totaling 60 samples; group three, which has a hue of 7.5YR and a value of 5 to 7 (~7.5YR Value 5-7) totaling 60 samples; and group four, which has a hue of 5YR and a value of 6 to 7 (~5YR Value 6-7) totaling 60 samples.

2.3 Calibration Model

For model construction, we utilized Partial Least Squares regression (PLSR). PLSR is commonly employed as a chemometric technique in Near-Infrared (NIR) analysis [5, 35–37]. The PLS regression is a multivariate regression modeling method suggested by [38]. This method can be employed to address the issue of co-linearity among independent variables. The PLSR can effectively extract latent variables (LVs) by excluding unexplainable information, ensuring that the LVs have the most dominant ability to explain the dependent variables [39].

In this study, a regression model was constructed for each soil color group, including group one (10YR Value 3-5), group two (10YR Value 6-7), group three (7.5YR Value 5-7), and group four (5YR Value 6-7). The NIR spectral data (1000 – 2,500 nm) were used as the independent variable, while the N, P, and K contents were used as the dependent variables. The LVs were extracted from the independent variables related to the dependent variables [40]. The established calibration models were assessed and quantified by validating using cross-validation. The model performance was then evaluated using statistical parameters such as the coefficient of determination (R^2) and root mean square error (RMSE). R^2 reflects the model's ability to interpret sample spectra, while RMSE refers to the errors between the predicted and actual nutrient contents. Generally, the closer the R^2 value is to 1, the closer the RMSE value is to 0, indicating better model performance [41].

3. Results and Discussion

3.1 Effects of soil color on spectral behavior

The spectral behavior of soil varies depending on a combination of factors such as reflectance intensity (albedo), absorption features (depth and amplitude), and spectral shape. These characteristics are influenced by the soil's physical, chemical, and mineralogical properties [42]. The absorbance behavior of four soil color groups is shown in Figure 1. The spectra of group one soil color (10YR 3-5) had a higher absorbance when compared to group two, group three, and group four. Based on the studies of soil series characters in group one samples, it is demonstrated that the soil is fine-textured soil (higher clay content) and poorly drained, adversely affecting the soil's color and resulting in a darker hue. The differences in soil particle size were the primary cause of the variations in absorbance intensity, where soils with higher clay content showed more absorbed energy across the spectrum [43]. The spectral characters of Group Two, Group Three, and Group Four have a similar and low absorbed energy compared to Group One. This shows that groups two, three, and four had a higher sand content than group one.

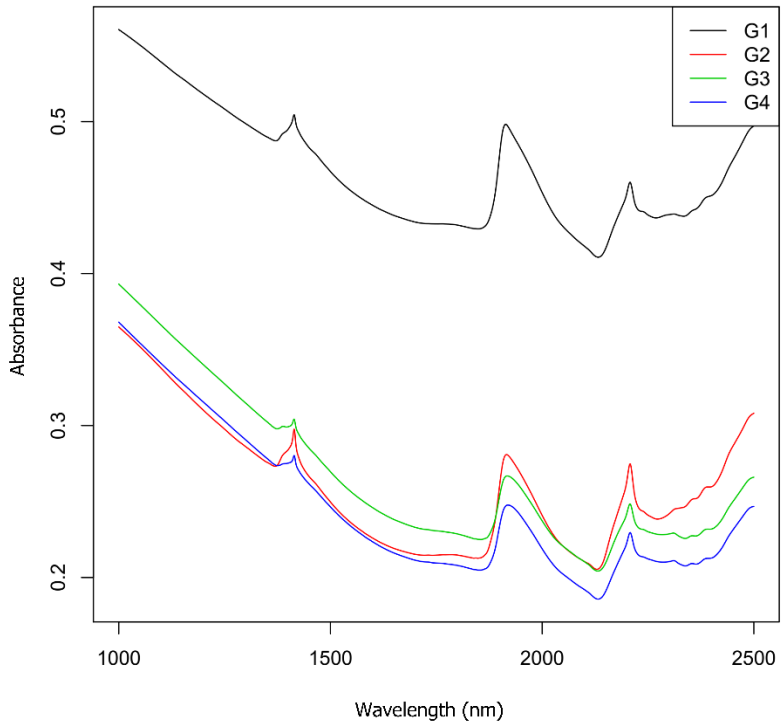


Figure 1. Spectra of the four soil color groups
G1 = soil color group one (10YR 3-5)
G2 = soil color group two (10YR 6-7)
G3 = soil color group three (7.5YR 5-7)
G4 = soil color group four (5YR 6-7)

However, the spectra of the four soil colors had similar behavior. The spectra exhibited robust absorption peaks at 1,400, 1,900, and 2,200 nm, attributed to molecular vibrations of hydroxyl (OH^-) groups [44, 45]. The absorption feature at 1,900 nm was more pronounced, possibly due to water (H_2O) in interstratified minerals [46]. Whiting et al. [45] have suggested that a stronger absorption intensity at 1,900 nm indicates the prevalence of structural H_2O in 2:1 minerals, such as montmorillonite and vermiculite. The absorption feature at 2,200 nm indicates kaolinite dominance [46]. The soil mineralogy has been characterized by 2:1 type clays, including illite, smectite, and vermiculite with hydroxy interlayers. However, most of the soil is dominated by 1:1-type clays [46–49]. The high activity of clay fraction (CFA) in some horizons was found to be associated with not only 2:1 clay but also 1:1 + 2:1 clay (interstratified minerals), as reported by [49].

Two types of minerals, namely 2:1 minerals or interstratified ones, exhibited a feature at 1400 nm, in addition to one sharper feature at 1900 nm and another with a more elongated shape at 2200 nm. The presence of both minerals simultaneously makes their identification difficult. However, in these soils, 2:1 clay minerals such as smectites and vermiculite are adsorbed at the surface by 1:1 minerals. The 2:1 minerals are found in the fine clay fraction (<0.2 mm) and are more likely to migrate in the soil due to the high content of exchangeable Na and Mg [46].

3.2 Accuracy of color to predict soil nutrients

Partial least squares regression (PLSR) analysis was performed on soil color and soil absorbance to predict the contents of total nitrogen (N), available phosphorus (P), and extractable potassium (K). The nutrient content prediction results for the four soil color groups are presented in Table 2. The modeling accuracy shows that the four soil color groups performed satisfactorily in predicting N, P, and K. The N prediction model showed an R^2 and RMSE ranging from 0.65 to 0.87 and 0.102 to 0.567 g kg^{-1}. The P prediction

model showed an R^2 and RMSE ranging from 0.35 to 0.87 and 7.713 to 15.314 mg kg^{-1}, and the K prediction model showed an R^2 and RMSE ranging from 0.47 to 0.77 and 14.944 to 45.237 mg kg^{-1}. Moreover, our results show that the soil color in group two (10YR Value 6-7) was the best model to predict N. The R^2 values were 0.87, and the RMSE values were 0.131 g kg^{-1}. The soil color in group four (5YR Value 6-7) was the best model to predict P and K. The R^2 values were 0.87 and 0.77, and the RMSE values were 7.713 mg kg^{-1} and 14.944 mg kg^{-1}. The models show that the soil color can predict soil nutrients. It has been shown in many research. For example, Franzmeier [50] reported correlations between soil organic matter and Munsell value and chroma with R^2 values of 0.48. Lindbo et al. [51] used a chroma meter to measure soil color, organic carbon, and hydromorphology correlations. They reported an R^2 value of 0.63 for the correlation between dry Munsell value and soil organic carbon. Konen et al. [20] used a chroma meter to develop correlations between soil color, organic carbon, and texture. They showed logarithmic correlations between reflectance, Munsell value, Munsell chroma, and soil organic carbon. The R^2 values of their correlations ranged from 0.68 to 0.77. Moreover, Liles et al. [52] reported that soil type and parent materials influenced the lightness of the soil. They analyzed the relationship between the lightness of the soil and the total C content in forest soil around northern California. Their findings demonstrated that the R^2 for the relationship between soil C% and the lightness value varied with different soil types and parent materials. For instance, the R^2 values were 0.34 for all samples, 0.83 for Inceptisols, 0.6 for Andisols, 0.036 for Alfisols, and 0.35 for Ultisols. Schulze et al. [21] highlighted those variations in regression equations that were significantly influenced by soil texture and landscape. Attempting to predict nutrient contents across diverse soil types and landscapes using a single equation is often challenging. Within a specific landscape, the primary soil-forming factors include topography and the texture of the parent material. However, when considering broader landscapes, the key factors shift to encompass parent materials and vegetation.

Table 2. Accuracy of the prediction model based on soil color for the three soil properties.

Soil color group	Soil nutrients	R^2	RMSE
Group one	N	0.75	0.146
(10YR Value 3-5)	P	0.80	15.314
	K	0.50	45.237
Group two	N	0.87	0.131
(10YR Value 6-7)	P	0.44	10.456
	K	0.67	15.102
Group three	N	0.65	0.567
(7.5YR Value 5-7)	P	0.35	8.000
	K	0.47	29.659
Group four	N	0.84	0.102
(5YR Value 6-7)	P	0.87	7.713
	K	0.77	14.944

R^2 = coefficient of determination
RMSE = root mean square error
N = total nitrogen (g kg^{-1})
P = available phosphorus (mg kg^{-1})
K = extractable potassium (mg kg^{-1})

4. Conclusion

The NIR models we developed for predicting soil nutrient contents (total nitrogen, available phosphorus, and extractable potassium) use soil color as a predictor. The prediction of total nitrogen content in soil color group two (10YR Value 6-7) outperformed the prediction for total nitrogen in other soil color groups. For available phosphorus and extractable potassium, the best predictions were obtained from soil color group four (5YR Value 6-7). This demonstrates that combining NIRs with soil color can predict soil nutrient contents accurately. This method is efficient and nondestructive. It serves as an alternative to

traditional approaches. In addition, although the model produced accurate predictions, its accuracy and robustness for future practical applications need to be validated in other study areas with more samples.

5. Acknowledgements

The authors gratefully acknowledge Jean-Michel Roger and all Soil Science staff, Faculty of Agriculture, Kasetsart University, for providing the research tools and assistance.

Author Contributions: Conceptualization, P.K, K.S and W.R.; methodology, K.S., W.R., P.K.; software, K.S.; validation, K.S., W.R., and P.K.; formal analysis, P.K; investigation, R.P.; resources, K.S., W.R.; writing—original draft preparation, P.K.; writing—review and editing, K.S., W.R. and P.K.; visualization, K.S., W.R. and P.K.; supervision, K.S.; funding acquisition, P.K. All authors have read and agreed to the published version of the manuscript.

Funding: The authors are grateful to the Office of the Ministry of Higher Education, Science, Research and Innovation, the Thailand Science Research and Innovation Fund through the Kasetsart University Reinventing University Program 2021, and the Agricultural Research Development Agency for financial support.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Cohen, M.J.; Prenger, J.P.; DeBusk, W.F. Visible-Near Infrared Reflectance Spectroscopy for Rapid, Nondestructive Assessment of Wetland Soil Quality. *J. Environ. Qual.* **2005**, *34*, 1422–1434. https://doi.org/10.2134/jeq2004.0353

[2] Viscarra Rossel, R.A.; Walvoort, D.J.J.; McBratney, A.B.; Janik, L.J.; Skjemstad, J.O. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*. **2006**, *131*, 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007

[3] Blanco, M.; Villarroya, I. NIR spectroscopy: A rapid-response analytical tool. *TrAC - Trends Anal. Chem*. **2002**, *21*, 240–250. https://doi.org/10.1016/S0165-9936(02)00404-1

[4] Batten, G.D. Plant analysis using near infrared reflectance spectroscopy: The potential and the limitations. Aust. *J. Exp. Agric*. **1998**, *38*, 697–706 . https://doi.org/10.1071/EA97146

[5] He, Y., Huang, M., García, A., Hernández, A., Song, H.: Prediction of soil macronutrients content using near-infrared spectroscopy. *Comput. Electron. Agric.* **2007**, *58*, 144–153. https://doi.org/10.1016/j.compag.2007.03.011

[6] Johnson, J.M.; Sila, A.; Senthilkumar, K.; Shepherd, K.D.; Saito, K. Application of infrared spectroscopy for estimation of concentrations of macro- and micronutrients in rice in sub-Saharan Africa. *F. Crop. Res.* **2021**, *270*, 108222. https://doi.org/10.1016/j.fcr.2021.108222

[7] Kim, Y.J.; Choi, C.H. The analysis of paddy soils in Korea using visible-near infrared spectroscopy for development of real-time soil measurement system. *J. Korean Soc. Appl. Biol. Chem.* **2013**, *56*, 559–565. https://doi.org/10.1007/s13765-013-3067-z

[8] Malmir, M.; Tahmasbian, I.; Xu, Z.; Farrar, M.B.; Bai, S.H. Prediction of soil macro- and micro-elements in sieved and ground air-dried soils using laboratory-based hyperspectral imaging technique. *Geoderma*. **2019**, *340*, 70–80. https://doi.org/10.1016/j.geoderma.2018.12.049

[9] Salazar, O.; Benvenuto, A.; Fajardo, M.; Fuentes, J.P.; Nájera, F.; Celedón, A.; Pfeiffer, M.; Renwick, L.L.R.; Seguel, O.; Tapia, Y.; Casanova, M. Evaluation of a miniaturized portable NIR spectrometer for the prediction of soil properties in Mediterranean central Chile. *Geoderma Reg.* **2023**, *34*, e00675. https://doi.org/10.1016/j.geodrs.2023.e00675

[10] Peng, Y.; Zhao, L.; Hu, Y.; Wang, G.; Wang, L.; Liu, Z. Prediction of soil nutrient contents using visible and near-infrared reflectance spectroscopy. ISPRS Int. *J. Geo-Information.* **2019**, *8*. https://doi.org/10.3390/ijgi8100437

[11] Munawar, A.A.; Yunus, Y.; Devianti, Satriyo, P. Calibration models database of near infrared spectroscopy to predict agricultural soil fertility properties. *Data Br.* **2020**, *30*, 105469. https://doi.org/10.1016/j.dib.2020.105469

[12] Ibáñez-Asensio, S.; Marqués-Mateu, A.; Moreno-Ramón, H.; Balasch, S. Statistical relationships between soil colour and soil attributes in semiarid areas. *Biosyst. Eng.* **2013**, *116*, 120–129. https://doi.org/10.1016/j.biosystemseng.2013.07.013

[13] Ketterings, Q.M.; Bigham, J.M. Soil Color as an Indicator of Slash-and-Burn Fire Severity and Soil Fertility in Sumatra, Indonesia. *Soil Sci. Soc. Am. J.* **2000**, *64*, 1826–1833. https://doi.org/10.2136/sssaj2000.6451826x

[14] Schmidt, S.A.; Ahn, C. Analysis of soil color variables and their relationships between two field-based methods and its potential application for wetland soils. *Sci. Total Environ.* **2021**, *783*, 147005. https://doi.org/10.1016/j.scitotenv.2021.147005

[15] Christensen, L.K.; Bennedsen, B.S.; Jørgensen, R.N.; Nielsen, H. Modelling nitrogen and phosphorus content at early growth stages in spring barley using hyperspectral line scanning. *Biosyst. Eng.* **2004**, *88*, 19–24. https://doi.org/10.1016/j.biosystemseng.2004.02.006

[16] Schwertmann, U. Relations between iron oxides, soil color, and soil formation. Soil Color. Proc. *Symp. San Antonio*. **1993**. 51–69.

[17] Simonson, R.W. Soil color standards and terms for field use-history of their development. Madison: Soil Society of America. **1993**.

[18] Torrent, J., Barrón, V.: The visible diffuse reflectance spectrum in relation to the color and crystal properties of hematite. *Clays Clay Miner.* **2003**, *51*, 309–317. https://doi.org/10.1346/CCMN.2003.0510307

[19] Brady, N.C.; Weil, R.R. The nature and properties of soils. New Jersey Prentice Hall. **2006**.

[20] Konen, M.E.; Burras, C.L.; Sandor, J.A. Organic Carbon, Texture, and Quantitative Color Measurement Relationships for Cultivated Soils in North Central Iowa. *Soil Sci. Soc. Am. J.* **2003**, *67*, 1823–1830. https://doi.org/10.2136/sssaj2003.1823

[21] Schulze, D.G.; Nagel, J.L.; Van Scoyoc, G.E.; Henderson, T.L.; Baumgardner, M.F.; Stott, D.E. Significance of organic matter in determining soil colors. , Madison: Soil Society of America. 1993.

[22] Munsell Color: Munsell Soil Color Charts, **2000**, Revised Washable Edition. , Gretagmacbeth, New Windsor, NY.

[23] Soil Science Division Staff. Soil survey manual. C. Ditzler, K. Scheffe, and H.C. Monger (eds.). USDA Handbook 18. , Government Printing Office, Washington, D.C. **2017**.

[24] Stoner, E.R.; Baumgardner, M.F. Characteristic Variations in Reflectance of Surface Soils. *Soil Sci. Soc. Am. J.* **1981**, *45*, 1161–1165. https://doi.org/10.2136/sssaj1981.03615995004500060031x

[25] Baumgardner, M.F.; Silva, L.R.F.; Biehl, L.L.; Stoner, E.R. Reflectance properties of soils. *Adv. Agron.* **1986**, 38, 1–44. https://doi.org/10.1016/S0065-2113(08)60672-0

[26] Shields, J.A.; Paul, E.A.; Arnaud, R.J.S.; Head, W.K. Spectrometric Measurment of Soil Color and its Relationship to Moisture and Organic Matter. *Can. J. Sci.* **1968**, *48*, 271–280.

[27] Condit, H.R. Spectral Reflectance of American Soils. *Photogramm Eng.* **1970**, *36*, 955–966.

[28] Moritsuka, N.; Matsuoka, K.; Katsura, K.; Sano, S.; Yanai, J. Soil color analysis for statistically estimating total carbon, total nitrogen and active iron contents in Japanese agricultural soils. *Soil Sci. Plant Nutr.* **2014**, 60, 475–485. https://doi.org/10.1080/00380768.2014.906295

[29] Mouazen, A.M.; Karoui, R.; Deckers, J.; De Baerdemaeker, J.; Ramon, H. Potential of visible and near-infrared spectroscopy to derive colour groups utilising the Munsell soil colour charts. *Biosyst. Eng.* **2007**, *97*, 131–143. https://doi.org/10.1016/j.biosystemseng.2007.03.023

[30] Gholizadeh, A.; Saberioon, M.; Viscarra Rossel, R.A.; Boruvka, L.; Klement, A. Spectroscopic measurements and imaging of soil colour for field scale estimation of soil organic carbon. *Geoderma*. **2020**, *357*, 113972. https://doi.org/10.1016/j.geoderma.2019.113972

[31] Viscarra Rossel, R.A.; Fouad, Y.; Walter, C. Using a digital camera to measure soil organic carbon and iron contents. *Biosyst. Eng*. **2008**, *100*, 149–159. https://doi.org/10.1016/j.biosystemseng.2008.02.007

[32] Jackson, P.E.; Krol, J.; Heckenberg, A.L.; Mientijes, M.; Staal, W. Determination of total nitrogen in food, environmental and other samples by ion chromatography after Kjeldahl digestion. *J. Chromatogr. A*. **1991**, *546*, 405–410. https://doi.org/10.1016/S0021-9673(01)93039-0

[33] Bray, R.H.; Kurtz, L.T.: Determination of total, organic and available forms of phosphorus in soils. **1945**.

[34] Jackson, K.W.; Chen, G. Atomic absorption, atomic emission, and flame emission spectrometry. *Anal. Chem*. **1996**, *68*, 231–256. https://doi.org/10.1021/a1960012l

[35] Shao, Y.; He, Y. Nitrogen, phosphorus, and potassium prediction in soils, using infrared spectroscopy. *Soil Res*. **2011**, *49*, 166–172. https://doi.org/10.1071/SR10098

[36] Pudełko, A.; Chodak, M. Estimation of total nitrogen and organic carbon contents in mine soils with NIR reflectance spectroscopy and various chemometric methods. *Geoderma*. **2020**, *368*. https://doi.org/10.1016/j.geoderma.2020.114306

[37] Chen, Z.; Ren, S.; Qin, R.; Nie, P. Rapid Detection of Different Types of Soil Nitrogen Using Near-Infrared Hyperspectral Imaging. *Molecules.* **2022**, *27*, 2017. https://doi.org/10.3390/molecules27062017

[38] Wold, S.; Martens, H.; Wold, H. A multivariate calibration problem in analytical chemistry solved by the PLS method. *Lect. Notes Math.* **1983**, *46*, 286–293.

[39] Wang, H.; Liu, Q.; TU, Y. Identification of Optimal Subspace from PLS Regression. *J. Beijing Univ.* Aeronaut. Astronaut. (In China). **2000**, *26*, 473.

[40] Zhang, Y., Li, M.Z., Zheng, L.H., Zhao, Y., Pei, X.: Soil nitrogen content forecasting based on real-time NIR spectroscopy. *Comput. Electron. Agric.* **2016**, *124*, 29–36. https://doi.org/10.1016/j.compag.2016.03.016

[41] He, H.J.; Wu, D.; Sun, D.W. Rapid and non-destructive determination of drip loss and pH distribution in farmed Atlantic salmon (Salmo salar) fillets using visible and near-infrared (Vis-NIR) hyperspectral imaging. *Food Chem*. **2014**, *156*, 394–401. https://doi.org/10.1016/j.foodchem.2014.01.118

[42] Demattê, J.A.M. Characterization and discrimination of soils by their reflected electromagnetic energy. *Pesqui. Agropecu. Bras*. **2002**, *37*, 1445–1458. https://doi.org/10.1590/S0100-204X2002001000013

[43] Sørensen, L.K.; Dalsgaard, S. Determination of Clay and Other Soil Properties by Near Infrared Spectroscopy. *Soil Sci. Soc. Am. J. 2005*, 69, 159. https://doi.org/10.2136/sssaj2005.0159

[44] Stenberg, B.; Viscarra Rossel, R.A.; Mouazen, A.M.; Wetterlind, J. Visible and Near Infrared Spectroscopy in Soil Science. *Adv. Agron.* **2010**, *107*, 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7

[45] Whiting, M.L.; Li, L.; Ustin, S.L. Predicting water content using Gaussian model on soil spectra. *Remote Sens. Environ*. **2004**, *89*, 535–552. https://doi.org/10.1016/j.rse.2003.11.009

[46] Demattê, J.A.M.; Horák-Terra, I.; Beirigo, R.M.; Terra, F. da S.; Marques, K.P.P.; Fongaro, C.T.; Silva, A.C.; Vidal-Torrado, P. Genesis and properties of wetland soils by VIS-NIR-SWIR as a technique for environmental monitoring. *J. Environ. Manage.* **2017**, *197*, 50–62. https://doi.org/10.1016/j.jenvman.2017.03.014

[47] Nascimento, A.F. do; Furquim, S.A.C.; Couto, E.G.; Beirigo, R.M.; Oliveira Júnior, J.C. de; Camargo, P.B. de; Vidal-Torrado, P. Genesis of textural contrasts in subsurface soil horizons in the Northern Pantanal-Brazil. *Rev. Bras. Ciência do Solo.* **2013**, *37*, 1113–1127. https://doi.org/10.1590/s0100-06832013000500001

[48] Nascimento, A.F.; Furquim, S.A.C.; Graham, R.C.; Beirigo, R.M.; Oliveira Junior, J.C.; Couto, E.G.; Vidal-Torrado, P. Pedogenesis in a Pleistocene fluvial system of the Northern Pantanal - Brazil. *Geoderma.* **2015**, 58–72. https://doi.org/10.1016/j.geoderma.2015.04.025

[49] De Arruda Oliveira Coringa, E.; Couto, E.G.; Torrado, P.V. Soil geochemistry of the northern pantanal, mato grosso, Brazil. *Rev. Bras. Cienc. do Solo*. **2014**, *38*, 1784–1793. https://doi.org/10.1590/s0100-06832014000600013

[50] Franzmeier, D. Relation of Organic Matter Content to Texture and Color of Indiana Soils. *1998*.

[51] Lindbo, D.L.; Rabenhorst, M.C.; Rhoton, F.E. Soil color, organic carbon, and hydromorphology relationships in sandy epipedons. *Quantifying Soil Hydromorphology*. **1998**, *54*, 95–105. https://doi.org/10.2136/sssaspecpub54.c6

[52] Liles, G.C.; Beaudette, D.E.; O'Geen, A.T.; Horwath, W.R. Developing predictive soil C models for soils using quantitative color measurements. *Soil Sci. Soc. Am. J.* **2013**, *77*, 2173–2181. https://doi.org/10.2136/sssaj2013.02.0057

ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

Comparative Analysis of DNN, GBT, and KNN Models for Network Intrusion Detection

Rattikan Viboonpanich[1], Amornvit Vatcharaphrueksadee[2*], and Wilairat Charoenmairungrueang[3]

[1] Faculty of Information Technology and Digital Innovation, North Bangkok University, Bangkok, 12130, Thailand
[2] Faculty of Information Technology and Digital Innovation, North Bangkok University, Bangkok, 12130, Thailand
[3] Faculty of Information Technology and Digital Innovation, North Bangkok University, Bangkok, 12130, Thailand
* Correspondence: amornvit.va@northbkk.ac.th

Citation:
Viboonpanich, R.; Vatcharaphrueksadee, A.; Charoenmairungrueang, W. Comparative analysis of DNN, GBT, and KNN models for network intrusion detection. *ASEAN J. Sci. Tech. Report.* **2024**, *27*(5), e252675. https://doi.org/10.55164/ajstr.v27i5.252675

Received: February 9, 2024
Revised: July 22, 2024
Accepted: August 8, 2024
Available online: August 27, 2024

Abstract: Network intrusion detection is critical to cybersecurity, aiming to identify and mitigate unauthorized access and attacks on computer systems and networks. This study evaluates the effectiveness of three machine learning techniques—deep neural networks (DNN), gradient boost trees (GBT), and k-nearest neighbors (KNN)—in detecting network intrusions. The performance of these models was assessed using a comprehensive dataset of 2,540,047 records encompassing 49 features across nine attack categories. The results indicate that GBT outperforms DNN and KNN in accuracy and robustness. These findings highlight the potential of GBT for enhancing intrusion detection systems and contribute valuable insights into the comparative performance of different machine learning algorithms in cybersecurity applications.

Keywords: Network Attacks Forecasting; UNSW-NB15 Dataset; Deep Neural Networks; Gradient Boost Trees; k-Nearest Neighbors

1. Introduction

In the current digital landscape, the security of data, computer systems, and networks is paramount due to the critical nature of the information they contain, ranging from healthcare and financial data to personal records [1]. The increasing reliance on digital storage has escalated the risks associated with data breaches and cyber intrusions, which can lead to significant economic damage and degrade the performance of information systems [1]. Therefore, developing advanced Intrusion Detection Systems (IDS) is essential for preserving data confidentiality and system integrity. This study makes significant contributions by evaluating the effectiveness of three advanced machine learning techniques—deep neural networks (DNN), gradient boost trees (GBT), and k-nearest neighbors (KNN)—in enhancing IDS. By comparing these techniques, the research provides valuable insights into their performance, aiding in developing more robust cybersecurity measures. Machine Learning (ML), with its adaptive learning capabilities from data, emerges as a potent tool for enhancing IDS by identifying and mitigating sophisticated cyber threats [2,3,4,5].

The application of ML in IDS development has attracted considerable attention due to its potential to improve intrusion detection efficacy significantly [6]. The UNSW-NB15 dataset, with its comprehensive coverage of various attack vectors such as Fuzzers, Analysis, Backdoors, DoS attacks, Exploits, and Viruses, provides an invaluable resource for IDS research and development, offering insights into potential vulnerabilities and the effectiveness of different detection strategies [7]. Previous studies have leveraged various ML algorithms to detect

cyber threats. For instance, Random Forest and Neural Networks have been used to identify cyber intrusions [22] accurately. Similarly, Naïve Bayes and Support Vector Machines (SVM) have effectively classified network attacks [23]. Moreover, deep learning models like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) have been employed for real-time malware detection in IoT networks [24]. Integrating k-nearest Neighbors (KNN) and Random Forests has also effectively differentiated between normal and attack traffic in network data [25]. This research harnesses advanced ML techniques, including Deep Neural Networks (DNN), Gradient Boost Trees (GBT), and k-nearest Neighbors (KNN), to engineer an IDS capable of effectively detecting a wide array of cyber threats. These techniques were selected for their demonstrated proficiency in pattern recognition and anomaly detection, which are crucial for timely and accurate threat identification [8-10].

This study aims to assess the performance of these ML-based IDS models in identifying and classifying diverse cyberattacks, utilizing the rich and varied UNSW-NB15 dataset. The findings highlight the capability of ML-powered IDS to efficiently detect and report malicious or abnormal activities, thereby enhancing the security measures of digital systems and networks. Specifically, this research makes significant contributions by providing a detailed comparative analysis of the accuracy and robustness of DNN, GBT, and KNN techniques in network intrusion detection. By highlighting the strengths and weaknesses of each method, the study offers valuable insights that can guide the development of more advanced and effective IDS solutions. Moreover, this research contributes to the body of knowledge by providing a comparative analysis of the accuracy of the employed ML techniques, offering insights that could inform future improvements and the development of more effective cybersecurity measures [8-10].

2. Materials and Methods

Applying Machine Learning (ML) techniques for forecasting network attacks within time series data is a focal point of this study, utilizing the comprehensive UNSW-NB15 dataset, encompassing over 2.54 million records [7].

2.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) represents a fundamental phase in data analytics, facilitating an in-depth understanding and extracting insights from datasets. This process involves a series of steps aimed at preparing the data for further analysis, including data selection and preparation for analytical processing, univariate analysis to understand individual variables, bi-/multivariate analysis to explore relationships between two or more variables, detection of aberrant and missing values to ensure data quality, outlier detection to identify and assess anomalies, and feature engineering to develop variables that enhance analytical depth and predictive capability [11]. This study utilized EDA to identify key features and patterns within the UNSW-NB15 dataset. The EDA process included the generation of statistical summaries and visualizations better to understand the distribution and relationships of the features. The results from EDA informed subsequent steps in feature selection and model development, ensuring that the most relevant data attributes were used in the machine learning models.

2.2 Correlation Heatmap

The Correlation Heatmap is a visual tool to depict the relationships between pairs of variables within the dataset, utilizing a color spectrum—often in cool, warm tones—to signify the strength and direction of correlations. Darker shades represent strong positive correlations, where an increase in one variable corresponds to an increase in another, while lighter shades denote strong negative correlations, indicating inverse relationships between variables. Intermediate shades suggest negligible or weak correlations, highlighting the absence of significant relationships. The heatmap is quantitatively underpinned by correlation coefficients ranging from -1 to 1, where 1 signifies a perfect positive correlation, -1 is a perfect negative correlation, and 0 denotes the absence of correlation [12]. This study generated the correlation heatmap to identify and visualize the relationships between different features within the UNSW-NB15 dataset. This analysis was crucial for feature selection, as it helped identify highly correlated features that could be redundant. Removing or combining these features optimizes the data for better performance in machine learning models. The insights from the correlation heatmap directly informed the feature engineering and selection process, ensuring that the models were trained on the most relevant and non-redundant features.

2.3 Machine Learning (ML)

Machine Learning, a pivotal subset of artificial intelligence, aims to develop computer systems capable of learning and improving autonomously from data and experience. By employing mathematical techniques and data analysis, ML crafts models capable of predictions, data classification, and other tasks using available datasets [13]. This study employed machine learning techniques to develop models for detecting network intrusions. The selected techniques—Deep Neural Networks (DNN), Gradient Boost Trees (GBT), and k-Nearest Neighbors (KNN)—were chosen for their demonstrated proficiency in handling large-scale data and their effectiveness in pattern recognition and anomaly detection. The performance of these models was systematically evaluated to determine their accuracy and robustness in identifying and classifying cyber threats within the UNSW-NB15 dataset. This comprehensive approach ensures that the most effective ML techniques are utilized to enhance the capabilities of intrusion detection systems.

2.4 Intrusion Detection Systems (IDS)

Intrusion Detection Systems (IDS) are integral to computer security designed to safeguard computer systems and networks from unauthorized access, data breaches, and potentially malicious activities. IDS monitors network traffic and system behavior to identify anomalies and security risks, thereby ensuring the integrity and confidentiality of information systems [6]. This study enhanced IDS effectiveness using machine learning techniques to improve detection rates and reduce false positives. By employing advanced algorithms such as DNN, GBT, and KNN, the IDS developed in this research aims to provide robust and accurate detection of a wide range of cyber threats. Integrating these ML techniques enables the IDS to learn from historical data and adapt to new threats, significantly enhancing its capability to protect network infrastructure.

2.5 Deep Learning and Deep Neural Networks (DNN)

Deep Learning, an advanced branch of ML, leverages Deep Neural Networks (DNNs) with multiple layers to extract complex features from data. DNNs have shown exceptional performance in tasks such as image recognition and natural language processing, automatically learning intricate, non-linear features from data for complex tasks [8]. The Multi-Layer Perceptron (MLP), or fully connected network, represents the simplest form of DNN, consisting of an input layer that matches the number of features, at least one hidden layer, and an output layer. Each neuron in a layer is fully connected to all neurons in the subsequent layer, facilitating comprehensive data analysis. MLP is suitable for tabular datasets, binary classification tasks, and basic regression problems, with an architecture that includes:

- An input layer with neurons corresponding to the number of features.
- Hidden layers, for instance, one with 512 neurons and another with 256 neurons, both employing ReLU activation functions.
- A Dropout layer with a rate of 0.5 to prevent model overfitting.
- An output layer with neurons equal to the number of classes using SoftMax (for classification) or a single neuron without activation (for regression).

Convolutional Neural Networks (CNNs) excel at processing grid-like topology data, such as images, through convolutional layers that automatically learn and adapt the spatial hierarchy of features from input images. This capability is pivotal for image classification, object detection, video analysis, and any task requiring spatial hierarchy in data. The typical CNN architecture includes:

- An input layer sized to match the image dimensions and channels (e.g., 224x224x3 for color images).
- Convolutional layers with small receptive fields (e.g., 3x3 filters) followed by ReLU activation.
- Max Pooling layers to reduce spatial dimensions.
- Additional convolutional layers with more filters than preceding layers, followed by ReLU.
- A fully connected layer to flatten and connect the output of previous layers to a dense layer.
- An output layer with neurons equal to the number of classes featuring a softmax activation function for classification tasks.

Recurrent Neural Networks (RNNs) and their variants like Long Short-Term Memory (LSTM) units and Gated Recurrent Units (GRU), are designed for sequential data processing, where each step is related to the previous one. Advanced RNNs like LSTM and GRU can learn long-term dependencies, making them

suitable for Natural Language Processing (NLP) applications such as language modeling, machine translation, speech recognition, and time series prediction. The architecture typically includes:

- An input layer with word embeddings or encoded vectors for each word/token in the sequence.
- LSTM/GRU layers capable of capturing long-term dependencies.
- A Dropout layer to mitigate overfitting, particularly in recurrent layers.
- A fully connected layer to interpret features identified by recurrent layers.
- An output layer with neurons equal to the number of classes for classification tasks, using SoftMax, or a single neuron for regression tasks.

In this study, the DNN architecture employed for intrusion detection consisted of an input layer, two hidden layers with 128 and 64 neurons, respectively, and an output layer for classification. Each hidden layer used the ReLU activation function to introduce non-linearity, and a dropout rate of 0.5 was applied to prevent overfitting. The network was trained using the Adam optimizer with a fine-tuned learning rate for optimal performance. This setup was chosen based on its proven effectiveness in handling high-dimensional data and its ability to learn complex patterns relevant to identifying network intrusions.

Each architecture can be further customized with additional layers, varying neuron counts, different activation functions, and other hyperparameters to suit the specific characteristics of the data and the problem at hand.

2.6 Gradient Boosting Trees (GBT)

Gradient Boosting is a learning technique that enhances prediction accuracy by integrating the predictions from multiple decision tree models. It continuously improves model predictions by focusing on the misclassified sample margins from previous iterations, known for its robustness and ability to handle various data efficiently [9]. This study employed GBT to develop a robust intrusion detection model. The GBT model was trained on the UNSW-NB15 dataset, utilizing its gradient-boosting algorithm to refine the model's accuracy iteratively. Hyperparameters such as the number of trees, learning rate, and maximum depth of the trees were fine-tuned to achieve optimal performance. The GBT's ability to handle high-dimensional data and focus on difficult-to-classify instances made it an excellent choice for enhancing the detection capabilities of the IDS. The model's performance was evaluated based on its accuracy, precision, recall, and F1 score, demonstrating its effectiveness in identifying and classifying network intrusions.

2.7 k-Nearest Neighbors (KNN)

The k-Nearest Neighbors algorithm is a classification method that assigns class labels based on the proximity of a given data point to others in the dataset. The principle is to compare a data point of interest with others to determine its similarity; the system classifies it based on the closest data points [10]. This study employed KNN to develop an intrusion detection model using the UNSW-NB15 dataset. The model was trained to maximize classification accuracy by identifying the optimal number of neighbors (k) and the distance metric (e.g., Euclidean distance). The KNN algorithm's ability to handle noisy data and its straightforward implementation made it suitable for this study. The model's performance was evaluated using accuracy, precision, recall, and F1-score metrics, demonstrating its efficacy in classifying network intrusions. Despite its simplicity, KNN provided competitive results, highlighting its potential as a reliable method for intrusion detection.

2.8 Performance Metrics for Classification Models (Confusion Matrix)

The Confusion Matrix is a tool used to assess the performance of classification models by comparing actual versus predicted data classifications, offering insights into the effectiveness of data categorization across different classes. Key metrics derived from the Confusion Matrix include:

- Accuracy: Measures the overall correctness of the model across all classes.
- Precision (P): Assesses the model's exactness, which is considered separately for each class.
- Recall (R): Evaluates the model's correctness, considering each class separately.
- F-measure (F): Provides a harmonic mean of precision and recall for the model, which is again considered separately for each class.

The Confusion Matrix, a table used to evaluate the efficacy of various ML classification models in conjunction with test datasets or real outcomes, facilitates this assessment by comparing actual values with predicted outcomes, summarizing the results in a matrix format that includes True Negatives (TN), False Positives (FP), False Negatives (FN), and True Positives (TP) as shown in Figure 1 [20,21]. This study used the Confusion Matrix to evaluate the performance of the DNN, GBT, and KNN models on the UNSW-NB15 dataset. The models' capabilities to accurately classify network intrusions were assessed by analyzing the Confusion Matrix. This analysis helped understand the models' strengths and weaknesses in distinguishing between normal and attack traffic, thereby comprehensively evaluating their performance. The evaluation metrics such as accuracy, precision, recall, and F-measure were calculated from the Confusion Matrix to determine the overall effectiveness of each model.

Estimated amount / Actual amount	yes	no	Total
yes	TP	FN	P
no	FP	TN	N
Total	P'	N'	$P+N$

Figure 1. This is a figure. Schemes follow the same formatting.

2.9 Related Research

In response to the pressing need for advanced predictive models in cybersecurity, this study proposes a novel methodological framework to enhance the forecasting accuracy of network attacks. Central to our approach is the application of three sophisticated Machine Learning (ML) techniques: Deep Neural Networks (DNN), Gradient Boost Trees (GBT), and k-Nearest Neighbors (kNN), each selected for their unique strengths and proven track record in the field of data analytics and cybersecurity.

Previous studies have demonstrated the effectiveness of various ML techniques for cyber threat detection:

Random Forests and Neural Networks have accurately detected cyber threats within network traffic data [22].

Naïve Bayes and Support Vector Machines (SVM) have proven effective in network attack classification, highlighting their utility in identifying cyber intrusions [23].

Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models have been successfully used for real-time malware detection in IoT networks [24].

k-Nearest Neighbors (kNN) and Random Forest algorithms have effectively distinguished between normal and attack traffic, showcasing their versatility in cybersecurity applications [25].

Recursive Feature Elimination (RFE) with Random Forest in edge computing environments has enhanced intrusion detection capabilities by eliminating redundant features [26].

This study builds upon these foundations by employing DNN, GBT, and kNN techniques in a comprehensive comparative analysis using the UNSW-NB15 dataset.

Deep Neural Networks (DNN) are at the forefront of our methodology, chosen for their capacity to model complex and non-linear relationships inherent in large-scale data. The depth and flexibility of DNN architectures make them particularly suited for capturing the multifaceted patterns of network attacks within the expansive UNSW-NB15 dataset, which encompasses over 2.54 million records.

Gradient Boost Trees (GBT) are incorporated to leverage their robust predictive performance and ability to handle diverse data types efficiently. The GBT algorithm, known for its precision and adaptability, is utilized to iteratively refine the model's accuracy, focusing on improving predictions for instances that are difficult to classify, thereby enhancing the overall model resilience against overfitting. k-Nearest Neighbors (kNN), with its intuitive classification mechanism based on the similarity of data points,

forms the third pillar of our methodology. This algorithm's simplicity and effectiveness in identifying patterns within data make it an invaluable tool for assessing the likelihood of network attacks, providing a complementary perspective to the more complex DNN and GBT models. This integrated methodological framework, encompassing DNN, GBT, and kNN algorithms, is designed to forecast network attacks with high accuracy and delve into the underlying dynamics and characteristics of such attacks within the time-series context of the UNSW-NB15 dataset. By harnessing the collective strengths of these ML techniques, our study aims to contribute significant insights into the detection and prediction of network attacks, paving the way for developing more secure and resilient network infrastructures.

3. Research Methodology

This study harnesses advanced Machine Learning (ML) methodologies, recognized for their autonomous learning and continuous performance optimization capabilities, to devise a sophisticated Intrusion Detection System (IDS). This system is designed to meticulously analyze and detect data breaches and intrusions within computer networks, addressing a pivotal challenge in cybersecurity that necessitates effective and reliable preventative measures. The research methodology encompasses the integration of three cutting-edge algorithms: Deep Neural Networks (DNN), Gradient Boost Trees (GBT), and k-Nearest Neighbors (KNN). The primary aim is to derive a model that exhibits maximal accuracy by leveraging each technique's unique strengths and predictive capabilities, thereby ensuring a robust and comprehensive approach to intrusion detection. The efficacy of the proposed IDS model is systematically visualized in Figure 2, which outlines the research framework spanning from data preparation to performance evaluation. This schematic representation delineates the sequential approach, beginning with partitioning the UNSW-NB15 dataset, progressing through the application of sophisticated ML algorithms, and culminating in the rigorous assessment of the model's classification accuracy and precision using a confusion matrix. The research framework also includes cross-validation to ensure the robustness of the model's performance across different subsets of data.

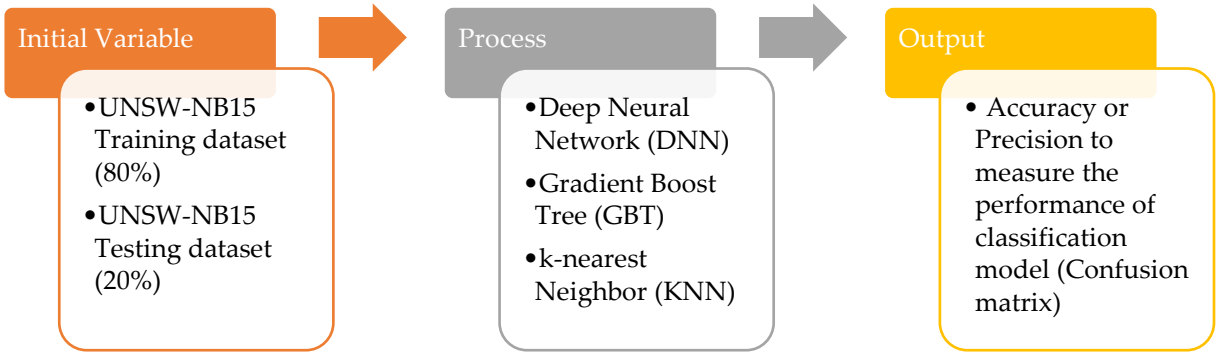


Figure 2. Research Framework

3.1 Data Set (Data Collection Method)

The UNSW-NB15 dataset, developed for Intrusion Detection System (IDS) research, aims to test and research the detection and prevention of computer and network intrusions, including the specifics of each attack type. Created by the University of New South Wales (UNSW) in Australia, the dataset provides a dichotomy of normal communication (labeled as '0') and attack data (labeled as '1'), with a total of 2,540,047 records featuring 49 key attributes relevant to network communications and attacks, such as IP addresses, port numbers, protocols, duration, byte counts, status, etc., across nine different types of attacks: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms [7]. For this study, the dataset was split into 80% training and 20% testing data to ensure a balanced representation and mitigate the impact of class imbalances. The training data was used to build and validate the models, while the testing data was used to evaluate their performance. Cross-validation techniques were employed further to ensure the robustness of the models' performance.

3.2 Data Preparation

This essential step in the data analysis ensures the data is primed for analysis and presentation. The process includes the following sub-steps:

- Selection of the relevant data for analysis from CSV file sources.
- Reading the data from the files into the analysis system using suitable tools or libraries such as Pandas in Python.
- Conducting a preliminary check to verify data integrity and correctness, including the completeness of rows and columns and identifying missing or abnormal values.
- Transforming the data into a format conducive to analysis, for instance, converting nominal data columns from the UNSW-NB15 dataset into numerical values and addressing incomplete data.
- Formatting the data into an analysis-friendly structure includes renaming columns, organizing the data sequence, and setting a format that enhances data representation.
- Managing missing data by excising it or applying imputation techniques to fill in gaps.
- Normalizing the data to ensure consistency across scales, such as converting values to a range of 0 to 1.
- Enriching the dataset with additional data creation or compilation from other sources if required.
- Saving the prepared data for analysis, where the formatted and cleansed data can be documented in a file format suitable for further analytical processes.

The UNSW-NB15 dataset is a comprehensive resource for testing network intrusion detection systems. Its primary features encompass network connections, attack mechanisms, network traffic, defense systems, endpoint systems, and connection timing, as shown in Table 1. This dataset is a crucial tool for testing and evaluating network intrusion detection models and simulating attack and communication patterns within network systems. It provides a realistic environment to assess detection strategies' effectiveness and train algorithms to recognize and respond to a wide range of cybersecurity threats.

In this study, the data preparation process included additional steps to enhance the quality and utility of the dataset: Data Cleansing: Techniques such as outlier detection and removal were employed to ensure the dataset's accuracy and reliability. Data cleansing also involved handling anomalies and inconsistencies within the dataset. Feature Selection: Key features were selected to improve model performance using techniques such as Recursive Feature Elimination (RFE) and correlation analysis. This step was critical in reducing the dimensionality of the dataset and eliminating redundant features. Data Splitting: The dataset was split into training and testing sets (80/20 split) to validate the models effectively and to ensure balanced representation. Cross-validation techniques were employed to enhance model robustness further and mitigate overfitting. Data Augmentation: Synthetic data was generated to address class imbalance, providing the training process was robust and comprehensive. This involved using techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to create a balanced training dataset.

Table 1. Features of the UNSW-NB15 Dataset.

No.	Name	Type	Description
1	srcip	nominal	Source IP address
2	sport	integer	Source port number
3	dstip	nominal	Destination IP address
4	dsport	integer	Destination port number
5	proto	nominal	Transaction protocol
6	state	nominal	Indicates to the state and its dependent protocol, e.g. ACC, CLO, CON, ECO, ECR, FIN, INT, MAS, PAR, REQ, RST, TST, TXD, URH, URN, and (-)
7	dur	Float	Record total duration
8	sbytes	Integer	Source to destination transaction bytes
9	dbytes	Integer	Destination to source transaction bytes
10	sttl	Integer	Source to destination time to live value

Table 1. Features of the UNSW-NB15 Dataset. (Continue)

No.	Name	Type	Description
11	dttl	Integer	Destination to source time to live value
12	sloss	Integer	Source packets retransmitted or dropped
13	dloss	Integer	Destination packets retransmitted or dropped
14	service	nominal	http, ftp, smtp, ssh, dns, ftp-data, irc and (-) if not much used service
15	Sload	Float	Source bits per second
16	Dload	Float	Destination bits per second
17	Spkts	integer	Source to destination packet count
18	Dpkts	integer	Destination to source packet count
19	swin	integer	Source TCP window advertisement value
20	dwin	integer	Destination TCP window advertisement value
21	stcpb	integer	Source TCP base sequence number
22	dtcpb	integer	Destination TCP base sequence number
23	smeansz	integer	Mean of the ?ow packet size transmitted by the src
24	dmeansz	integer	Mean of the ?ow packet size transmitted by the dst
25	trans_depth	integer	Represents the pipelined depth into the connection of http request/response transaction
26	res_bdy_len	integer	Actual uncompressed content size of the data transferred from the server's http service.
27	Sjit	Float	Source jitter (mSec)
28	Djit	Float	Destination jitter (mSec)
29	Stime	Timestamp	record start time
30	Ltime	Timestamp	record last time
31	Sintpkt	Float	Source interpacket arrival time (mSec)
32	Dintpkt	Float	Destination interpacket arrival time (mSec)
33	tcprtt	Float	TCP connection setup round-trip time, the sum of 'synack' and 'ackdat'.
34	synack	Float	TCP connection setup time, the time between the SYN and the SYN_ACK packets.
35	ackdat	Float	TCP connection setup time, the time between the SYN_ACK and the ACK packets.
36	is_sm_ips_ports	Binary	If source (1) and destination (3) IP addresses equal and port numbers (2)(4) equal then, this variable takes value 1 else 0
37	ct_state_ttl	Integer	No. for each state (6) according to the specific range of values for source/destination time to live (10) (11).
38	ct_flw_http_mthd	Integer	No. of flows with methods such as Get and Post in http service.
39	is_ftp_login	Binary	If the ftp session is accessed by the user and password, then 1 else 0.
40	ct_ftp_cmd	integer	No flows that have a command in ftp session.
41	ct_srv_src	integer	No. of connections that contain the same service (14) and source address (1) in 100 connections according to the last time (26).
42	ct_srv_dst	integer	No. of connections that contain the same service (14) and destination address (3) in 100 connections according to the last time (26).
43	ct_dst_ltm	integer	No. of connections of the same destination address (3) in 100 connections according to the last time (26).
44	ct_src_ ltm	integer	No. of connections of the same source address (1) in 100 connections according to the last time (26).

Table 1. Features of the UNSW-NB15 Dataset. (Continue)

No.	Name	Type	Description
45	ct_src_dport_ltm	integer	No connections of the same source address (1) and the destination port (4) in 100 connections according to the last time (26).
46	ct_dst_sport_ltm	integer	No connections of the same destination address (3) and the source port (2) in 100 connections according to the last time (26).
47	ct_dst_src_ltm	integer	No connections of the same source (1) and the destination (3) address in 100 connections according to the last time (26).
48	attack_cat	nominal	The name of each attack category. In this data set, nine categories, e.g., Fuzzers, Analysis, Backdoors, DoS Exploits, Generic, Reconnaissance, Shellcode, and Worms
49	Label	binary	0 for normal and 1 for attack records

3.3 Model Development

In the progression of data analysis and machine learning model development, this research utilized the Python programming language, capitalizing on its robust analytical libraries, including Pandas [33], NumPy [34], Seaborn [35], and Matplotlib [36] for effective data manipulation and visualization. These tools were integral to the experimental comparison of various models to ascertain the most productive algorithm for our specific application. The Scikit-learn library was also employed to implement and optimize the machine-learning algorithms.

The study focused on optimizing model parameters to maximize classification accuracy to accommodate the intricacies of the UNSW-NB15 dataset, which is comprised of diverse network intrusion events. The chosen modeling techniques encompassed Deep Neural Networks (DNN), Gradient Boost Trees (GBT), and k-Nearest Neighbors (KNN). The specific architectures and parameters of these models were fine-tuned to suit the dataset characteristics:

- Deep Neural Networks (DNN): The DNN model consisted of an input layer, two hidden layers with 128 and 64 neurons, respectively, and an output layer for classification. Each hidden layer used the ReLU activation function, and a dropout rate 0.5 was applied to prevent overfitting. The model was trained using the Adam optimizer with a learning rate fine-tuned for optimal performance.
- Gradient Boost Trees (GBT): The GBT model was trained with hyperparameters such as the number of trees, learning rate, and maximum depth, which were fine-tuned using grid search to achieve the best performance.
- k-Nearest Neighbors (KNN): The KNN model was developed by identifying the optimal number of neighbors (k) and the distance metric (e.g., Euclidean distance) through cross-validation to maximize classification accuracy.

For training and validation, the dataset was partitioned into an 80% training set and a 20% testing set using the train_test_split() function from Pandas, as shown in Figure 3 [37]. Cross-validation techniques were employed further to ensure the robustness of the models' performance. This bifurcation was critical to ensure both the robust training of the models and their subsequent evaluation against an independent test set, thus enabling a thorough assessment of the predictive prowess of the developed models.

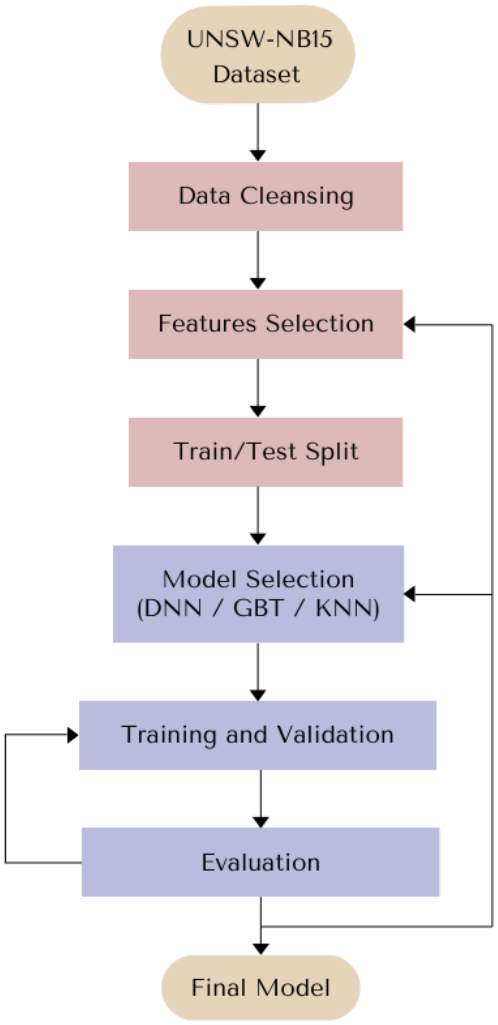


Figure 3. Machine Learning Model Development Flowchart

3.4 Model Evaluation

Following the segmentation of the dataset into training and testing sets, the models trained on 80% of the data were applied to perform network intrusion analysis. The remaining 20% served to evaluate the predictive accuracy of the models. Cross-validation techniques were employed to ensure that the models' performance was robust and generalizable across different subsets of data.

This study employed Deep Neural Networks (DNN), Gradient Boost Trees (GBT), and k-Nearest Neighbors (KNN) to ascertain the efficacy of each model. The evaluation was conducted using several key performance metrics:

- Accuracy: Measures the overall correctness of the model across all classes.
- Precision: Assesses the model's exactness by calculating the ratio of correctly predicted positive observations to the total predicted positives.
- Recall: Evaluate the model's ability to identify all relevant instances by calculating the ratio of correctly predicted positive observations to all observations in the actual class.
- F1-Score: Provides a harmonic mean of Precision and Recall, offering a single metric to evaluate the model's performance.
- Confusion Matrix: Provides a comprehensive view of the model's classification performance by displaying the true positives, false positives, true negatives, and false negatives.

The performance metrics for the models were calculated according to the formulas specified in Table 2 encompasses a range of evaluative criteria. These metrics were derived from the confusion matrix to provide a detailed understanding of each model's strengths and weaknesses in classifying network intrusions.

The evaluation revealed that the GBT model outperformed the DNN and KNN models in accuracy and robustness. The significance of GBT's superior performance was confirmed through statistical tests and comparative analysis. Despite the simplicity of the KNN model, it provided competitive results, highlighting its potential as a reliable method for intrusion detection.

Table 2. Performance Metrics Used in the Evaluation Process.

No.	Metric	Equation
1	Accuracy and recognition rate	$\frac{\text{True Positive} + \text{True Negative}}{\text{Total Observations}}$
2	Error rate and misclassification rate	$\frac{\text{False Positive} + \text{False Negative}}{\text{Total Observations}}$
3	Sensitivity, True positive rate, and recall	$\frac{\text{True Positive}}{\text{Positive Observations}}$
4	Specificity, True negative rate	$\frac{\text{True Negative}}{\text{Negative Observations}}$
5	Precision	$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$
6	Recall	$\frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$
7	F1 score	$\frac{2 \text{ x Precision x Recall}}{\text{Precision} + \text{Recall}}$

3.5 Model Deployment

The developed models utilizing Deep Neural Networks (DNN), Gradient Boost Trees (GBT), and k-Nearest Neighbors (KNN) are slated for operational deployment, aimed at analyzing and detecting intrusions within computer and network systems. This stage will see the transition from theoretical models to practical, deployable software solutions that can be integrated into existing cybersecurity infrastructures for enhanced real-time analysis and threat detection. The deployment process involves several critical steps to ensure the models' effective integration and functionality in real-world environments:

- Integration with Existing Systems: The models will be integrated into current IDS frameworks, ensuring compatibility and seamless operation alongside existing security measures.
- Real-time Data Processing: The deployed models will process network traffic in real time, continuously monitoring suspicious activities and potential threats. This requires robust data handling and low-latency processing capabilities.
- Model Update Mechanisms: To maintain high detection accuracy, mechanisms for updating the models with new data and retraining them periodically will be implemented. This ensures the models remain effective against evolving cyber threats.
- Scalability and Performance Optimization: The deployment will consider scalability to efficiently handle large volumes of network traffic. Performance optimization techniques will be employed to minimize resource usage while maximizing detection speed and accuracy.

- Security and Privacy Considerations: Ensuring the security and privacy of the data being processed by the models is paramount. Measures will be taken to protect sensitive information and comply with relevant data protection regulations.

By following these deployment steps, the models will be effectively transitioned from experimental setups to operational IDS solutions capable of providing robust, real-time network protection. The deployment will be continuously monitored and adjusted based on feedback and performance metrics to ensure optimal operation.

4. Results of Data Analysis

In this study, a comprehensive dataset of 100 GB, containing 49 distinct features, was analyzed. Following the data preparation phase, the dataset was consolidated into 2,540,047 records across four integrated data files, as shown in Table 3. The amalgamation of various Data Frames, supplemented with descriptive annotations for each feature, facilitated a clearer interpretation of the dataset, ensuring each column's contents were accurately represented.

Table 3. This table shows the ratio of the data in each data file.

Data File	Number of observations (records)	Number of Features	Ratio
DF1	700,001	49	27.56%
DF2	700,001	49	27.56%
DF3	700,001	49	27.56%
DF4	440,044	49	17.32%

Histograms were generated to illustrate the distribution of labels within the dataset, thereby aiding in examining label frequencies pertinent to the Intrusion Detection System (IDS) dataset. The dataset predominantly consisted of 'normal' records (label '0'), with a count of 2,218,764, in contrast to 'attack' instances (label '1'), which amounted to 321,283 records, as shown in Figure 4.

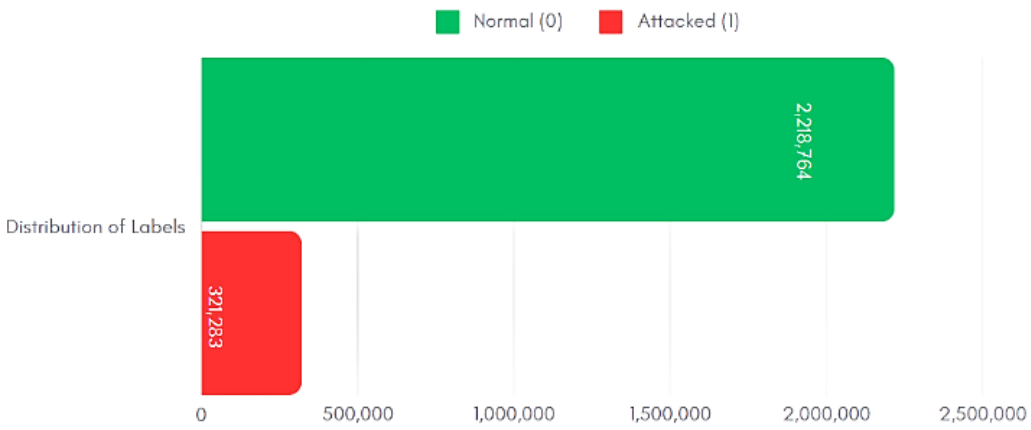


Figure 4. Distribution of Labels of the Intrusion Detection System (IDS) dataset

After the initial analysis, columns categorized as 'Object' types were excluded due to their non-numeric nature. This refinement resulted in a dataset composed solely of numerical columns, enhancing its suitability for analytical and computational processing. This step was pivotal in ensuring the dataset's compatibility with the requirements of intrusion detection systems and other analytical endeavors, focusing on data size, type, integrity, and validation. The construction of a Correlation Heatmap, utilizing the Seaborn (sns) and Matplotlib (plt) libraries, enabled the calculation of correlations between features within the Data Frame. This heatmap graphically depicted the correlation matrix for each feature pair, providing a visual representation of the relationships between variables. The heatmap analysis indicated darker cells, with values nearing +1, signified a strong positive correlation, suggesting a direct relationship between feature pairs. In

contrast, lighter cells, with values approaching -1, denoted a strong negative correlation, indicating an inverse relationship between certain features. This analysis aspect was instrumental in elucidating the negative correlations between specific feature pairs within the dataset.

The provided text does have an academic orientation but can be enhanced for precision and formality. Here's a refined version maintaining the original structure:

4.1 Evaluation of Deep Neural Network (DNN) Classification

The evaluation of the Deep Neural Network (DNN) model revealed an accuracy rate of 87.3%. The model architecture incorporated two hidden layers, with the first and second layers comprising 128 and 64 nodes, respectively. Each node within these layers employed a Rectified Linear Unit (ReLU) activation function to facilitate data differentiation and augment model complexity. The training process, conducted over 50 epochs, utilized the 'adam' optimizer alongside a specified loss function. Techniques for regularization and a dropout layer, implemented via TensorFlow, were critical in minimizing overfitting by adjusting network weights dynamically. Despite these measures, the model's precision, recall, and F1-score indicated certain limitations in accurately detecting specific classes. To enhance model performance, further scrutiny of the training dataset is advisable, in addition to expanding the training data volume and potentially adjusting the number of layers and nodes to achieve higher accuracy.

The confusion matrix for the DNN model, as shown in Figure 5 (right panel), illustrates the model's performance in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Specifically, the DNN model achieved 443,790 true positives, 64,220 false positives, 0 true negatives, and 0 false negatives. This analysis provided insights into the areas where the model excelled and struggled.

- True Positives (TP): The model correctly identified 443,790 instances of attacks.
- False Positives (FP): The model incorrectly identified 64,220 normal instances as attacks.
- True Negatives (TN): The model did not identify any normal instances correctly.
- False Negatives (FN): The model did not miss any attack instances.

This confusion matrix analysis reveals a significant imbalance in the model's ability to identify normal instances (TN) correctly. The model's high number of false positives indicates a tendency to overpredict attacks, which could lead to numerous false alarms in a practical deployment. Conversely, the model's perfect detection of attack instances (no false negatives) suggests strong performance in identifying threats.

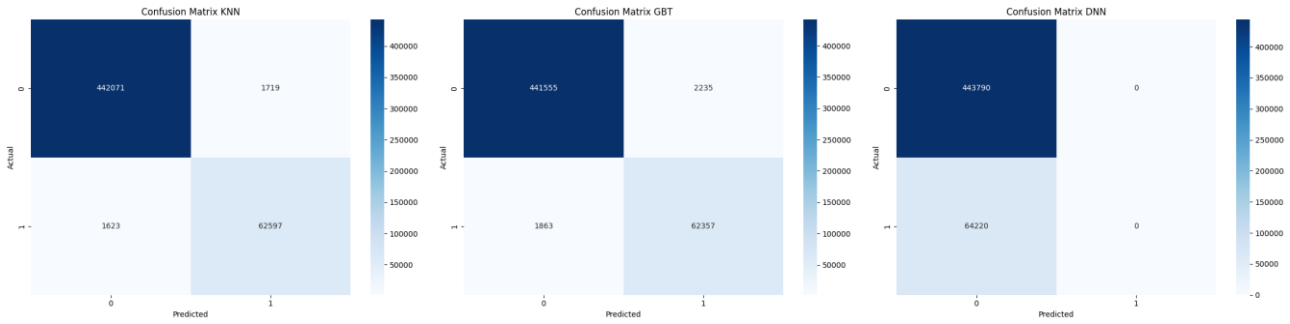


Figure 5. Confusion Matrices for DNN, GBT, and KNN Models

4.2 Gradient Boost Tree (GBT) Classification Performance

The Gradient Boost Tree (GBT) model exhibited superior performance with an accuracy of 99.1%, a precision of 96.5%, a recall of 97%, and an F-measure of 98.2%. This model was constructed and assessed using the Gradient Boosting Classifier from Scikit-Learn, which was tailored for classification tasks. The SimpleImputer function was utilized to manage missing values within the dataset, employing a mean imputation strategy to substitute missing entries with average values. This methodology significantly contributed to the model's high precision and efficacy in classifying diverse classes within the dataset.

The GBT model's architecture involved multiple decision trees, where each tree was built sequentially to correct the errors of the previous trees. The hyperparameters, including the number of trees, learning rate, and maximum depth, were fine-tuned using grid search to optimize model performance. The final model included 100 trees, a learning rate of 0.1, and a maximum depth of 3, balancing complexity and performance.

The confusion matrix for the GBT model, as shown in Figure 5 (middle panel), illustrates the model's performance in terms of true positives, false positives, true negatives, and false negatives. Specifically, the GBT model achieved 441,555 true positives, 2,235 false negatives, 1,863 false positives, and 62,357 true negatives. This analysis provided insights into the areas where the model excelled and struggled.

- True Positives (TP): The model correctly identified 441,555 instances of attacks.
- False Positives (FP): The model incorrectly identified 1,863 normal instances as attacks.
- True Negatives (TN): The model correctly identified 62,357 normal instances.
- False Negatives (FN): The model missed 2,235 instances of attacks.

Furthermore, statistical tests were conducted to confirm the significance of the GBT model's superior performance over the DNN and KNN models. The GBT model's robustness and high accuracy make it a reliable choice for intrusion detection in network systems.

4.3 k-Nearest Neighbor (KNN) Classification Outcomes

The k-Nearest Neighbor (KNN) model demonstrated a notable accuracy of 99%, precision at 96.2%, recall at 96.1%, and an F-measure of 97.8%. The results underscore the model's proficiency in accurately classifying and detecting dataset classes, highlighting the effectiveness of the KNN approach in classification tasks pertinent to network intrusion detection scenarios.

The KNN model was developed by identifying the optimal number of neighbors (k) and the distance metric (e.g., Euclidean distance) through cross-validation to maximize classification accuracy. Despite the simplicity of the KNN model, it provided competitive results, highlighting its potential as a reliable method for intrusion detection.

The confusion matrix for the KNN model, as shown in Figure 5 (left panel), illustrates the model's performance regarding true positives, false positives, true negatives, and false negatives. Specifically, the KNN model achieved 442,071 true positives, 1,719 false negatives, 1,623 false positives, and 62,597 true negatives. This analysis provided insights into the areas where the model excelled and struggled.

- True Positives (TP): The model correctly identified 442,071 instances of attacks.
- False Positives (FP): The model incorrectly identified 1,623 normal instances as attacks.
- True Negatives (TN): The model correctly identified 62,597 normal instances.
- False Negatives (FN): The model missed 1,719 instances of attacks.

Table 4 compares performance metrics comprehensively, including accuracy, precision, recall, and F-measure, across the DNN, GBT, and KNN algorithms. This comparative analysis facilitates a nuanced understanding of the efficacy of each algorithm within the context of the study's objectives.

Table 4. Comparison of performance metrics across DNN, GBT, and KNN algorithms

Algorithm	Accuracy	Precision	Recall	F1-Score
Deep Neural Network (DNN)	87.3%	84.6%	84.5%	85.2%
Gradient Boost Tree (GBT)	99.1%	96.5%	97.0%	98.2%
K-nearest Neighbor (KNN)	99.0%	96.2%	96.1%	97.8%

5. Discussion

This study conducted a comparative evaluation of algorithmic performances, focusing on Deep Neural Networks (DNN), Gradient Boost Trees (GBT), and k-Nearest Neighbors (KNN) within the domain of machine learning. The methodology entailed a comprehensive compilation of training and testing data from empirical sources and literature reviews, followed by the experimental design executed in Python, utilizing specific libraries tailored to each algorithmic model. Upon consolidating the data and refining the feature sets across all three methodologies, a detailed comparative analysis was initiated to ascertain the most effective approach regarding accuracy and precision.

Experimental outcomes revealed that the DNN model exhibited an accuracy of 87.3%. The confusion matrix for the DNN model showed 443,790 true positives, 64,220 false positives, 0 true negatives, and 0 false negatives. This suggests that while the DNN model is highly effective at detecting attacks (high recall), it struggles with a high rate of false positives and does not correctly identify any normal instances. This result aligns with Fuat Türk's research, which applied DNN within network intrusion detection frameworks,

achieving remarkable accuracy levels of 98.6% and 98.3% for binary and multi-class detection, respectively [27]. Aleesa, Ahmed, and Thanoun's findings further corroborate the superior efficacy of DNN in binary classification tasks over multi-class scenarios, with test set accuracies reaching 99.22% for binary classifications, surpassing the 99.59% accuracy rate for multi-class classifications and demonstrating a binary classification loss of 1.56%, as opposed to a 0.92% loss in multi-class categorizations [28].

The GBT model registered the highest accuracy at 99.1%, with 441,555 true positives, 2,235 false negatives, 1,863 false positives, and 62,357 true negatives. The low false positive and false negative rates indicate a balanced performance, with the model accurately identifying both attack and normal instances. This is consistent with Zhou et al.'s insights, which emphasized the equilibrium between detection efficacy and operational speed, employing a 70%/30% dataset split and securing a GBT outcome of 93.13% [29]. In parallel, research by Faker, Osama & Dogdu, and Erdogan highlighted the synergistic integration of Big Data and deep learning techniques to enhance the performance of intrusion detection systems, attaining a maximum accuracy of 97.92% [30].

The KNN model demonstrated an accuracy rate of 99%, with 442,071 true positives, 1,719 false negatives, 1,623 false positives, and 62,597 true negatives. The results underscore the model's proficiency in accurately classifying and detecting dataset classes, highlighting the effectiveness of the KNN approach in classification tasks pertinent to network intrusion detection scenarios. This finding is mirrored in Kasongo's research, where the accuracy for multi-class classification using KNN improved from 70.09% with 42 features to 72.30% with 19 features optimized via the XGBoost method, underscoring XGBoost's contribution to bolstering the predictive capacity of the algorithm [31]. Rahman's investigation, utilizing standalone models within the Weka Explorer framework with dedicated training and test datasets, yielded approximately 100% accuracy [32]. Aziz's study further illustrated an enhancement in dataset performance using the NB classifier, elevating from 76.04% to 92.57%.

In essence, this comparative analysis elucidates the robust capabilities of advanced machine learning algorithms in refining the accuracy and precision of intrusion detection mechanisms, thereby indicating significant prospects for future optimizations and applications in the cybersecurity domain.

Strategies for Improvement:

- Address Class Imbalance: Adjust the training process to improve the model's ability to correctly identify normal instances, potentially through techniques such as class weighting or oversampling of normal instances.
- Feature Engineering: Enhance the feature set to better discriminate between typical and attack instances.
- Model Architecture: Experiment with more profound or complex network architectures to improve the model's overall performance.

These strategies for improvement aim to address the limitations identified in the current models and enhance their practical applicability in real-world cybersecurity scenarios. By refining the training process, improving the feature set, and experimenting with different model architectures, future research can build on the findings of this study to develop more robust and effective intrusion detection systems.

6. Conclusion and Future works

The findings from this study elucidate the effectiveness of Deep Neural Networks (DNN), Gradient Boost Trees (GBT), and k-Nearest Neighbors (KNN) algorithms in the realm of network intrusion detection and data classification. These results are instrumental in informing the development of future network intrusion detection systems, emphasizing the superior accuracy demonstrated by the GBT and KNN algorithms. Such insights pave the way for advancing detection and prevention mechanisms, potentially leading to the innovation of adaptive models capable of autonomously adjusting to new system behaviors or emerging threats.

Moreover, the applicability of DNN in identifying a diverse array of intrusions, including DDoS attacks, SQL Injections, and Zero-Day exploits, underscores the adaptability and strength of these algorithms within cybersecurity applications. Nonetheless, the issue of imbalanced data presents a significant challenge to the reliability of classification results, as disproportionate class distributions can lead to biased accuracy measures, often manifested in an elevated Null Accuracy. Incorporating additional evaluative metrics, such

as the Area Under the Curve (AUC) derived from Receiver Operating Characteristic (ROC) analysis, offers a more comprehensive assessment of model efficacy, enhancing the strategic development of system security measures.

The study's results indicate that the Gradient Boost Tree (GBT) model demonstrated the highest accuracy at 99.1%, with significant performance across various metrics. The k-Nearest Neighbors (KNN) model also performed exceptionally well, with an accuracy of 99%. In contrast, despite its high recall, the Deep Neural Network (DNN) model showed a significant number of false positives, suggesting areas for improvement in its application for network intrusion detection.

In conclusion, this research has laid a foundational framework for the application of Deep Neural Networks (DNN), Gradient Boost Trees (GBT), and k-Nearest Neighbors (KNN) in the domain of network intrusion detection, offering promising insights into their efficacy and potential areas for enhancement. The exploration of these algorithms, within the context of this study, reveals significant opportunities for advancing the precision, adaptability, and real-time capabilities of intrusion detection systems. Future research directions, as delineated, encompass a broad spectrum of possibilities ranging from algorithmic refinement and imbalanced data management to integrating machine learning models with emerging technologies and developing comprehensive, automated response mechanisms. These avenues hold the promise of elevating the robustness and efficiency of network security frameworks and contribute to the broader discourse on the role of advanced computational techniques in cybersecurity. As we progress, the continuous iteration and expansion of this research domain will undoubtedly play a pivotal role in shaping resilient and adaptive cybersecurity infrastructures capable of confronting the ever-evolving landscape of cyber threats.

Future research should address class imbalance, enhance feature engineering, and experiment with more complex model architectures to improve overall performance. Additionally, exploring the integration of machine learning models with real-time data processing technologies can further enhance the efficacy of intrusion detection systems. By building on the findings of this study, future work can develop more robust, adaptive, and efficient cybersecurity solutions capable of mitigating a wide array of threats.

7. Acknowledgements

We want to express our sincere gratitude to North Bangkok University for their generous financial support and resources throughout this research project. We are also grateful to the faculty, staff, and fellow researchers for their invaluable expertise, encouragement, and constructive feedback, which have significantly contributed to our understanding of cybersecurity and the application of exploratory data analysis techniques. Their support has played a crucial role in completing our studies and advancing knowledge in this field.

Author Contributions: Conceptualization, A.V. and W.C.; methodology, R.V.; software, R.V.; validation, A.V.; formal analysis, A.V.; investigation, A. V. and W.C.; resources, R.V.; data curation, R.V.; writing—original draft preparation, R.V.; writing—review and editing, A.V.; visualization, A.V.; supervision, A.V.; project administration, A.V.; funding acquisition, A.V.

Funding: This research is subsidized by North Bangkok University.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Avital, M.; Bittencourt, L. F.; Santos, J. L.; Marín, D. B.; Rojas, L. C. Global DDoS Threat Landscape Report **2019**.

[2] Axelsson, S. The Base-Rate Fallacy and the Difficulty of Intrusion Detection. *ACM Trans. Inf. Syst. Secur*. **2000**, *3*(3), 186–205.

[3] Tavallaee, M.; Bagheri, E.; Lu, W.; Ghorbani, A. A. A Detailed Analysis of the KDD CUP 99 Data Set. *Proc. 2009 IEEE Symp. Comput. Intell. Secur. Def. Appl.*, **2009**, 1–6.

[4] Creech, G.; Hu, J.; Williams, R. A Survey and Comparison of Distributed Intrusion Detection System. *J. Netw. Comput. Appl.* **2014**, *40*, 12–33.

[5] Nassar, M.; Alazab, M.; Venkatraman, S. Anomaly Intrusion Detection System: A Comprehensive Review. *IEEE Access* **2019**, *7*, 166152–166188.

[6] Garcia, S.; Grill, M.; Stiborek, J.; Zunino, A. An Empirical Comparison of Botnet Detection Methods. *Comput. Secur*. **2014**, *45*, 100–123.
[7] Moustafa, N.; Slay, J. UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set). In Proceedings of the Military Communications and Information Systems Conference (MilCIS), 2015; IEEE: **2015**.
[8] Singh, M.; Kaur, M.; Singh, S. Intrusion Detection System Using Deep Learning Techniques: A Review. **2019**.
[9] Yang, X.; Wu, M.; Wei, W.; Li, L. A Boosted Trees Algorithm for Network Intrusion Detection. **2019**.
[10] Zhang, Z.; Luo, L.; Zhang, L.; Zhang, Y. Network Intrusion Detection System Based on k-Nearest Neighbor Algorithm. **2019**.
[11] Tukey, J. W. Exploratory Data Analysis; **1977**.
[12] Ashok, K.; Kumar, S.; Kumari, A.; Saini, A. Edge Computing Based IDS Detecting Threats Using Machine Learning and PyCaret. 2023, DOI: 10.1109/CISES58720.2023.10183591.
[13] Mitchell, T. M. Machine Learning; McGraw Hill: **1997**.
[14] Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction; Springer Science & Business Media: **2009**.
[15] Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning; MIT Press: **2016**.
[16] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems, **2012**; pp 1097–1105.
[17] LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* 2015, *521*(7553), 436–444.
[18] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems, **2012**; pp 1097–1105.
[19] LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* 2015, *521* (7553), 436–444.
[20] Duangklang, P.; Kruakaew, R. Models for Automatic Aircraft Type Prediction. *NKRAFA J. Sci. Technol*. **2019**.
[21] Confusion Matrix - An Overview. Available online: https://www.sciencedirect.com/topics/engineering/confusion-matrix (accessed on June 8, 2022).
[22] Using Machine Learning Techniques Random Forest and Neural Network to Detect Cyber Attacks. **2023**, DOI: 10.14293/pr2199.000059.v1.
[23] Patrick, S. Machine Learning Based Network Attacks Classification. **2023**, DOI: 10.1109/icpeca56706.2023.10075818.
[24] Jingwen, W.; Peilong, L. MalIoT: Scalable and Real-time Malware Traffic Detection for IoT Networks. **2023**, arXiv:2304.00623.
[25] Michael, G. A Machine-Learning Procedure to Detect Network Attacks. J. Complex Netw. **2023**, DOI: 10.1093/comnet/cnad017.
[26] Ashok, K.; Kumar, S.; Kumari, A.; Saini, A. Edge Computing Based IDS Detecting Threats Using Machine Learning and PyCaret. **2023**, DOI: 10.1109/CISES58720.2023.10183591.
[27] Fuat, T. Analysis of Intrusion Detection Systems in UNSW-NB15 and NSL-KDD Datasets with Machine Learning Algorithms. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi 2023, DOI: 10.17798/bitlisfen.1240469.
[28] Aleesa, A.; Thanoun, M.; Mohammed, A.; Sahar, N. DEEP-INTRUSION DETECTION SYSTEM WITH ENHANCED UNSW-NB15 DATASET BASED ON DEEP LEARNING TECHNIQUES. J. Eng. Sci. Technol. **2021**, *16*, 711–727.
[29] Zhou, Y.; Han, M.; Liu, L.; He, J.S.; Wang, Y. Deep Learning Approach for Cyberattack Detection. In IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS); IEEE: **2018**, pp 262–267.
[30] Faker, O.; Dogdu, E. Intrusion Detection Using Big Data and Deep Learning Techniques. ACM: **2019**, DOI: 10.1145/3299815.3314439.
[31] Kasongo, S. M.; Sun, Y. Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset. *J. Big Data* **2020**, *7*, 105, DOI: 10.1186/s40537-020-00379-6.
[32] Rahman, O.; Quraishi, M. A. G. Experimental Analysis of Random Forest, K-Nearest Neighbor and Support Vector Machine Anomaly Detection. **2019**, DOI: 10.13140/RG.2.2.19998.18245.
[33] McKinney, W. Pandas, Python Data Analysis Library. **2015**.
[34] Oliphant, T. E. Guide to Numpy; Trelgol Publishing: **2006**, 1, 85.

[35] Khandare, A.; Agarwal, N.; Bodhankar, A.; Kulkarni, A.; Mane, I. Analysis of Python Libraries for Artificial Intelligence. In Intelligent Computing and Networking: Proceedings of IC-ICN 2022; Springer Nature Singapore: Singapore, **2023**, pp 157–177.
[36] Ari, N.; Ustazhanov, M. Matplotlib in Python. In 2014 11th International Conference on Electronics, Computer and Computation (ICECCO); IEEE: **2014**, pp 1–6.
[37] Sundaram, J.; Gowri, K.; Devaraju, S.; Gokuldev, S.; Jayaprakash, S.; Anandaram, H.; Thenmozhi, M. An Exploration of Python Libraries in Machine Learning Models for Data Science. In Advanced Interdisciplinary Applications of Machine Learning Python Libraries for Data Science; IGI Global: **2023**, pp 1–31.

ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

Optimizing Organic Fertilization for Marguerite Daisy (*Argyranthemum frutescens*): Impact of Application Rate and Frequency on Growth and Yield

Chamaiporn Anuwong[1], Phissanu Kaewtaphan[2], and Pattrarat Teamkao[3*]

[1] School of Agricultural Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand
[2] School of Agricultural Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand
[3] School of Agricultural Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand
* Correspondence: pattrarat.te@kmitl.ac.th

Citation:
Anuwong, C.; Kaewtaphan, P.; Teamkao, P. Optimizing organic fertilization for marguerite daisy (*Argyranthemum frutescens):* Impact of application rate and frequency on growth and yield. *ASEAN J. Sci. Tech. Report.* **2024**, *27*(5), e253200. https://doi.org/10.55164/ajstr.v27i5.253200

Article history:
Received: March 14, 2024
Revised: July 26, 2024
Accepted: August 8, 2024
Available online: September 7, 2024

Abstract: Edible flowers are a new market for horticulture plants. Beyond their attractive shapes and colors, this group of plants contains secondary metabolites that benefit human health. This study aims to investigate the growth and flower yield of marguerite daisies using organic fertilizer at different rates and frequencies. The experiment was designed as a factorial, completely randomized design with two factors: fertilizer rate (0.5, 1.0, and 1.5 times compared to the total nitrogen content in chemical fertilizer) and frequency of organic fertilizer application (every 30 and 15 days). Slow-release chemical fertilizer (Osmocote 13-13-13) was used as a control. The experiment reveals that the rate of organic fertilizer application significantly affected the growth and flower yield of marguerite daisy. Still, the frequency of organic fertilizer application did not significantly affect it. Application at 1.5 times yielded the most significant growth and flower production compared to 0.5 and 1 times of application. When comparing the results of organic fertilizer application with 13-13-13 chemical fertilizer, it was found that applying organic fertilizer 1.0 times and 1.5 times every 15 or 30 days resulted in similar plant growth and flower size as with the chemical fertilizer ($p>0.05$). However, chemical fertilizer produced the highest chlorophyll index (SPAD), accumulated flower buds, and flower blooming per pot ($p<0.05$). The plant requires more than 1.5 times the organic fertilizer application to achieve flower production equivalent to chemical fertilizer.

Keywords: Fertilizer ratios; Timing of fertilizer application; Optimum fertilizer; Edible flower

1. Introduction

Flowers are rich in natural antioxidants, including flavonoids, anthocyanins, and many other compounds [1,2]. The demand for flowers has increased due to their attractive colors, appearance, and taste, leading to their use in the perfume and food industries. Flowers used for culinary purposes, known as "edible flowers," have become more popular because of their aroma, texture, flavor, color, shape, and phytochemical composition [1,2]. Edible flowers are commonly used as garnish for salads, soups, desserts, and beverages. Examples of flowers used in this way include roses, marigolds, hibiscus, calendula, and daisy [3]. The genus *Argyranthemum* is a notable source of secondary metabolites. Maguerite daisy (*Argyranthemum frutescens*) exhibits antimicrobial activity against Gram-positive

and Gram-negative bacteria and cytotoxic activity against HeLa and Hep-2 cell lines [4].

The traditional cultivation of edible flowers for food purposes relies on organic production. Organic production is a system that promotes and enhances the health and biodiversity of agroecosystems. Organic management practices combine traditional methods, innovation, and science to promote environmental and economic sustainability [5]. Moreover, organic production utilizes fewer additives and fertilizers to minimize hazards to consumers and environmental health. The nutrient sources in organic production rely on organic materials, both in fresh or dry form and composted materials. Manure, including urine from poultry, pigs, and cattle, is the primary source of nutrients from animal origin. To prevent nitrate contamination, farmyard manure, solid animal excrement, and liquid animal excrement should be used with care [6]. Composting organic materials into stable, humidified, and pathogen-free material is another source of nutrients in organic practices. Organic fertilizers made from organic materials release nutrients slowly. The application of organic fertilizer is usually done at sowing or transplanting stages.

Organic fertilizer improves soil physiochemical properties and increases total organic carbon, water-soluble carbon, and soil microbial diversity and activity [7, 8]. Applying bio-organic fertilizer at a rate of 10,000 kg/ka with micro-moistening irrigation resulted in greater net photosynthesis rate, water use efficiency, flower yield, and improved nutritional quality of edible rose compared to bio-organic fertilizer at a rate of 15,000 kg/ka with the same water supply method [9]. The split-split application of organic manure increased individual fruit weight over split and single-dose applications. The rate of organic manure application of 30 t/ha gave eggplant a higher number and fruit weight than a lower dose of organic manure (10 and 20 t/ha). However, the application of 30 t/ha gave the highest levels of vitamin B1 and B2, although this was not statistically different from 20 t/ha [10].

Presently, edible flowers are commonly grown soilless in potting media or hydroponically in greenhouses or outdoors. Typically, the controlled-release fertilizer Osmocote 13-13-13 is preferred for cultivating ornamental plants. This chemical fertilizer is re-applied every 2 or 3 months. Meanwhile, organic fertilizers contain nutrients mostly in organic form and act as slow-release fertilizers, providing nutrients in lower amounts over an extended period. Organic fertilizer could substitute for slow-release chemical fertilizer, potentially reducing production costs and nutrient leaching. The present study focuses on the growth and flower yield of marguerite daisy using organic fertilizer at different rates and frequencies of application.

2. Materials and Methods

2.1 Study area

The experiment was conducted under greenhouse conditions at the School of Agricultural Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, from August to September 2023.

2.2 Experimental design

The experiment was designed as a factorial in a completely randomized design with two factors. The first factor was the rate of organic fertilizer application (5, 10, and 15 g/pot equal to 0.5, 1, and 1.5 times compared to the total nitrogen content in chemical fertilizer), and the second factor was the frequency of fertilizer application (every 30 and 15 days). A slow-release chemical fertilizer (Osmocote 13-13-13) was used as a control at a rate of 1 g/pot. The application rate and nutrient content in each treatment are presented in Table 1. The organic fertilizer was made from composted soybean meal, eucalyptus bark, and pineapple peel, with a ratio of 3:2:1 for 75 days. The compost was turned every week, and moisture was maintained at 60% until the end of the composting process. The pH was measured by a pH meter. Total organic matter was determined by the Walkley and Black method. Total nitrogen was determined using the Kjeldahl method. Total P_2O_5 was determined by the indirect method. Total K_2O was determined by the flame photometric method. CaO and MgO were determined using double acid digestion followed by the atomic adsorption spectrophotometric method. Total sulfur was measured using the turbidimetric method. The properties of the organic fertilizer are detailed in Table 2.

One-month-old marguerite daisy plantlets were transplanted into potting media, with two plantlets per 6-inch round black plastic pot. During transplanting, 70% of the organic fertilizer was applied, and the remaining 30% was applied every 30 or every 15 days according to the treatment. In contrast, the total amount of chemical fertilizer was applied at the time of transplantation. Daily watering was conducted, and the plants were grown for 60 days.

Table 1. Application rate and nutrient content for the treatments in the study.

Treatment	Fertilizer application (g/pot)				Total fertilizer application (g/pot)	Nutrient in fertilizer (g/pot)		
	Transplant	15 DAT*	30 DAT	45 DAT		N	P_2O_5	K_2O
0.5T_30D	3.5	0	1.5	0	5	0.065	0.025	0.05
0.5T_15D	3.5	0.5	0.5	0.5	5	0.065	0.025	0.05
1.0T_30D	7	0	3	0	10	0.13	0.05	0.1
1.0T_15D	7	1	1	1	10	0.13	0.05	0.1
1.5T_30D	10.5	0	4.5	0	15	0.195	0.075	0.15
1.5T_15D	10.5	1.5	1.5	1.5	15	0.195	0.075	0.15
13-13-13	1	0	0	0	1	0.13	0.13	0.13

* DAT = Day after transplanting

Table 2. Properties of organic fertilizer used in the study.

Parameter	pH	OM (%)	Total N (%)	Total P_2O_5 (%)	Total K_2O (%)	Total CaO (%)	Total MgO (%)	Total S (%)
Value	7.6	21.2	1.3	0.5	1.0	10.7	2.3	0.9

2.3 Plant growth and flower yield

The growth of marguerite daisy, including plant height and canopy diameter, was recorded in centimeters on days 30 and 60 of the experiment. The chlorophyll index (SPAD) was measured on fully expanded leaves using a Konica Minolta SPAD-502 Plus on days 30 and 60 of the experiment. The plants were cut at ground level, with the upper part as a shoot and the remaining part as a root. The fresh weight and dry weight (after oven-drying at 70ºC for 48 hours) of both the shoot and root of the maguerite daisy were measured at the end of the experiment (day 60).

Flower bud and flower blooming were initially estimated one month after transplanting, and measurements were taken every week until 60 days after transplanting. The number of days until the first blooming was recorded, and flower diameter was measured using fully-bloomed flowers using a vernier caliper.

2.4 Statistical analysis

Analysis of variance (ANOVA) for the factorial design was conducted. Compared with chemical fertilizer, the comparison of means for interactions was estimated using Duncan's Multiple Range Test (DMRT) with a completely randomized design. All analyses were performed using SPSS 17.

3. Results and Discussion

3.1 Plant growth

The organic fertilizer application rate significantly affected every plant growth parameter, but the intensity of use did not affect plant growth (Table 3-4). The organic fertilizer application rate affected plant height, canopy diameter, SPAD, shoot, and root weight. Applying 1.5 times resulted in the most significant plant height, canopy diameter, and SPAD, although it was not statistically different from the results obtained with 1.0 times (Table 3). Furthermore, applying 1.5 times led to the highest shoot fresh and dry weights, followed by 1.0 and 0.5 times, respectively (Table 4). Additionally, the application of 1.0 times resulted in the highest root fresh and dry weight, although there was no statistical difference compared to the application of 1.5 times of organic fertilizer.

Applying a higher dose of organic fertilizer tended to decrease root weight (Table 4). This finding follows Bi and Evans [11], which found a reduction in the root rating of marigolds with high doses of broiler chicken litter-based organic fertilizer, 4-2-2, and 3-3-3, compared to low to medium doses. Moreover, the plants showed symptoms associated with excessive fertilization, possibly due to higher electrical conductivity. Nitrogen plays a crucial role in chlorophyll manufacture through photosynthesis, promoting green leafy growth, and supporting fruit and seed development—phosphorus aids in transferring energy throughout the plant, particularly for root development and flowering. Potassium is essential for photosynthesis and regulates numerous

metabolic processes for growth and fruit and seed development. Greater applications of organic fertilizer also increase plant nutrient intake, leading to improved growth in line with the rate of application increment.

The frequency of use affected plant growth and biomass. A trend indicated that application every 30 days resulted in more significant plant growth except for root weight, which showed a trend toward higher values with application every 15 days. It appears that the rate of organic fertilizer application is more influential than the intensity of application. Nutrients in organic fertilizer are mainly organic, which undergoes slow mineralization for plant utilization. The release of available nitrogen from organic fertilizers varies depending on factors such as temperature, duration, the C/N ratio of the materials, and the stabilization process [12, 13]. For instance, final mineralized nitrogen from poultry manure can range from 3 to 75% [12]. Compost and vermicompost release nitrogen in the 4.0-16.5% range, with organic nitrogen mineralized at a lower rate (-0.9-7.7%) [13]. Nitrogen release can be observed in the 25-45% range during 35 days of incubation of organic fertilizers, such as turkey manure and mushroom substrate [14].

Table 3. Canopy diameter, plant height, and SPAD value of marguerite daisy fertilized with different rates and frequencies of organic fertilizer.

	Plant height (cm)		Canopy diameter (cm)		SPAD	
	30 DAT*	**60 DAT**	**30 DAT**	**60 DAT**	**30 DAT**	**60 DAT**
Fertilizer rate						
0.5T	12.20 b	15.67 b	12.35 b	14.20 c	23.92 c	21.34 b
1.0T	17.52 a	20.20 a	16.05 a	20.40 b	30.64 b	24.01 a
1.5T	17.37 a	21.00 a	14.82 a	23.70 a	33.15 a	24.97 a
Frequency						
30D	15.82	19.40	14.72	19.61	29.35	23.68
15D	15.57	18.51	14.09	19.26	29.12	23.20
Fertilizer rate	**	**	**	**	**	**
Frequency	ns	ns	ns	ns	ns	ns
Fertilizer rate x Frequency	ns	ns	ns	ns	ns	ns
CV (%)	17.2	14.5	13.2	22.1	13.9	8.4

* DAT = Day after transplanting.
ns = non significant, **= significantly different at $p < 0.01$. Means within the same column followed by the same letter are not significantly different by DMRT.

3.2 Flower development

The organic fertilizer application rate significantly affected flower quality, but the application intensity did not. The application rate of 1.5 times resulted in the highest accumulation of flower buds per pot (84.5 buds), accumulated flowers blooming per pot (20.58 flowers), day-to-first flower blooming (9.17 days), and flower diameter (2.78 cm) (Table 5). This was followed by the application rates of 1.0 and 0.5 times, which were statistically different.

An increase in the organic fertilizer rate increased flower numbers and improved flower quality while reducing the time spent flowering. During the flowering or fruiting stage, higher phosphorus and potassium levels are required to support flower and fruit formation. Moreover, nitrogen affects the flower's life cycle, including vegetative and reproductive phases. Flower size, stem length, number of flowers per plant, and color were reduced by nitrogen deficiency. Therefore, achieving the optimum level of nitrogen supply in each growth stage is crucial for flower crop production [15]. Different plants have varying nutrient requirements at different stages of growth. During the vegetative stage, plants require higher nitrogen levels to promote leaf and stem development. Meanwhile, higher levels of phosphorus and potassium are needed to support flower and fruit formation during the flowering or fruiting stage.

Table 4. Shoot fresh weight, shoot dry weight, root fresh weight, and root dry weight of marguerite daisy fertilized with different rates and frequencies of organic fertilizer.

		Shoot weight (g)		Root weight (g)	
		Fresh weight	**Dry weight**	**Fresh weight**	**Dry weight**
Fertilizer rate					
	0.5T	3.69 c	1.19 c	3.61 b	1.34 b
	1.0T	9.22 b	2.77 b	9.39 a	3.39 a
	1.5T	16.14 a	4.38 a	7.53 a	2.23 ab
Frequency					
	30D	10.20	3.02	5.60	2.35
	15D	9.85	2.71	8.20	2.42
Fertilizer rate		**	**	**	**
Frequency		ns	ns	ns	ns
Fertilizer rate x Frequency		ns	ns	ns	*
CV (%)		60.3	51.4	47.8	55.2

ns = non significant, *= significantly different at $p < 0.05$, **= significantly different at $p < 0.01$. Means within the same column followed by the same letter are not significantly different by DMRT.

Table 5. The flower quality of marguerite daisy is fertilized at different rates and frequencies of organic fertilizer.

	Accumulate flower bud/pot (bud)	**Accumulated blooming/pot (flower)**	**Day to blooming (day)**	**Flower diameter (cm)**
Fertilizer rate				
0.5T	29.67 c	5.92 c	11.00 a	2.4 b
1.0T	58.41 b	15.50 b	11.83 a	2.64 a
1.5T	84.5 a	20.58 a	9.17 b	2.78 a
Frequency				
30D	56.39	14.16	10.61	2.67
15D	58.67	13.83	10.72	2.54
Fertilizer rate	**	**	*	**
Frequency	ns	ns	ns	ns
Fertilizer rate x Frequency	ns	ns	ns	ns
CV (%)	45.9	50.9	22.7	9.2

ns = non significant, *= significantly different at $p < 0.05$, **= significantly different at $p < 0.01$. Means within the same column followed by the same letter are not significantly different by DMRT.

3.3 Comparison between organic fertilizer and slow-released chemical fertilizer

The interaction between the rate and intensity of organic fertilizer application was compared with a slow-released chemical fertilizer (Osmocote 13-13-13) that is typically used in flower production for the growth of marguerite daisy over 60 days, as shown in Table 6. Application rates of 1 time and 1.5 times resulted in canopy diameter, shoot fresh and dry weight, root fresh weight, and dry weight that were not statistically different from those observed with chemical fertilizer use. Additionally, applying 1.5 times of organic fertilizer every 30 days resulted in greater plant height than chemical fertilizer application, with statistically significant differences. Chemical fertilizer application also led to higher SPAD values than organic fertilizer use. Furthermore, the application of organic fertilizer showed a yellower leaf color (Fig. 1). Lalk et al. [16] reported that conventional fertilizer (Osmocote® 15-9-12 applied 20 g/pot) resulted in higher values for plant growth, leaf SPAD, fruit yield, and photosynthetic rate with strawberries than organic fertilizer (5N-1.3P-3.3K applied 60 g/pot). Nutrients in organic fertilizers are in organic form and must mineralize for the nutrients to be available for plant uptake, resulting in a slow release of nutrients. Gaskell et al. [17] proposed that large

quantities and continuous application of organic fertilizers are required to achieve certain fertility and soil organic matter levels for optimal yield in the organic farming of strawberries.

SPAD values were significantly correlated with nitrogen and phosphorus concentrations in grapevine leaves. In contrast, SPAD values were correlated with nitrogen, calcium, potassium, and magnesium levels in apples [18] and nitrogen and magnesium content in blueberry plants (*Vaccinium corymbosum* L.) [19]. A strong correlation was observed among SPAD readings, the total yield, and the marketable yield of sweet potatoes [20]. Leaf chlorophyll content at 79 days after sowing correlated well with rice grain yield [21].

Table 6. Growth comparison of marguerite daisy fertilized with different rates and frequencies of organic fertilizer compared with Osmocote 13-13-13.

Treatment	Canopy diameter (cm)	Plant height (cm)	SPAD	Shoot weight (g)		Root weight (g)	
				Fresh weight	Dry weight	Fresh weight	Dry weight
0.5T_30D	14.37 c	15.80 c	21.74 cd	4.07 b	1.30 c	2.86 b	1.61 b
0.5T_15D	14.03c	15.53 c	20.94 d	3.12 b	1.02 c	4.75 ab	0.94 b
1.0T_30D	20.13 b	20.13 ab	23.51 bc	9.25 ab	2.85 bc	9.74 a	4.03 a
1.0T_15D	20.67 ab	20.27 ab	24.50 b	9.20 ab	2.69 bc	9.04 a	2.75 ab
1.5T_30D	24.33 ab	22.27 a	25.77 b	17.29 a	4.90 a	5.40 ab	1.41 b
1.5T_15D	23.07 ab	19.73 b	21.17 b	14.98 a	3.86 ab	9.65 a	3.06 ab
13-13-13	24.93 a	19.87 b	30.36 a	9.73 ab	2.75 bc	9.51 a	2.93 ab
F-test	**	**	**	**	**	**	**
CV (%)	23.3	13.4	12.7	55.8	47.8	43.4	49.8

**= significantly different at $p < 0.01$. Means within the same column followed by the same letter are not significantly different by DMRT.

Applying chemical fertilizer resulted in the highest accumulation of buds and flowers per pot (Table 7). However, there was no statistically significant difference in flower diameter between chemical and organic fertilizers, except for the application rate of 0.5 times every 15 days (Table 7, Fig. 2). The day-to-first flower blooming did not show statistically significant differences between treatments. However, applying organic fertilizer 1.5 times every 15 days tended to have the fastest blooming rate.

Table 7. Flower comparison of marguerite daisy fertilized with different rates and frequencies of organic fertilizer compared with Osmocote 13-13-13.

Treatment	Accumulated bud/pot (bud)	Flower diameter (cm)	Accumulate blooming/pot (flower)	Day to blooming (day)
0.5T_30D	32.17 d	2.51 ab	6.67 d	10.17
0.5T_15D	27.17 d	2.28 b	5.17 d	11.83
1.0T_30D	51.83 c	2.75 a	15.83 bc	11.50
1.0T_15D	65.00 c	2.53 ab	15.17 c	12.1
1.5T_30D	85.17 b	2.74 a	20.00 bc	10.17
1.5T_15D	83.83 b	2.81 a	21.17 b	8.17
13-13-13	109.00 a	2.78 a	28.67 a	10.8
F-test	**	**	**	ns
CV (%)	48.0	12.7	54.2	22.9

ns = non significant, **= significantly different at $p < 0.01$. Means within the same column followed by the same letter are not significantly different by DMRT.

The lower flower yield of the organic fertilizer treatment compared to the chemical fertilizer may result from the different forms of nitrogen and lower phosphorus content of the organic fertilizer (Table 1). Nitrogen significantly affects plant growth and flower quality. Although the total nitrogen content in organic and chemical fertilizers is the same, they differ in their forms. Nitrogen in organic fertilizer mainly exists in

the form of organic nitrogen, while Osmocote 13-13-13 contains nitrogen in the form of ammonium nitrogen (NH_4-N) at 7.6% and nitrate nitrogen (NO_3-N) at 5.4% [22]. The fertilizer takes 13.0 days and 10.7 days to release 50% of NH_4-N and NO_3-N, respectively. About 80% of NH_4-N and NO_3-N are released within 20 days. Additionally, 65% potassium and 35% phosphorus are released within 20 days [22]. On average, NH_4-N, NO_3-N, phosphorus, and potassium released from Osmocote 13-13-13 are 99.0, 108.7, 68.9, and 99.1%, respectively [22]. It is recommended to reapply Osmocote every 3 months. Niedzinski [14] reported that nitrogen release ranges from 25 to 45% during a 35-day incubation period for organic fertilizers (turkey manure and mushroom substrate), while mineral fertilizers release 55 to 95% of nitrogen.

Phosphorus is another limiting nutrient in organic fertilizer. The organic fertilizer contained 0.5% of P_2O_5, whereas Osmocote 13-13-13 contained 13% of P_2O_5. Phosphorus supports energy transfer throughout the plant for root development and flowering. The amount of phosphorus is related to various yield-governing parameters, such as the number of branches, flowers per plant, flower diameter, and flower weight in marigolds [23].

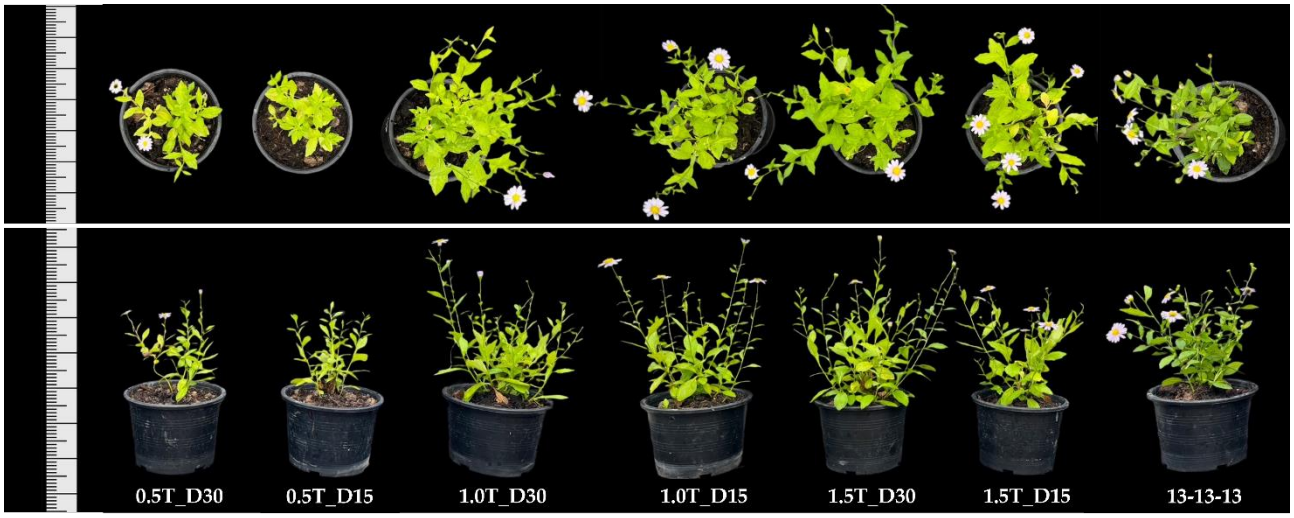


Figure 1. Characteristics of marguerite daisy growth during the experiment.

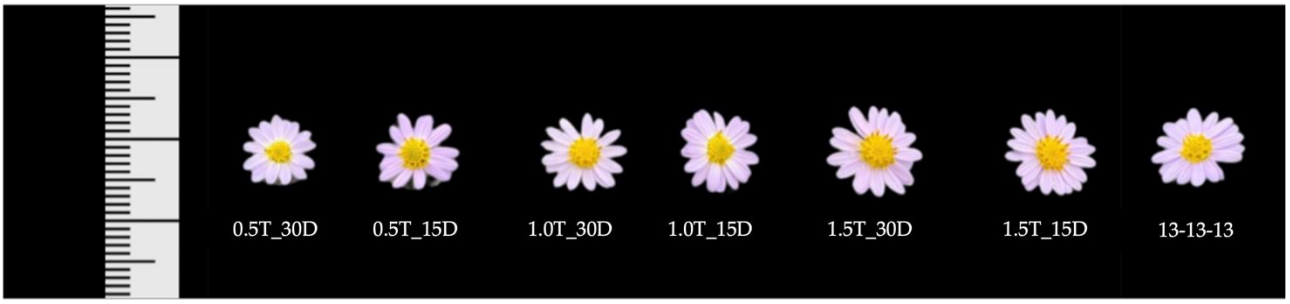


Figure 2. Size comparison of marguerite daisy flower growth with different treatments.

Plant nutrients play an essential role in growth and production. Fertilizers that provide nutrients in different forms and quantities produce varied plant growth and yields. Moreover, fertilizer is a factor that influences secondary metabolites in plants. Indian aster (*Kalimeris indica*) growing in biocomposite (derived from fruits and vegetables) exhibited higher contents of anthocyanins, flavonoids, sugars, organic acids, aldehydes, and alcohol compared to those grown in organic fertilizer derived from animal manure [24]. Furthermore, the flowers were more purple and had a fruity aroma, making them more attractive for consumption [24]. The quality of marguerite daisies, including secondary metabolites in flowers grown under organic production, requires further research.

In the present study, the application rate strongly affected marguerite daisy growth and flower yield. Applying 1.5 times the nitrogen content in slow-release chemical fertilizer resulted in similar growth and

flower quality compared to chemical fertilizer. However, applying 1.5 times of organic fertilizer was inadequate to supply the flower yield and achieve a leaf SPAD level equal to chemical fertilizer. Therefore, more than 1.5 times of organic fertilizer may be required to increase flower yield, and it may be necessary to supplement with liquid fertilizer to increase phosphorus levels.

4. Conclusions

Plant nutrients play an essential role in growth and production. Fertilizers that provide nutrients in different forms and quantities produce varying plant growth and yields. The higher rate of organic fertilizer application increases marguerite daisy growth and flower yield, with the highest yield observed with 1.5 times application. The intensity of organic fertilizer application, whether every 30 days or every 15 days, did not affect the plant's growth and flower production. Applying organic fertilizer at 1.5 times the rate of Osmocote 13-13-13 resulted in similar growth and flower quality of marguerite daisy compared to chemical fertilizer. However, it yielded lower flower and leaf SPAD levels than chemical fertilizer. Suggestions indicate that applying more than 1.5 times the amount of organic fertilizer or supplementing with foliar application of bio-extract to increase nutrient supply could result in marguerite daisy growth and yield comparable to chemical fertilizer.

5. Acknowledgements

This research was supported by the School of Agricultural Technology, King Mongkut's Institute of Technology Ladkrabang. The author would like to thank Dr. Krichanont Iyapunya for providing the organic fertilizer material.

Author Contributions: Conceptualization, C.A., P.K., and P.T.; methodology, C.A., and P.T.; formal analysis, P.T.; investigation, C.A.; writing—original draft preparation, P.T.; writing—review and editing, P.T., C.A., P.K.

Funding: This research was funded by the School of Agricultural Technology, King Mongkut's Institute of Technology Ladkrabang.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Prabawati, N.B.; Oktavirina, V.; Palma, M.; Setyaningsih, W. Edible flowers: antioxidant compounds and their functional properties. *Horticulturae*. **2021**, *7*(4), 66. https://doi.org/10.3390/horticulturae7040066

[2] Purohit, S.R.; Rana, S.S.; Idrishi, R.; Sharma, V.; Ghosh, P. A review on nutritional, bioactive, toxicological properties and preservation of edible flowers. *J. Future Foods.* **2021**, *4*, 100078. https://doi.org/10.1016/j.fufo.2021.100078

[3] Netam, N. Edible flower cultivation: A new approach in floriculture industry. *The Pharma Innovation Journal*. **2021**, *10*(3), 857-859. https://doi.org/10.22271/tpi.2021.v10.i3l.5896

[4] Gonzalez, A.G.; Estevez-Reyes, R.; Estevez-Braun, A.; Ravelo, A.G.; Jimenez, I.A.; Bazzocchi, I.L.; Aguilar, M.A.; Moujir, L. Biological activities of some Argyranthemum species. *Phytochemistry*. **1997**, *45*(5), 963-967. https://doi.org/10.1016/S0031-9422(97)00063-0

[5] Fernandez, J.A.; Ayastuy, M.E.; Belladonna, D.P.; Comezana, M.M.; Contreras, J.; Mourao, I.M.; Orden, L.; Rpdriguez, R.A. Current trends in organic vegetable crop production: practices and technique. *Horticulturae*. **2022**, *8*, 893. https://doi.org/10.3390/horticulturae8100893

[6] Jones. C.S.; Drake, C.W.; Hruby, C.E.; Schilling, K.E.; Wolter, C.F. Livestock manure driving stream nitrate. *Ambio*. **2019**, *48*(10), 1143-1153. https://doi.org/10.1007/s13280-018-1137-5

[7] Singh, V.; Sharma, S.; Kumar, P.; Bhardwaj, S.; Gautam, H. Conjoint application of bio-organic and inorganic nutrient sources for improving cropping behaviour, soil properties and quality attributes of apricot (*Prunus armeniaca*). *Indian J. Agric. Sci.* **2010**, *80*, 981-987.

[8] Wang, J.; Li, X.; Xing, S.; Ma, Z.; Hu, S.; Tu, C. Bio-organic fertilizer promotes plant growth and yield and improves soil microbial community in continuous monoculture system of *Chrysanthemum morifolium* cv. Chuju. *Int. J. Agric. Biol.* **2017**, *19*, 563-568. https://doi.org/10.17957/IJAB/15.0339

[9] Liu, X.; Zhang, Y.; Jiang, Z.; Yue, X.; Liang, J.; Yang, Q.; Li, J.; Li, N. Micro-moistening irrigation combined with bio-organic fertilizer: An adaptive irrigation and fertilization strategy to improve soil

environment, edible Rose yield, and nutritional quality. *Ind. Crops Prod*. **2023**, *196*, 116487. https://doi.org/10.1016/j.indcrop.2023.116487

[10] Agbo, C.U.; Chukwudi, P.U.; Ogbu, A.N. Effects of rates and frequency of application of organic manure on growth, yield and biochemical composition of *Solanum melongena* L. (cv. "Nawa local") fruits. *J Anim Plant Sci.* **2012**, *14* (2), 1952-1960.

[11] Bi, G.; Evans, W.B.; Spiers, J.M.; Witcher, A. Effects of organic and inorganic fertilizers on marigold growth and flowering. *Hort. Science*. **2010**, *45*(9), 1373-1377. https://doi.org/10.21273/HORTSCI.45.9.1373

[12] Geisseler, D.; Smith, R.; Cahn, M.; Muramoto, J. Nitrogen mineralization from organic fertilizers and composts: Literature survey and model fitting. *J. Environ. Qual.* **2021**, *50*(6), 1325-1338. https://doi.org/10.1002/jeq2.20295

[13] Vione, E.L.B.; Drescher, G.L.; da Silva, L.S.; Giacomini, S.J.; Aita, N.T.; da Silva, A.A.K.; Prigol, L.H.F. Nitrogen mineralization of compost and vermicompost from different animal manure and its recovery by lettuce using 15N. *SSRN*. **2023**. http://dx.doi.org/10.2139/ssrn.4385890

[14] Niedzinski, T.; Sierra, M.J.; Labetowicz, J.; Noras, K.; Cabrales, C.; Millan, R. Release of nitrogen from granulate mineral and organic fertilizers and its effect on selected chemical parameters of soil. *Agronomy*. **2021**, *11*, 1981. https://doi.org/10.3390/agronomy11101981

[15] Ruamrungsri, S.; Panjama, K.; Ohyama, T.; Inkham, C. Nitrogen in flowers. In T. Ohyama & K. Inubushi (Eds.), *Nitrogen in Agriculture*. IntechOpen. 2021. https://doi.org/10.5772/intechopen.98273

[16] Lalk, G.T.; Bi, G.; Stafne, E.T.; Li, T. Fertilizer type and irrigation frequency affect plant growth, yield, and gas exchange of containerized strawberry cultivars. *Technology in Horticulture*. **2023**, *3*, 1-8. https://doi.org/10.48130/TIH-2023-0003

[17] Gaskell, M.; Bolda, M.P.; Muramoto, J.; Daugovish, O. Strawberry nitrogen fertilization from organic nutrient sources. *Acta Hortic*. **2009**, *842*, 385-88. https://doi.org/10.17660/ActaHortic.2009.842.74

[18] Porro, D.; Dorigatti, C.; Stefanini, M; Ceschini, A. Use of SPAD meter in diagnosis of nutritional status in apple and grapevine. *Acta Hortic*. **2001**, *564*, 243-252. https://doi.org/10.17660/ActaHortic.2001.564.28

[19] Pinzón-Sandoval E.H.; Balaguera-López, H.E.; Almanza-Merchán, P.J. Evaluation of SPAD index for estimating nitrogen and magnesium contents in three blueberry varieties (*Vaccinium corymbosum* L.) on the Andean Tropics. *Horticulturae.* **2023**, *9*, 269. https://doi.org/10.3390/horticulturae9020269

[20] Pepó, P. Correlation analysis of the SPAD readings and yield of sweet potato (*Ipomoea batatas L.*) under different agrotechnical conditions. *Aust. J. Crop Sci*. **2020**, *14*(05), 761-765. https://doi.org/10.21475/ajcs.20.14.05.p2124

[21] Ramesh, K.; Chandrasekaran, B.; Balasubramanian, T.; Bangarusamy, U.; Sivasamy, R.; Sankaran, N. Chlorophyll dynamics in rice (*Oryza sativa*) before and after flowering based on SPAD (chlorophyll) meter monitoring and its relation with grain yield. *J. Agron. Crop Sci*. **2002**, *188*, 102-105. https://doi.org/10.1046/j.1439-037X.2002.00532.x

[22] Broschat, T.K.; Moore, K.K. Release rates of ammonium-nitrogen, nitrate-nitrogen, phosphorus, potassium, magnesium, iron, and manganese from seven controlled-release fertilizers. *Commun. Soil Sci. Plant*. **2007**, *38*, 843-850. https://doi.org/10.1080/00103620701260946

[23] Dangi., S.P.; Aryal, K.; Magar, P.S.; Bhattarai, S.; Shrestha, D.; Gyawali, S.; Basnet, M. Study on effect of phosphorus on growth and flowering of marigold (*Tagetes erecta*). *JOJ Wildl Biodivers.* **2019**, *1*(5), 555571. https://doi.org/10.19080/JOJWB.2019.01.555571

[24] de Morais, J.S.; Cabral, L.; da Costa, W.K.A.; Uhlmann, L.O.; Lima, M.S.; Noronha,M.F.; dos Santos, S.A.; Madruga, M.S.; Olegario, L.S.; Wagner, R.; Sant'Ana, A.S.; Magnani, M. Chemical and volatile composition, and microbial communities in edible purple flowers (*Torenia fournieri* F. Lind.) cultivated in different organic systems. *Food Res. Int*. **2022**, *162*, 111973. https://doi.org/10.1016/j.foodres.2022.111973

ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

Enhancing Linear Regression Through Neighbor-based Similarity Analysis

Chinnawat Chetcharungkit[1*]

[1] Department of Mathematics, Faculty of Science, Kasetsart University, Bangkok, 10903, Thailand
* Correspondence: fscicwc@ku.ac.th

Citation:
Chetcharungkit, C. Enhancing linear regression through Neighbor-based similarity analysis. *ASEAN J. Sci. Tech. Report.* **2024**, *27*(5), e251974. https://doi.org/10.55164/ajstr.v27i5.251974

Article history:
Received: December 6, 2023
Revised: August 8, 2024
Accepted: July 30, 2024
Available online: September 7, 2024

Abstract: When working with real-world datasets characterized by complex and non-linear relationships, the limitations of non-complex machine learning models like linear regression become evident. In response to addressing this technical problem, we propose a novel algorithm to enhance linear regression without the necessity of complex mathematical or statistical expressions. Instead, the algorithm segments data into multiple subgroups or neighbors, each with its best-fitting line. The primary objective of this approach is to enable more accurate predictions for unseen data points by utilizing the most similar neighbors and their corresponding linear regression lines, with the support of k-nearest neighbors. Empirical evidence from three publicly available housing price datasets demonstrates the algorithm's effectiveness in improving traditional linear regression models.

Keywords: Linear modeling; machine learning; K-nearest neighbor; boosting algorithm

1. Introduction

In today's era of precise and rapid advancements in computational capabilities, various advanced machine learning techniques, like modified gradient boosting trees and deep learning, have emerged to improve model performance significantly [1, 2]. These sophisticated methods can model complex non-linear relationships in data and accommodate various data types such as images and text [3, 4]. With growing research, these techniques have led to predictive modeling across finance, healthcare, and marketing domains. Nonetheless, amidst the ongoing global warming crisis, it is recommended that complex models like deep learning be used only for essential purposes as they consume a lot of energy. This results in significant carbon dioxide emissions. For instance, it is found that training the bidirectional encoder representations from the transformers language model (BERT LM), an advanced deep learning model, can emit as much carbon dioxide as a year's home energy consumption [5]. This suggests that complex models should be used only when necessary.

In contrast to complex models, linear regression is a simple machine learning method that models the relationship between dependent variables and one or more independent variables by fitting a linear equation to observed data. While linear regression may not effectively handle data with non-linear relationships like deep learning, its role in data science persists. It remains a fundamental and indispensable method in statistics, data analysis, and financial engineering [6]. Its enduring importance lies in simplicity and interpretability. Since linear regression offers rapid implementation, requiring minimal computational

resources and enabling swift model development compared to complex architectures of deep learning models, employing linear regression could extend global efforts to mitigate the threat of the global warming crisis. Also, due to its simplicity, individuals can easily understand and use linear regression, even with limited knowledge of machine learning. Moreover, linear regression demonstrates superiority when the relationships between variables are primarily linear. This is an example of a scenario where deep learning models cannot rival its performance. Yet, in data analysis, it becomes evident that many real-world datasets exhibit complex and non-linear relationships. For example, stock price data over time do not follow a straight line [7]. When we face such intricate data, it becomes essential to acknowledge the limitations of linear regression, as it may not be the most suitable approach. Nonetheless, rather than discarding linear regression, an attempt to enhance its capabilities to deal with data exhibiting non-linear patterns is intriguing. The success of this approach could eliminate the need for employing deep learning unless it is genuinely essential.

To date, numerous studies have sought to enhance regression models. The first approach belongs to a class of feature engineering intending to make the data more suitable and informative. This enables the model to learn patterns and make more precise predictions. The works in [8-10] are examples of this approach. However, feature engineering demands domain knowledge and a significant amount of time since we must carefully analyze the relationships between input features and target variables. This task can become challenging, especially when dealing with large input feature sets. The second approach is the so-called regularization. This technique adds a penalty term to the model's loss function [11-14]. The last approach introduces non-linearity to linear regression, such as using basis functions [15] and certain classes of statistical models like generalized linear models (GLMs) [16]. Unfortunately, the drawbacks of the latter two are that a solid knowledge of mathematics and statistics and coding skills are highly required. This complexity is further compounded when using standard platforms like scikit-learn in Python [17], as modifying the library's code as suggested by the mentioned papers might be nontrivial. Of course, these techniques may not suit every artificial intelligence (AI) developer, especially beginners.

Therefore, this research aims to develop a novel, user-friendly algorithm that can enhance the performance of linear regression models without involving the intricacies of daunting mathematical expressions, which might pose challenges to certain AI developers. Instead, we achieve this by incorporating a simple model like K-nearest neighbor (KNN) into our proposed algorithm. The algorithm's implementation is also designed to be straightforward, ensuring accessibility even for individuals not well-versed in coding. Three housing price datasets in [18-20] have been selected to assess the proposed algorithm's effectiveness. The choice of housing price data is due to its significance in various aspects. The first obvious reason is that they are standard and popular datasets for regression tasks. Apart from the technical reason, housing prices substantially impact the economy, as the real estate market plays a vital role in driving economic growth and stability [21]. In terms of modeling, some studies in [22-26] have attempted to construct machine learning models for housing price prediction. However, these studies relied on various common machine learning tools such as exploratory data analysis, linear regression, artificial neural networks, support vector machines, and gradient boosting trees, and none of them introduced novel methodologies.

2. Background

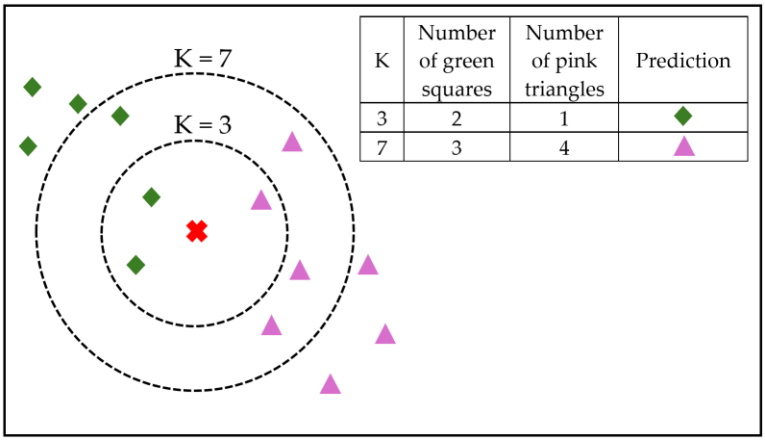
We commence by offering a concise overview of K-Nearest Neighbors (KNN), linear regression, and principal component analysis (PCA), as these methodologies will be subsequently applied in this research.

2.1 K-Nearest Neighbors (KNN)

K-nearest neighbors (KNN) is a straightforward and instinctive machine learning technique in classification and regression tasks. It is categorized within the class of instance-based, non-parametric approaches. The central idea behind KNN is that data points with similar features tend to belong to the same class or exhibit similar behaviors [15]. To determine similarity, in this study, we use the so-called Euclidean norm or L^2-norm defined as follows:

$$\|\boldsymbol{x}-\boldsymbol{y}\|_2=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Intuitively, the closer two vectors are, the smaller their L^2-norm is. Hence, the predictive outcome for classification tasks is the class label, determined by selecting the most frequently occurring class among the K nearest neighbors. To clearly understand how KNN works, Figure 1 demonstrates how the K-Nearest Neighbors (KNN) algorithm classifies a new data point (the red cross). If K is set to be 3, the three nearest neighbors (inside the smaller dashed circle) include 2 green squares and 1 pink triangle, resulting in a prediction of a green square due to the majority. However, if K = 7, the seven nearest neighbors (inside the larger dashed circle) include 3 green squares and 4 pink triangles, leading to a prediction of a pink triangle instead. It can be noted that the value of the hyperparameter K plays a crucial role in determining the classification outcome by considering different sets of nearest neighbors. To find an optimal value of K, we can test various K values and choose the one that provides the best performance on a test set.



K	Number of green squares	Number of pink triangles	Prediction
3	2	1	◆
7	3	4	▲

Figure 1. The illustration of how the K-Nearest Neighbors (KNN) algorithm works to classify new data point, represented by the red cross, with K = 3 and 7.

2.2 Linear Regression

Let $\boldsymbol{X}$ be an input matrix in $\mathbb{R}^{n \times d}$ and $\boldsymbol{y}$ a target vector in $\mathbb{R}^n$. Linear regression aims to find a vector of parameters $\boldsymbol{w}$ in $\mathbb{R}^d$ such that $\|\boldsymbol{Xw} - \boldsymbol{y}\|_2^2$ is minimized as much as possible. It is well-known that the solution to this optimization problem is $\boldsymbol{w} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}$ [27]. Note that since $\boldsymbol{w}$ is a result of matrix multiplication, it becomes clear that linear regression might exhibit poor performance in cases where the underlying pattern is not linear.

2.3 Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique used in data analysis and machine learning to transform a dataset into a new coordinate system, where the data's variance is maximized along the new axes. For example, if the original data has 200 input features, PCA can transform the dataset to reduce the number of features to any desired amount, such as 50. Moreover, with PCA, we can ensure that the new dataset with 50 input features effectively represents the original dataset.

PCA identifies the principal components (linear combinations of the original features) that capture the most significant variation in the data. These principal components are ordered by importance, allowing for the retention of the most informative components while reducing the dimensionality of the dataset. The algorithm for dimensionality reduction using PCA is as follows [15]:

Input Variables:

- **X**: Original standardized data matrix with dimensions $m \times n$ (where m is the number of samples and n is the number of features).
- k: Desired number of principal components, i.e., the number of features of the newly transformed data.

PCA Algorithm:

1. Calculate the covariance matrix (**Σ**) of **X.**

$$\mathbf{\Sigma} := \frac{1}{m}\mathbf{X}^{\mathrm{T}}\mathbf{X}$$

Decompose **Σ** using eigen-decomposition as a product of an orthogonal matrix **V** whose columns are the real-orthonormal eigenvectors of **Σ**, and a diagonal matrix **Λ** whose entries are all the eigenvalues of **Σ**.

$$\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$$

2. Sort the eigenvalues in **Λ** in descending order and rearrange the eigenvectors in **V** accordingly (if necessary).
3. Choose the top k eigenvectors corresponding to the largest k eigenvalues to form the matrix of principal components.

$$\mathbf{V}_k = \mathbf{V}[:, 1{:}k]$$

where $\mathbf{V}[:, 1{:}k]$ refers to the selection of the first k columns from the matrix **V**.

4. The output (transformed) data is given by $\mathbf{X}\mathbf{V}_k$, which is a matrix of dimension $m \times k$.

PCA is generally admitted as a valuable tool for simplifying complex data, visualizing patterns, and removing multicollinearity in feature sets, making it a fundamental technique in data preprocessing and exploratory data analysis. In this study, PCA is an optional tool that is not integrated into any part of the proposed algorithm. Instead, we only employ it as a tool for data preprocessing to reduce the dimensionality of data to avoid prolonged computational processing.

3. Motivation and Methods

Before presenting the pseudocode algorithm of our proposed method, let us first delve into the idea behind it.

3.1 Motivation

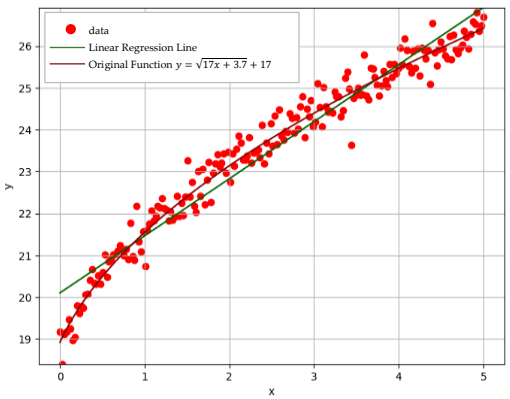


Figure 2. Scatter plot of 200 data points (red) with errors deviating from the equation $y = \sqrt{17x + 3.7} + 17$ (dark red) and the regression line (green).

In Figure 2, we simulate a scenario where the true model represents the relationship between the x-axis and y-axis through the equation $y = \sqrt{17x + 3.7} + 17$. After that, the 200 data points generated from the equation but with added noise or errors represent the data we collected. We also plot the linear regression line (in green) fitted from this data. Notably, the linear regression model did not perform well in this case, clearly due to the underlying non-linear pattern presented in the data.

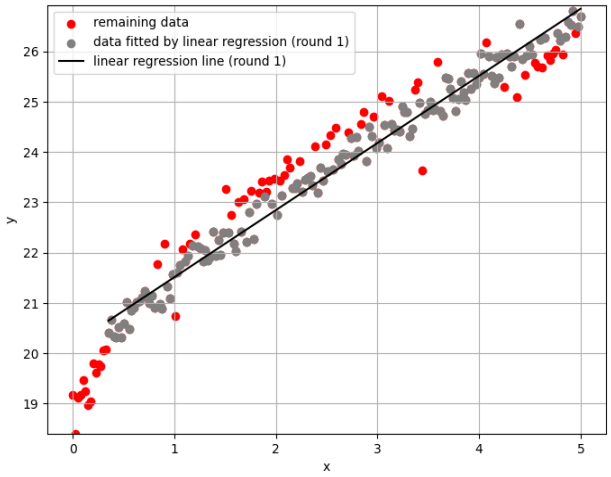


Figure 3. Round-1 linear regression line (black) fitted from the filtered data points (grey dots).

However, from Figure 3, if we pick only data points where the error between the predicted value and the actual value is less than the 70th percentile of the error observed in the entire dataset (in grey), the very linear regression line called round-1 linear regression line (in black), perform better with this filtered data.

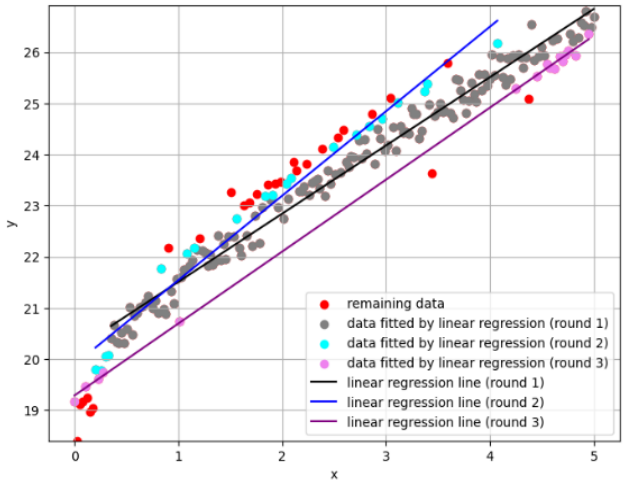


Figure 4. Round-2 and round-3 linear regression lines fitted from the corresponding filtered data points.

Next, we iterate the above step and yield the round-2 linear regression line (depicted in blue in Figure 4). It is obtained by fitting the remaining data in Figure 3, and subsequently, we select only the best-fitted data points (shown in cyan), as illustrated in Figure 4. We then repeat this process for one more round, yielding the round-3 linear regression line (in purple) and its corresponding data (in violet). Of course, this process will continue until all the data has been accounted for. Every data point has been assigned to an appropriate subgroup or neighbor and its corresponding linear regression line. At the end of this process, all the data is categorized into multiple neighbors, each with its own well-predicted linear regression line. When predicting the value of an unseen data point, we first determine which neighbor the new data point best fits through the k-nearest neighbor (KNN) model. Then, we employ that neighbor's corresponding linear regression line to receive the prediction. For example, suppose that the KNN model identifies a new data point resembling the

cyan neighbor's pattern. Then, we utilize the blue linear regression line to forecast the expected outcome of this unseen data. The above approach would make more accurate predictions based on the characteristics of similar data points within the same neighborhood. These concepts form the foundation of the algorithm proposed by this research, as shown in Table 1.

Table 1. The proposed algorithm

Pseudocode	
Input	
iter	the number of iterations to obtain multiple neighbors of data and their corresponding linear regression lines
n_perc	the *n*th percentile in the dataset
num_knn	the number of nearest neighbors to use for k-nearest neighbors (KNN) model
X_train	training data
y_train	target variable of the training data
X_test	test data
Output	
$X_1,\ldots,X_{iter}$	a sequence of neighbors of the data in each iteration
$l\_reg_1,\ldots,l\_reg_{iter}$	a sequence of the linear regression models corresponding to the neighbors of the data
y_pred	the predicted values from *X_test* by the proposed algorithm

Step 1: Find all the neighbors and build their corresponding linear regression models.

```
X = X_train
y = y_train
for i = 1 to iter :
    l_reg.fit(X, y)
    pred = l_reg.predict(X)
    dist = |pred - y|
    threshold_dist = n_perc in dist data
    if i ≠ iter :
        X_i = X whose dist ≤ threshold_dist
        l_reg_i = l_reg
        X = X not containing X_i
        y = y which corresponds to X
    else:
        X_i = X
        l_reg_i = l_reg
```

Step 2: Predict the target values.

```
# Determine the neighbor to which each data point in X_train belongs.
y_knn = [ i  for x in X_train  if x belongs to X_i ])

# Build a KNN model
KNN.fit(X_train, y_knn, n_neighbors = num_knn)

# Forecast the outcome values
for  x_j  in  X_test :
    neighbor =  KNN.predict(x_j)      # To get the neighbor to which x_j belongs
    y_pred_j = l_reg_neighbor.predict(x_j)    # Use the corresponding linear regression of neighbor to forecast
```

Table 1 shows three parameters for the algorithm: *iter*, *n_perc*, and *num_knn*. To find the optimal value for each parameter, we can experiment with various combinations of their settings to find the most effective one. This process is commonly referred to as "grid search." Iter aims to set the number of iterations required to obtain multiple neighbors of the data and their corresponding linear regression lines. N*_perc* removes data points that poorly fit the linear regression line. As for the last parameter, *num_knn* acts as a hyperparameter for the KNN model. Recall that KNN allows us to identify the most suitable neighbor to which any unseen data point belongs and then determine the linear regression line for prediction.

While KNN is integrated, this does not introduce additional complexity to the proposed algorithm. It can be noted that each line of code presented in Table 1 is just a routine command typically employed by data scientists. Furthermore, no modifications to the existing library code are required. All of these are to ensure that the implementation remains trouble-free for users.

Before applying the proposed algorithm to the datasets, it is vital to mention that Figure 4 serves only the purpose of visualizing how the algorithm works, even though concerns about overfitting might exist. However, this concern may not be as significant when dealing with real-world data because the simulated data in Figure 4 has only a quadratic pattern, less complex than any actual data. Also, linear regression is not a complex model. In particular, the overfitting issue is problematic because the model cannot accurately predict unseen data. Hence, if our algorithm can produce a model with better performance than the baseline model, it is legitimate to say the proposed algorithm is successful.

3.2 Methods

Now, we will employ the proposed algorithm to the three selected datasets. The process involves the following steps. For simplicity in evaluating the effectiveness of the proposed algorithm, some conventional steps like exploratory data analysis, data imputation to fill missing values, feature selection, and feature engineering will be omitted.

3.2.1 Data Preparation

The three datasets: house prices [18], California housing prices [19], and Boston house prices [20], underwent conventional techniques like dropping features with many missing values and removing rows containing missing values. All the categorical features were transformed using one-hot encoding (if necessary). Furthermore, in the case of the first dataset, the PCA algorithm in section 2.3 was applied due to its large number of input features (297) resulting from one-hot encoding. Setting the number of principal components to 150, the PCA-transformed data will have only 150 input features. Table 2 presents an overview of the preprocessing steps performed on each dataset during this phase.

Table 2. Summary of data characteristics and data preprocessing steps of the three datasets.

Item	House Prices	California Housing Prices	Boston House Prices
Size of raw data	1,094×76	20,433×10	506×14
Dropping column	Alley, PoolQC, MiscFeature, FireplaceQu, and Fence	None	None
Number of categorical features	42	1	0
PCA	Yes (150 principal components)	No	No
Number of input features	150	13	13
Target feature	SalePrice	median_house_value	MEDV

3.2.2 Train-test Split

Each dataset is split into training and test sets with a ratio of 67:33. The target feature for each dataset is shown in the final row of Table 2. This means we have *X_train, X_test, y_train*, and *y_test* for each housing dataset.

3.2.3 Model Training and Prediction

The proposed algorithm fits the input features (*X_train*) and the output feature (*y_train*) of the training dataset for each of the three datasets. However, since the proposed algorithm comprises three hyperparameters, *iter*, *n_perc*, and *num_knn*, different combinations of these values can yield various models with varying performance levels. Consequently, the optimal combination of these hyperparameter values is crucial. In this paper, to obtain the best parameter configuration for the algorithm, we loop through all the combinations of the following settings :

- *iter* ∈ {1, 2, 3 ,4, 5, 6, 7, 8},
- *n_perc* ∈ {0.3, 0.5, 0.7, 0.75},
- *num_knn* ∈ {1, 3, 17, 31, 59, 93}.

In more detail, we exhaustively tried all the 8×4×6 = 192 possible combinations from the given set of each hyperparameter to build 192 regression models and then predict the target values (*y_pred*) from the test datasets of the three housing datasets.

It is worth mentioning that the range or values of each hyperparameter can be set arbitrarily. The broader the ranges we specify, the higher the chance of finding more optimal parameters; however, this will require more time. This process involves a trade-off between the thoroughness of the search and the time required to perform it. While there is no guarantee of finding the absolute optimal solution, this approach ensures we identify the most optimal combination within the given set of hyperparameters.

3.2.4 Model Evaluation

Following the generation of 192 sets of predicted values (*y_pred*) of each housing dataset from the previous step, we compute several standard regression evaluation metrics using the *y_pred* values and the *y_test* values from subsection 3.2.2. These metrics will select the best model among the 192 models. The evaluation metrics include root mean square error (RMSE), R^2-score, mean absolute percentage error (MAPE), and mean absolute error (MAE), as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2}{n}} \tag{2}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{3}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\tilde{y}_i}{y_i}\right| \times 100 \tag{4}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i| \tag{5}$$

where y_i is the ith actual output value (the ith value of *y_test*), $\tilde{y}_i$ is the ith predicted value (the ith value of *y_pred*), $\bar{y}$ is the average of the actual output values, and n is the number of rows of the test set. Note that the four metrics share the common goal of measuring how well a model's predictions align with the actual observed values, although their formulae are different, and the best model is the one with the highest R^2-score and the lowest RMSE, MAPE, and MAE.

4. Results and Discussion

The most optimal parameter configurations for *iter*, *n_perc*, and *num_knn* across the three selected housing price datasets, along with the corresponding values of the four evaluation metrics for regression, are shown in Table 3. The evaluation metrics obtained from the traditional (baseline) linear regression models to assess the effectiveness of the proposed algorithm are also provided in this table.

Table 3. Best parameter configurations and the four evaluation metrics across the three datasets.

Item		House Prices	California Housing Price	Boston House Prices
Best of (*iter*, *n_perc*, *num_knn*)		(3, 0.75, 17)	(5, 0.3, 59)	(6, 0.5, 1)
RMSE	Baseline	51,186.151	66,572.595	4.907
	Proposed	47,448.611	65,484.518	4.260
R^2	Baseline	0.535	0.665	0.747
	Proposed	0.601	0.676	0.809
MAPE	Baseline	0.155	0.280	0.189
	Proposed	0.139	0.274	0.162
MAE	Baseline	24,534.293	48,767.613	3.671
	Proposed	22,408.018	47,832.910	3.094

Table 4. Percentage change in the evaluation metric values for the proposed models compared to the baseline models across the different datasets.

Dataset	Metric	Percentage Change (%)
House Prices	RMSE	-7.302
	R^2	12.336
	MAPE	-10.323
	MAE	-8.667
California Housing Prices	RMSE	-1.634
	R^2	1.654
	MAPE	-2.143
	MAE	-1.917
Boston House Prices	RMSE	-13.185
	R^2	8.300
	MAPE	-14.286
	MAE	-15.718

Among the three datasets, the Boston housing dataset shows the most robust linear pattern, with the highest R^2-score from the baseline model. In contrast, the House prices dataset displays the weakest linear pattern. To see a clear comparison, we present Table 4, visually represented by Figure 5. The table provides the percentage changes in the four performance metrics for the proposed models compared to the baseline models across the three housing price datasets. Note that when it comes to RMSE, MAPE, and MAE, negative percentages show enhancements in performance, whereas positive ones indicate deteriorations. Yet, this

situation becomes otherwise for R^2. This means that the proposed algorithm empirically enhances the performance of the traditional models.

Specifically, according to Figure 5, it is evident that the proposed algorithm significantly enhances the performance of the traditional linear regression model, achieving the best results in the case of the Boston house prices dataset, while the improvement is the least pronounced in the California housing price dataset. Furthermore, regardless of the underlying data pattern, the proposed model can effectively augment or replace the traditional linear regression model. This is evidenced by the substantial percentage changes in the evaluation metrics for the house prices dataset, which rarely exhibits a linear underlying pattern.

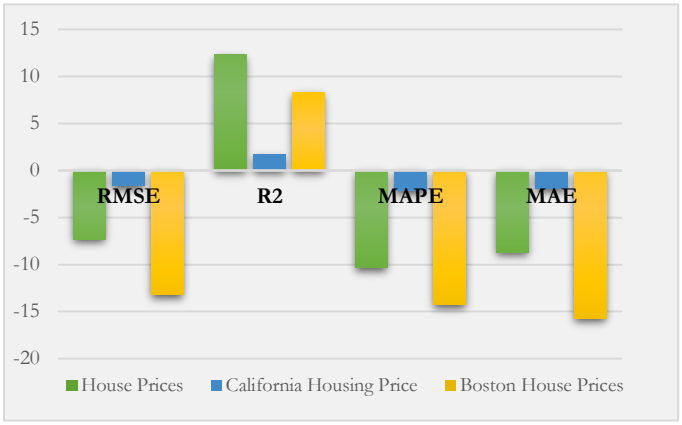


Figure 5. Percentage changes in evaluation metric values.

Examining the most challenging dataset for the proposed algorithm, the California housing price dataset, our initial observation reveals a small degree of linearity through an R^2-score of 0.665 from the baseline model. This first indicates a more complex underlying pattern on this dataset. Also, the optimal value of num_knn is extremely high, up to 59, reflecting that many data points are proximate. These issues lead to difficulty partitioning data into multiple neighbors and constructing best-fitting linear regression lines. Consequently, the proposed algorithm demonstrates its lowest effectiveness on this dataset. These findings show that this dataset would be better suited for advanced models like gradient-boosted trees or deep learning, even though the proposed algorithm can be deployed to enhance performance with modest improvements. From the results, it is evident that the proposed algorithm can significantly enhance linear regression. This improvement is particularly beneficial for fields that highly rely on linear regression, such as the Capital Asset Pricing Model (CAPM) in finance [6], or those who value linear regression's simplicity and interpretability. Yet, in some practical applications where precision is critical, it is advisable to consider other advanced regression models—such as support vector machines, gradient boosting trees, neural networks, and the proposed algorithm—to ensure that the most accurate and practical model is selected for real-world scenarios.

5. Conclusions

The study thoroughly evaluated the algorithmic approach to address the technical problem of linear regression's limitations when applied to non-linear data using three distinct housing price datasets: House Prices, California Housing Prices, and Boston House Prices. The results demonstrate that the proposed algorithm significantly improves the performance of the traditional linear regression model by decreasing RMSE, MAPE, and MAE metrics while increasing the R^2 score across all the datasets. The success of our proposed algorithm would be due to its capacity to break down complex datasets into smaller and manageable neighbors and then construct individual linear regression models for each of them. This approach allows us to capture finer patterns and relationships within the data, resulting in more accurate predictions with the support of KNN. The utility of applying the proposed algorithm improves the performance of linear regression models with a mathematically effortless approach. Still, it also contributes to the model selection

process by indicating the suitability of inviting more complex models like deep learning in appropriate scenarios. Moreover, AI developers could gain advantages from the proposed algorithm by freely substituting their preferred model for linear regression. This might be another approach to enhance the performance of any existing reliable models.

6. Acknowledgements

This work is well supported by Department of Mathematics, Faculty of Science, Kasetsart University, for providing the facilities to conduct the research.

Author Contributions: Conceptualization, methodology, investigation, formal analysis, writing—original draft preparation, writing—review and editing, C.C.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Moore, A.; Bell, M. XGBoost, A novel explainable AI technique, in the prediction of myocardial infarction: A UK Biobank cohort study. *Clinical Medicine Insights: Cardiology*, **2022**, *16*, 1-6. https://doi.org/10.1177/11795468221133611

[2] Khan, M. S.; Salsabil, N.; Alam, M. G. R., et al. CNN-XGBoost fusion-based affective state recognition using EEG spectrogram image analysis. *Scientific Reports* **2002**, *12*, 14122. https://doi.org/10.1038/s41598-022-18257-x

[3] Moraru, L.; Sistaninejhad, B.; Rasi, H.; Nayeri, P. A Review Paper about Deep Learning for Medical Image Analysis. *Computational and Mathematical Methods in Medicine*, **2023**, https://doi.org/10.1155/2023/7091301

[4] Shorten, C.; Khoshgoftaar, T. M.; Furht, B. Text Data Augmentation for Deep Learning. *Journal of Big Data*, **2021**, *8*, 101-135. https://doi.org/10.1186/s40537-021-00492-0

[5] Dodge, J.; Prewitt, T.; Combes, R. T. D., et al. Measuring the carbon intensity of AI in cloud instances. In 2022 ACM Conference on Fairness, Accountability, and Transparency, **2022**, 1877-1894. https://doi.org/10.1145/3531146.3533234

[6] Fama, E. F.; French, K. R. The Capital Asset Pricing Model: Theory and Evidence. *Journal of Economic Perspectives*, **2004**, *18*(3), 25-46. https://doi.org/10.1257/0895330042162430

[7] Singh, T.; Kalra, R.; Mishra, S., et al. An efficient real-time stock prediction exploiting incremental learning and deep learning. *Evolving Systems*, **2022**, *14*, 919-937. https://doi.org/10.1007/s12530-022-09481-x

[8] Walberg, H. J.; Rasher, S. P. Improving Regression Models. *Journal of Educational Statistics*, **1976**, *1*, 253-277. https://doi.org/10.2307/1164786

[9] Uddin, M. F.; Lee, J.; Rizvi, S., et al. Proposing Enhanced Feature Engineering and a Selection Model for Machine Learning Processes. *Applied Sciences*, **2018**, *8*(4). https://doi.org/10.3390/app8040646

[10] Qiao, Z.; Wang, C. H.; Liu, J. Y. Integrating Feature Engineering with Deep Learning to Conduct Diagnostic and Predictive Analytics for Turbofan Engines. *Mathematical Problems in Engineering*, **2022**. https://doi.org/10.1155/2022/9930176

[11] Deisenroth, M. P.; Faisal, A. A.; Ong, C. S. Mathematics for Machine Learning. Cambridge, UK: Cambridge University Press. **2020**.

[12] Arashi, M.; Roozbeh, M.; Hamzah, N. A., et al. Ridge regression and its applications in genetic studies. *PLoS ONE*, **2021**, 16. https://doi.org/10.1371/journal.pone.0245376

[13] Michel, V.; Gramfort, A.; Varoquaux, G., et al. Total Variation Regularization Enhances Regression-Based Brain Activity Prediction. *In First Workshop on Brain Decoding: Pattern Recognition Challenges in Neuroimaging*, **2021**, 9-12. https://doi.org/10.1109/WBD.2010.13

[14] Samavat, A.; Khalili, E.; Ayati, B., et al. Deep Learning Model with Adaptive Regularization for EEG-Based Emotion Recognition Using Temporal and Frequency Features. *IEEE Access*, **2022**, *10*, 24520-24527. https://doi.org/10.1109/ACCESS.2022.3155647

[15] Marsland, S. Machine Learning: An Algorithmic Perspective, 2nd ed. New York: Taylor & Francis Group, **2014**, 158-160.
[16] Dunn, P. K.; Smyth, G. K. Generalized linear models with examples in R. New York: Springer, **2018**. 211-233.
[17] Scikit-learn. (n.d.). Scikit-learn: Machine Learning in Python. Available: https://scikit-learn.org/stable/ [Accessed: 15 Jul 2023].
[18] Kaggle. (n.d.). House Prices: Advanced Regression Techniques. Available: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data [Accessed: 1 Jul 2023].
[19] Kaggle. (n.d.). California Housing Prices. Available: https://www.kaggle.com/datasets/camnugent/california-housing-prices [Accessed: 1 Jul 2023].
[20] Kaggle. (n.d.). Boston House Prices. Available: https://www.kaggle.com/datasets/vikrishnan/boston-house-prices [Accessed: 1 Jul 2023].
[21] Wang, H. Q.; Liang, L. Q. How Do Housing Prices Affect Residents' Health? New Evidence From China. Frontiers in Public Health, **2022**, *9*, 816372. https://doi.org/10.3389/fpubh.2021.816372
[22] Mukhlishin, M. F.; Saputra, R.; Wibowo, A., et al. Predicting house sale price using fuzzy logic, Artificial Neural Network and K-Nearest Neighbour. In Proceedings of the 2017 1st International Conference on Informatics and Computational Sciences, **2017**, 171-176. https://doi.org/10.1109/ICICOS.2017.8276357
[23] Phan, T. D. Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. In Proceedings of the 2018 International Conference on Machine Learning and Data Engineering, **2018**, 35-42. https://doi.org/10.3390/land11112100
[24] Truong, Q.; Nguyen, M.; Dang, H., et al. Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science*, **2020**, 433-442. https://doi.org/10.1016/j.procs.2020.06.111
[25] Gupta, P.; Zhang, Q. Housing Price Prediction Based on Multiple Linear Regression. Scientific Programming, **2021**. https://doi.org/10.1155/2021/7678931
[26] Tanamal, R.; Minoque, N.; Wiradinata, T., et al. House Price Prediction Model Using Random Forest in Surabaya City. *TEM Journal*, **2023**, *12*, 126-132. https://doi.org/10.18421/TEM121-17
[27] Goodfellow, I. J.; Bengio, Y.; Courville, A. Machine Learning Basics. In Deep Learning. Cambridge: MIT Press, **2016**, 96-146.

ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

Product of Hollow Concrete Blocks Mixed with Rice Husk Ash and Cassava Fermentation Waste

Prachoom Khamput[1], Chookiat Choosakul[2*], Tawich Klathae[3], Suporn Rittipuakdee[4], and Sunun Monkeaw[5]

[1] Faculty of Engineering, Rajamangala University of Technology Thanyaburi, Pathum Thani, 12110, Thailand;
[2] College of Industrial Technology and Management, Rajamanagala University of Technology Srivijaya, Nokhon Si Thammarat, 80210,
[3] College of Industrial Technology and Management, Rajamanagala University of Technology Srivijaya, Nokhon Si Thammarat, 80210, Thailand
[4] College of Industrial Technology and Management, Rajamanagala University of Technology Srivijaya, Nokhon Si Thammarat, 80210, Thailand
[5] Faculty of Engineering, Rajamanagala University of Technology Phra Nakhon, Bangkok, 10300, Thailand
* Correspondence: chookiat.c@rmutsv.ac.th

Citation:
Khamput, P.; Choosakul, C.; Klathae, T.; Rittipuakdee, S.; Monkeaw, S. Product of hollow concrete blocks mixed with rice husk ash and cassava fermentation waste *ASEAN J. Sci. Tech. Report.* **2024**, *27*(5), e253838. https://doi.org/10.55164/ajstr.v27i5.253838.

Article history:
Received: April 26, 2024
Revised: August 16, 2024
Accepted: August 17, 2024
Available online: September 7, 2024

Abstract: This research focuses on enhancing the properties of hollow concrete blocks by incorporating rice husk ash and cassava fermentation waste. The study replaces 15% of the cement with rice husk ash from a Bag Filter source and substitutes 0.5-10% of the stone dust by weight with cassava fermentation waste from Ajinomoto (Thailand) Co., Ltd. The hollow concrete blocks were formed using a cement-to-stone dust ratio of 1:10 and a water-to-binder ratio (W/B) of 0.625, with 7 x 19 x 39 cm dimensions. The test results indicate that the density of the hollow concrete blocks decreases with the addition of rice husk ash compared to control hollow concrete blocks. However, when a small amount of cassava fermentation waste is added, the weight and density of the hollow concrete blocks increase. Additionally, the moisture content of the rice husk ash mixed blocks is lower than that of the control hollow concrete blocks. Still, it increases proportionally with the inclusion of cassava fermentation waste. Combining rice husk ash and cassava fermentation waste leads to higher water absorption values in the hollow concrete blocks. Moreover, the compressive strength of the rice husk ash mixed hollow concrete blocks is greater than that of the control hollow concrete blocks. However, the addition of cassava fermentation waste reduces compressive strength.

Keywords: Cassava Fermentation; Cassava pulp; Hollow Concrete Blocks; Rice Husk;

1. Introduction

Hollow concrete blocks, often referred to in the market as "brick blocks," come in both load-bearing and non-load-bearing varieties. They are typically hollow in appearance and are popular due to their affordability and availability. These blocks are advantageous in construction because they allow for quick work and time savings due to their large size. Commercial brick blocks are made from a mixture of Portland cement, crushed stone, sand, and water, which are pressed into block forms using a high-frequency shaker [1]. Standard sizes for hollow concrete blocks include 0.07 x 0.19 x 0.39 meters, 0.09 x 0.19 x 0.39 meters, and 0.14 x 0.19 x 0.39 meters [2]. Cement is a key component in the production of hollow concrete blocks. It is produced by burning limestone at high temperatures of about 1,500 degrees Celsius, generating significant carbon dioxide emissions. Cement production accounts for approximately 8% of global greenhouse gas emissions annually. Reducing these emissions presents a

challenge, prompting a global push to develop technologies that lower carbon output. This includes reducing the use of cement, a major contributor to emissions, and substituting it with alternative materials that provide similar properties. Embracing low-carbon concrete or alternative materials to replace traditional concrete is essential for achieving net-zero carbon emissions by 2050 [3].

Rice husk ash is a byproduct of paddy milling and the combustion of rice husks for energy, producing approximately 4.6 million tons annually [4]. This low-density material offers excellent heat insulation and reduces weight while enhancing properties when incorporated into building materials. Rice husk ash is a pozzolanic material, which, when mixed with cement, can increase the compressive strength of cement or concrete by more than 10% [5]. In hollow load-bearing concrete masonry, used for constructing exterior and interior walls and various structural applications, rice husk ash is particularly beneficial. [6]. According to a study by Kinkachon et al. [7]. rice husk ash can be used as a raw material for producing lightweight bricks, with a recommended proportion of 10% by weight. Additionally, using rice husk ash in concrete mixtures at up to 25% can significantly delay the onset of reinforcement corrosion within the concrete [8].

Tapioca starch production generates byproducts such as cassava peel and cassava pulp. Cassava pulp constitutes up to 10% of the fresh cassava tubers used in production [9]. This pulp has a high moisture content of about 70-80% by weight, making it difficult to utilize and an excellent food source for microorganisms. If left untreated, it can lead to microbial degradation in the environment, causing unpleasant odors that disturb surrounding communities [10]. Research by Boontositrakul, Suweero, and Weeranukul [11] found that replacing stone dust with 3% by weight of cassava tree chips in rice husk ash hollow concrete blocks can reduce the density or weight of the blocks and improve their heat insulation properties.

This research investigates the potential of using rice husk ash, sourced from the bag filter area, as a substitute for Portland cement and cassava starch fermentation residue to replace stone dust in producing hollow concrete blocks. This approach seeks to enhance the value of these waste materials by incorporating them into construction products, thereby reducing environmental pollution and lowering the costs associated with hollow concrete block production. Additionally, this strategy aims to develop new, cost-effective construction materials that can be effectively utilized in various building projects.

2. Materials and Methods

2.1 Research materials

This research utilizes Portland cement Type 1, which meets the standards of ASTM C-150 Type 1 [12]. Stone dust is sun-dried and then sifted through a No. 4 sieve. It has a specific gravity of 2.62, consistent with the research results of [13] as shown in Figure 1. It is used as a coarse aggregate. Rice husk ash (RHA), sourced from the bag filter area of Ajinomoto (Thailand) Co., Ltd., is sieved through No. 325 according to ASTM E11 [14], with a sieve opening size of 45 micrometers. This ash has a specific gravity of 2.07, which is consistent with the research results of [15], as determined by ASTM C188 [16] tests (see Figure 2), and is used to replace Portland cement Type 1 partially. The cassava fermentation waste (CFW), also obtained from Ajinomoto (Thailand) Co., Ltd. and shown in Figure 3, is used as an aggregate to replace stone dust partially. Tap water is used to facilitate the binder's reaction.

Figure 1. Stone Dust.

Figure 2. Rice husk ash.

Figure 3. Cassava fermentation waste.

2.2 Mixture design and sample preparation

The ratio for hollow concrete blocks mixed with rice husk ash is adopted from previous studies [11], using a cement-to-stone dust ratio of 1:10. A water-to-binder ratio (W/B) of 0.625 is used to develop hollow concrete blocks mixed with rice husk ash and cassava fermentation waste. Cement is partially replaced with rice husk ash from bag filter sources, and stone dust is partially replaced with cassava fermentation waste. The main ratio selected for mixing involves replacing 15% of the cement with rice husk ash (RHA15), as this ratio yields the highest compressive strength. Cassava fermentation waste is used to replace 0.5% to 10% of the stone dust by weight (CFW05-CFW10) to compare the properties of the developed hollow concrete blocks with control hollow concrete blocks. A total of 13 mixture ratios are used for forming hollow concrete blocks, as shown in Table 1.

Table 1 Hollow concrete blocks mixture rate by weight.

Mix	Binder		Aggregates		W/B
	Cement	Rice husk ash	Stone dust	CFW	
Control	1.00	0.00	10.00	0.00	0.625
RHA15	0.85	0.15	10.00	0.00	0.625
CFW05	0.85	0.15	9.95	0.05	0.625
CFW1	0.85	0.15	9.90	0.10	0.625
CFW2	0.85	0.15	9.80	0.20	0.625
CFW3	0.85	0.15	9.70	0.30	0.625
CFW4	0.85	0.15	9.60	0.40	0.625

Table 1. Hollow concrete blocks mixture rate by weight. (Continue)

Mix	Binder		Aggregates		W/B
	Cement	Rice husk ash	Stone dust	CFW	
CFW5	0.85	0.15	9.50	0.50	0.625
CFW6	0.85	0.15	9.40	0.60	0.625
CFW7	0.85	0.15	9.30	0.70	0.625
CFW8	0.85	0.15	9.20	0.80	0.625
CFW9	0.85	0.15	9.10	0.90	0.625
CFW10	0.85	0.15	9.00	1.00	0.625

Hollow concrete blocks mixed with rice husk ash are formed to a size of 0.07 x 0.19 x 0.39 meters using the designed ratios and a vibrating press, as shown in Figure 4. As depicted in Figure 5, the blocks are then cured or dried in a shaded and well-ventilated area for the specified test period.

Figure 4. Vibrating hollow concrete blocks compactor.

Figure 5. Hollow concrete blocks ready to be cured.

2.3 Test Method

1. The density of hollow concrete blocks is tested according to ASTM C140 standards [17] after a curing period of 28 days.

2. The moisture content and water absorption of hollow concrete blocks are tested according to ASTM C426 standards [18] after a curing period of 28 days.

3. The compressive strength of hollow concrete blocks is tested according to ASTM C140 standards [17] at the curing ages of 7, 14, 21, and 28 days, as shown in Figure 6.

Figure 6. Compressive Strength Test of Hollow concrete blocks.

4. The practical use of hollow concrete blocks is tested by constructing walls and plastering them with mortar. This process is used to evaluate the durability and characteristics of the plastered wall surface.

3. Results and Discussion

3.1 Test results of weight per lump and density of hollow concrete blocks

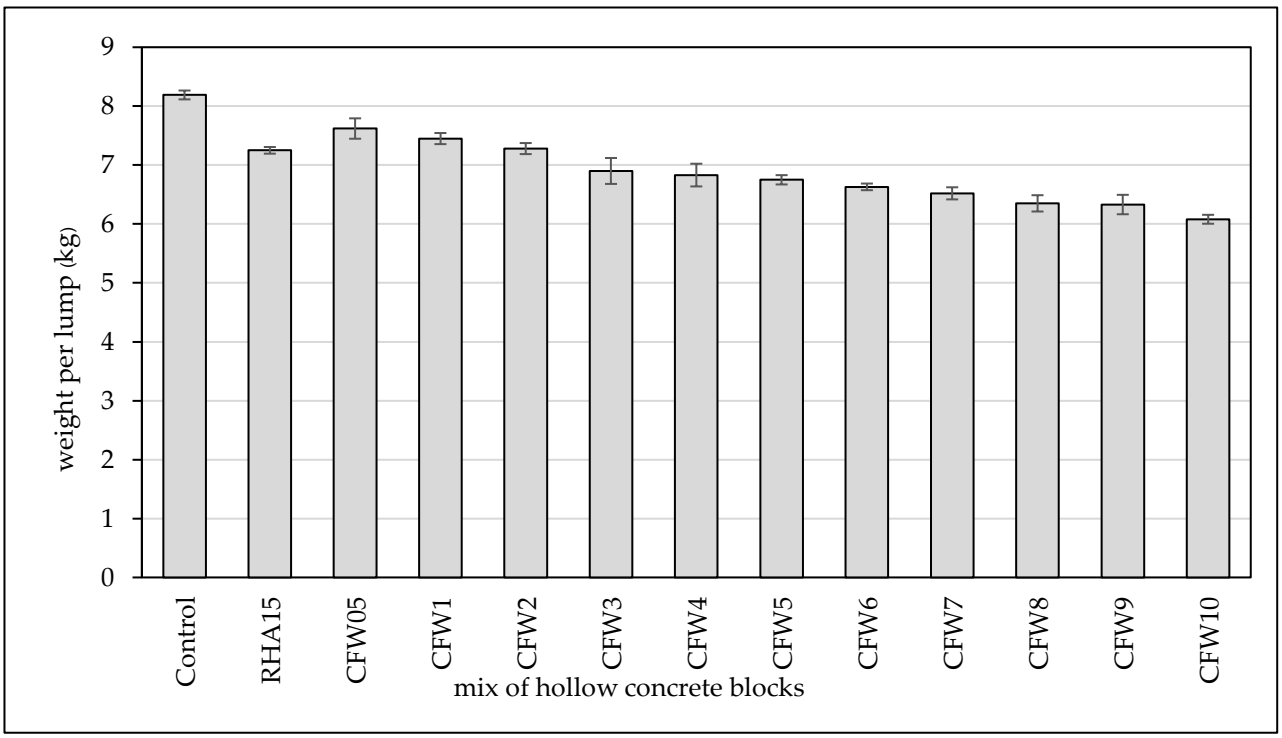


Figure 7. Weight per lump of hollow concrete blocks at the age of 28 days.

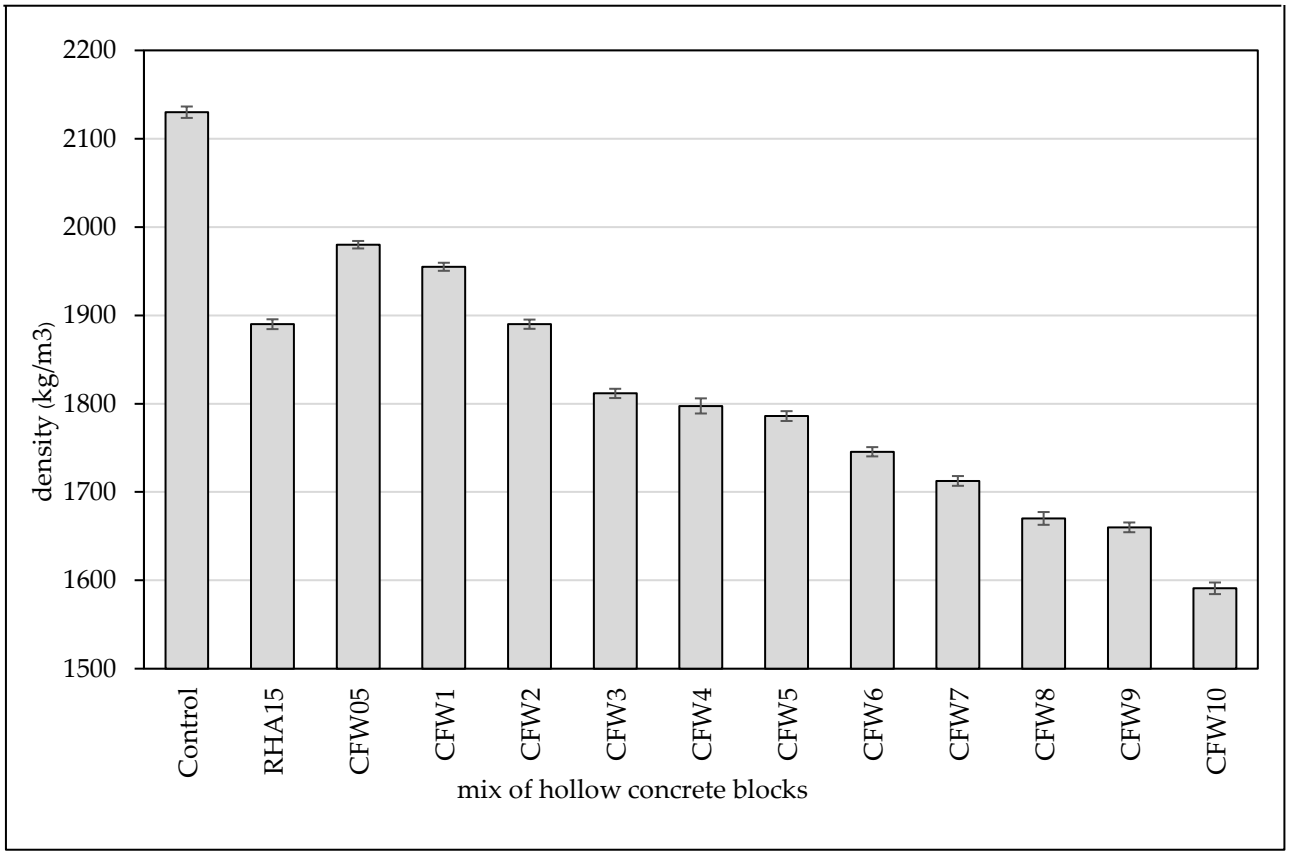


Figure 8. The density of hollow concrete blocks at the age of 28 days.

Figures 7 and 8 show that incorporating cassava fermentation waste into hollow concrete blocks mixed with fine-sized rice husk ash (from the Bag Filter source) affects both the weight and density of the blocks. The weight and density per unit increase when a small amount of cassava fermentation waste is added to the rice husk ash concrete blocks. However, when larger quantities of cassava fermentation waste are used, the weight and density decrease significantly, resulting in lower values than concrete blocks mixed only with rice husk ash.

Comparing these blocks to conventional hollow concrete blocks, rice husk ash and rice husk ash with cassava fermentation waste mixtures exhibit lower weight and density per unit. This is because conventional hollow concrete blocks typically use stone dust or limestone chips with a density of 2,611 kg/m^3 [19]. In contrast, rice husk ash has a bulk density ranging from 540–860 kg/m^3 and a compact density ranging from 1,120–1,500 kg/m^3 [5, 20]. As a result, hollow concrete blocks made with rice husk ash have lower weight and density than conventional blocks. Most conventional hollow concrete blocks and those made with rice husk ash have larger aggregate particles (larger than a No. 4 sieve), resulting in a porous texture. When cassava fermentation waste smaller than a No. 4 sieve is added in the right amount, it contributes to a denser texture, increasing the concrete block's density [21]. However, excessive cassava fermentation waste can negatively impact the blocks' binder content, alignment, and adhesion, leading to reduced weight and density. The effects of mixing new versus old tapioca starch into rice husk ash concrete blocks are similar in weight and density. According to ASTM C129 [22], the density classification of concrete blocks shows that the control concrete blocks are categorized as normal weight. Concrete blocks mixed with rice husk ash (RHA15) and those mixed with rice husk ash and cassava fermentation waste (CFW05-CFW9) are classified as medium weight. Meanwhile, concrete blocks that include 10% cassava fermentation waste (CFW10) are classified as lightweight.

3.2 Test results of moisture content and water absorption of hollow concrete blocks

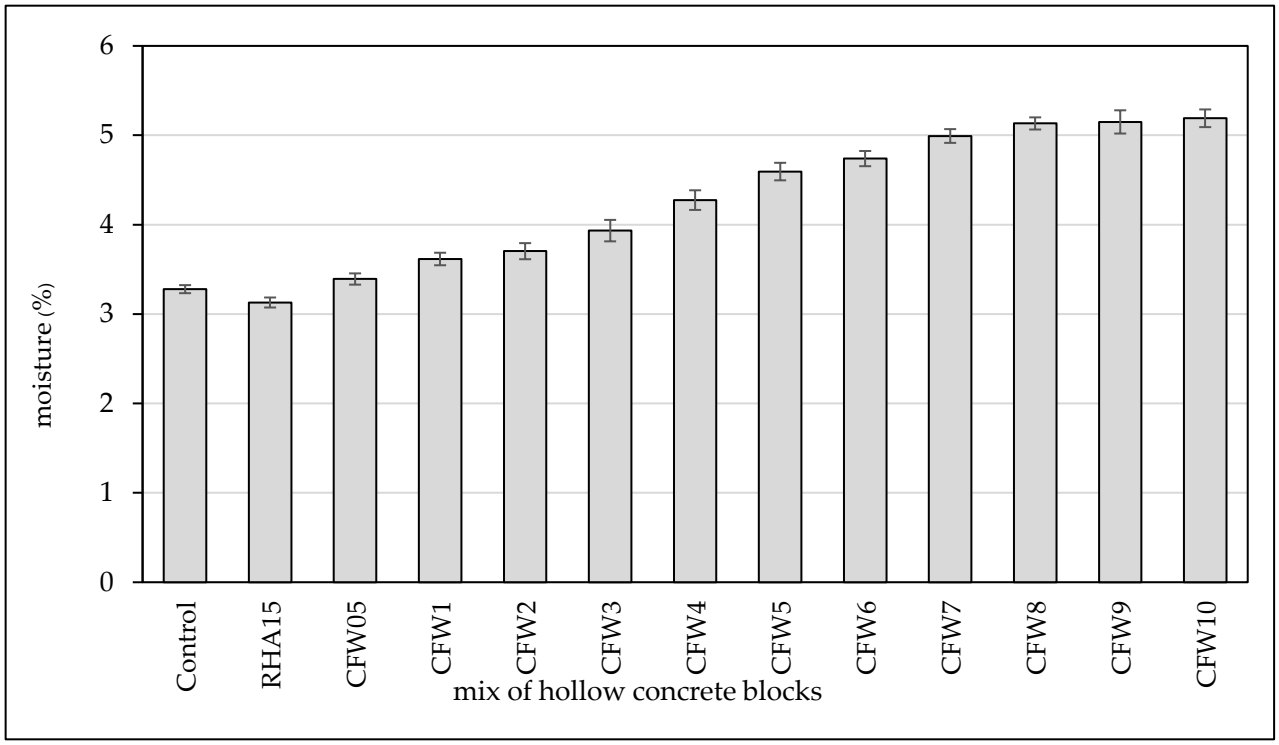


Figure 9. The moisture content of hollow concrete blocks at the age of 28 days.

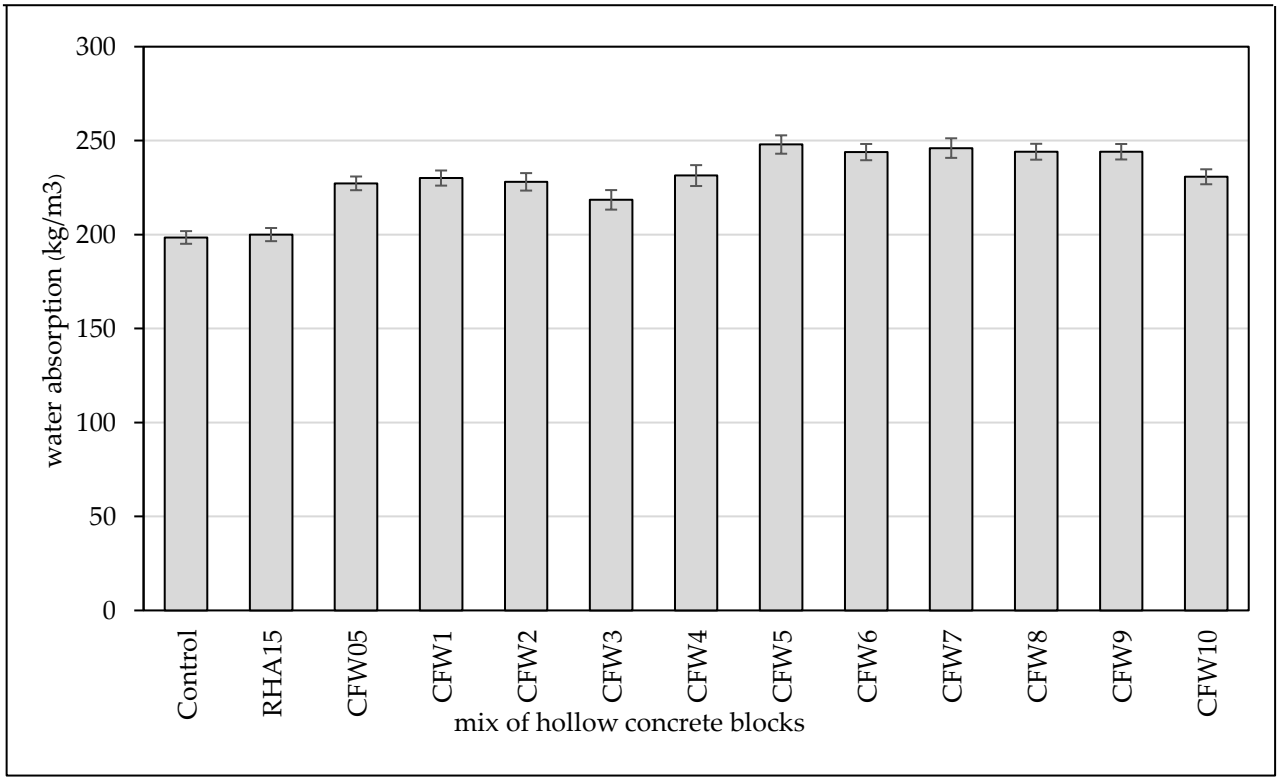


Figure 10. Water absorption of hollow concrete blocks at the age of 28 days.

Figures 9 and 10 reveal that hollow concrete blocks mixed with rice husk ash have lower moisture content but higher water absorption values than conventional hollow concrete blocks. When a small amount

of cassava fermentation waste is added to these rice husk ash blocks, the moisture content increases, and water absorption decreases. However, increasing the amount of cassava fermentation waste in the hollow concrete blocks results in higher moisture content and water absorption values. This is because the texture of rice husk ash hollow concrete blocks varies with different amounts of cassava fermentation waste; dense concrete has lower water absorption, while porous concrete has higher water absorption [20]. The water absorption value is a property of hollow concrete blocks that indicates their plastering capacity. Hollow concrete blocks suitable for construction should have low water absorption to minimize cracking in plastered walls. Blocks with high water absorption can cause the mortar (cement, sand, and water) to lose moisture, leading to an incomplete hydration reaction between cement and water. This can result in small, invisible cracks that may develop into larger ones. When comparing the dry shrinkage value to the ASTM C129 [22] standard for non-load-bearing concrete blocks, it was found to meet the specification that the total linear drying shrinkage should not exceed 0.065%. According to the TIS 57-2017 [23] standard for water seepage, which varies based on the type and density of concrete blocks, it was found that the control concrete block, as well as those mixed with rice husk ash and 0.5-4% cassava fermentation waste, met the standard criteria. However, concrete blocks mixed with rice husk ash and 5-10% cassava fermentation waste did not meet the criteria.

3.3 Hollow Concrete Blocks Compressive Strength Test Results

Figure 11 shows that hollow concrete blocks mixed with rice husk ash have significantly higher compressive strength than conventional hollow concrete blocks. This increase in strength is due to the pozzolanic reaction, where the silica in the rice husk ash reacts with calcium hydroxide ($Ca(OH)_2$) produced during the hydration of cement and water [5, 24].

$$2SiO_2 + 3Ca(OH)_2 \xrightarrow{H_2O} 3CaO.2SiO_2.3H_2O \quad (1)$$

Calcium silicate hydrate ($3CaO{\cdot}2SiO_2{\cdot}3H_2O$ or C-S-H) is crucial for strengthening concrete [5, 24-25]. Consequently, hollow concrete blocks mixed with rice husk ash exhibit higher compressive strength than conventional blocks. The comparison results indicate that rice husk ash hollow concrete blocks (RHA15) achieve the highest compressive strength, followed by those mixed with small amounts of cassava fermentation waste (CFW05). The compressive strength gradually decreases as the amount of cassava fermentation waste increases. Previous research has shown that sugar content can slow down the setting process of concrete and reduce its compressive strength [26-29]. The equation is as follows:

$$\underset{\text{(tapioca)}}{(C_6H_{12}O_5)n} + nH_2O \xrightarrow{\text{enzyme}} \underset{\text{(sugar)}}{nC_6H_{12}O_6} \xrightarrow{\text{yeast}} \underset{\text{(ethanol)}}{2nC_2H_5OH} + 2nCO_2 \quad (2)$$

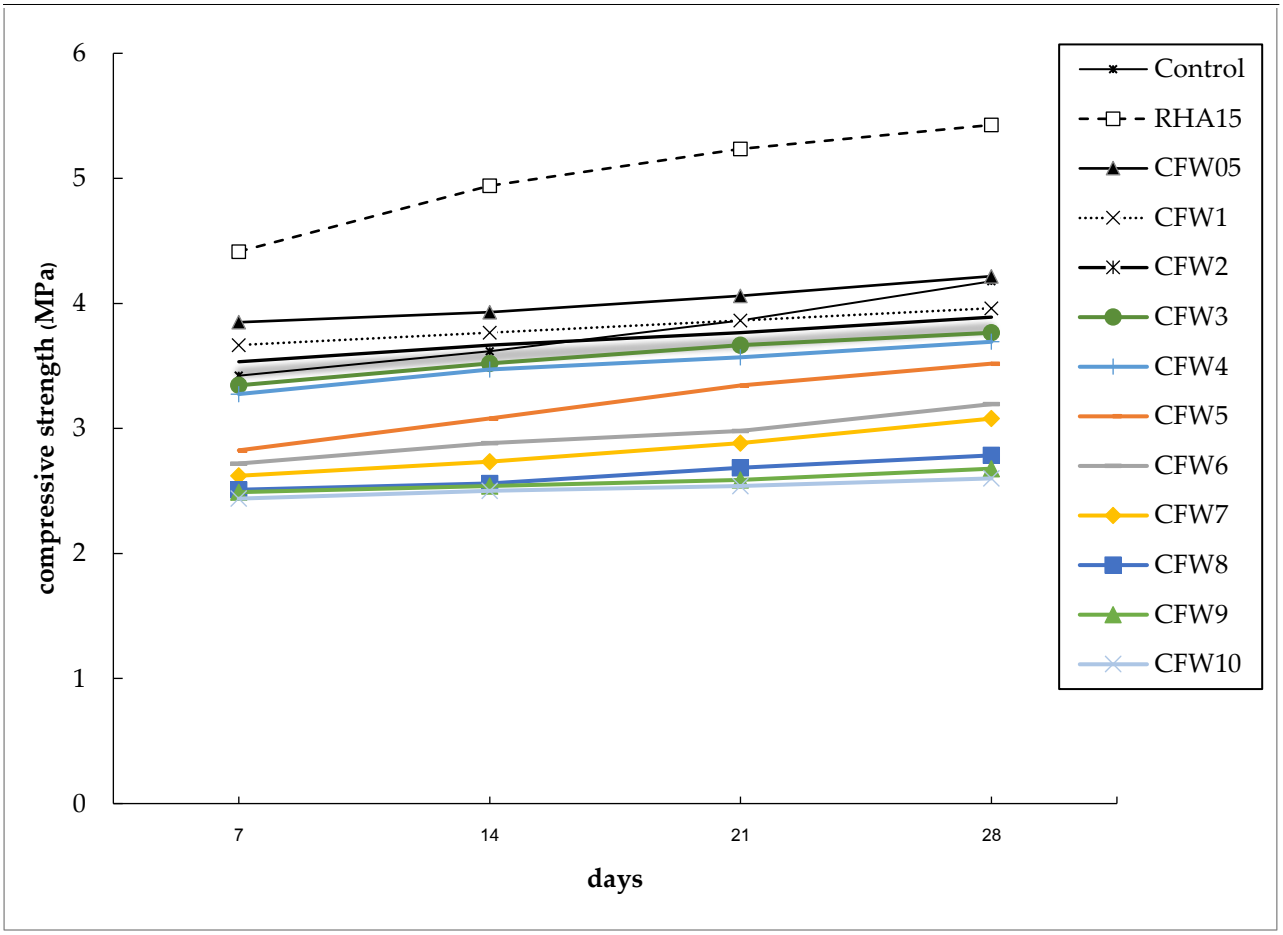


Figure 11. Compressive strength of mixed hollow concrete blocks.

A comparison of the compressive strength of hollow concrete blocks mixed with cassava fermentation waste against the ASTM C129 [22] standard for non-load-bearing concrete masonry units—which requires an average compressive strength of at least 4.14 megapascals for five blocks—revealed that only the rice husk ash hollow concrete blocks without cassava fermentation waste met this criterion. However, adding small amounts of cassava fermentation waste to hollow concrete blocks resulted in compressive strength similar to conventional blocks. Although the compressive strength of all tested hollow concrete blocks was lower than the ASTM C129 [22] requirement, these blocks can still be used in general applications. The ASTM C129 [22] standard does not mandate that all hollow concrete block products meet the criteria. Furthermore, the typical use of hollow concrete blocks in wall construction does not require them to withstand compressive loads greater than those of the walls built with them. Therefore, if rice husk ash hollow concrete blocks mixed with cassava fermentation waste are to be used, it is recommended to select blocks with a compressive strength of at least 4.14 megapascals, such as those with small amounts of cassava fermentation waste in the CFW05 ratio. ASTM C129 [22] stipulates an average compressive strength value for five blocks, with no individual block having a compressive strength lower than 3.45 megapascals [30]. This indicates the minimum compressive strength required for practical use as a building wall, ensuring that the compressive strength of all blocks is above the standard, even if the average exceeds it.

3.4 Practical Test Results of Hollow Concrete Blocks

When rice husk ash hollow concrete blocks mixed with cassava fermentation waste in the CFW05 ratio were tested for wall construction and plastered with mortar, as shown in Figures 12 and 13, it was concluded that these blocks perform similarly to ordinary hollow concrete blocks in wall construction.

Figure 12. Testing the use of hollow concrete blocks by wall formation.

Figure 13. Hollow concrete blocks walls that have already been plastered.

3.5 Cost calculation results of hollow concrete blocks

The cost of hollow concrete blocks mixed with rice husk ash and cassava fermentation waste can be determined by accounting for the costs of raw materials, labor, electricity, management, and depreciation of machinery and equipment. However, the preliminary cost calculation considers only the cost of raw materials. This calculation is based on the weight per block at 28 days of curing, with rice husk ash and cassava fermentation waste considered costless raw materials. The results of this calculation are summarized in Table 2.

From the raw material costs detailed in Table 2, it is evident that hollow concrete blocks mixed with rice husk ash and cassava fermentation waste are less expensive than conventional hollow concrete blocks, with savings ranging from $0.013 to $0.030 per block (compared to $0.085 per conventional block). Given that these blocks meet the ASTM C129 [22] standards, the CFW05 ratio, which includes the highest amount of cassava fermentation waste, offers substantial cost reductions. The raw material cost for these blocks is $0.072 per block, representing a reduction of $0.013, or 15.29%, compared to the cost of raw materials for standard hollow concrete blocks. This preliminary cost estimate is based on the weight per block and mixture ratios and may differ from actual production costs. Therefore, checking the standard prices of raw materials in different regions for more precise cost assessments is advisable.

Table 2. Cost of raw material cost of rice husk ash hollow concrete blocks mixed with cassava fermentation waste.

Mix	cost (Dollar)	Different cost (Dollar)	Different cost (%)
Control	0.085	0.00	0
RHA15	0.069	-0.016	-18.82
CFW05	0.072	-0.013	-15.29
CFW1	0.070	-0.015	-17.65
CFW2	0.068	-0.017	-20.00
CFW3	0.064	-0.021	-24.71
CFW4	0.063	-0.022	-25.88
CFW5	0.063	-0.022	-25.88
CFW6	0.061	-0.024	-28.24
CFW7	0.060	-0.025	-29.41
CFW8	0.058	-0.027	-31.76
CFW9	0.057	-0.028	-32.94
CFW10	0.055	-0.03	-35.29

Remark Cement price from building materials store, Pathum Thani
Price of stone dust from stone mill Sila Theptawan, Saraburi
Water supply price from Provincial Waterworks Authority, Pathum Thani

4. Conclusions

Developing hollow concrete blocks mixed with rice husk ash and cassava fermentation waste revealed that cassava fermentation waste is unsuitable for inclusion in hollow concrete blocks or other cement-based products. This is due to the presence of sugary residues from the enzymatic reaction of tapioca starch, which prevents the concrete from hardening and reduces its compressive strength. Among the various tested ratios, the CFW05 ratio, which includes the highest amount of cassava fermentation waste, showed potential for practical application. In this ratio, the hollow concrete blocks weighed 7.62 kilograms, with a density of 1,980 kilograms per cubic meter, a moisture content of 3.39 percent, water absorption of 227 kilograms per cubic meter, and a compressive strength of 4.20 megapascals. Although this ratio meets the ASTM C129 [22] standard for non-load-bearing concrete masonry units, it only meets individual blocks' compressive strength requirements. It does not meet the average compressive strength requirement of 4.14 megapascals for five blocks. Nonetheless, this ratio is still suitable for use in building wall construction.

5. Acknowledgements

We want to thank Ajinomoto (Thailand) Co., Ltd.

Author Contributions: Conceptualization, P.K; methodology, P.K., T.K., S.R.; software, T.K., S.M.; validation, S.R., S.M.; data curation, P.K., C.C.; writing—original draft preparation, P.K., C.C.; writing—review and editing, P.K., C.C.; visualization, P.K., C.C.; supervision, P.K.. All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] iEnergy GURU. Concrete masonry unit. **2015**. Available online: https://ienergyguru.com/2015/09/concrete-masonry-unit/. (8 November 2020).

[2] Lertwattanaruk, P.; Suntijitto, A. Properties of Natural fiber cement materials containing coconut coir and oil palm fibers for residential building applications. *Construction and Building Materials* **2015**, *94*, 664-669.

[3] CarbonCure. Innovative CO_2 Technologies. **2020**. Available online: https://www.carboncure.com /technologies/. (12 November 2020).

[4] Weeranukul, P.; Suweero, K. Development of cement boards from coconut shell ash for energy and environment conservation. *Engineering and Applied Science Research* **2016**, *43*, 173-175. (in Thai)

[5] Jindaprasert, P.; Jaturapitakkul, C. *Cement pozzolan and concrete*. 7th ed; ACI Partners with Thailand Concrete Association, Bangkok, **2013**. (in Thai)

[6] Hollow load-bearing concrete masonry units. Thai Industrial Standards Institute, TIS. 57-2017, Ministry of Industry, Bangkok, **2017**. (In Thai)

[7] Kinkachon, T.; Kunlawong, S.; Pattum, J.; Chaikhan, S.; Thongdamrongtham, S.; Chareerat, T. Effect of rice husk ash used as pozzolan material on properties of lightweight bricks. *Journal of Science and Science Education (JSSE)* **2022**, *5(2)*, 182-190.

[8] Suwanmaneechot, P.; Chalee, W. Time to Initial Corrosion of Steel Reinforcement in Concrete Containing Rice Husk-Bark Ash under Marine Environment, Proceedings of 19th National Convention on Civil Engineering, Khonkaen, Thailand, 14-16 May 2014, pp. 830-836. (in Thai)

[9] Ditkunchaimongkol, N. The study of efficiency and cost of cassava pulp hydrolysis by acid and enzymes. *Report of academic conferences and presentations of research results national and international national group science* **2015**, *1*(6), 249-259.

[10] Duangsrisen, W.; Treeamnuk, T.; Sukthang, N.; Arjharn, W. A Study of Drying Cassava Pulp Using a Rotary Screen Dryer. *Thai Society of Agricultural Engineering Journal* **2012**, *19*(1), 7-13. (in Thai)

[11] Boontositrakul, K.; Suweero, K.; Weeranukul, P. Using Cassava Pit Waste as Light Weight Aggregate for Hollow Load Bearing Concrete Masonry Mixed with Rice Husk Ash Product. *Journal of Engineering, RMUTT* **2020**, *18*(1), 13-22. (in Thai)

[12] Standard Specification for Portland Cement. Annual Book of ASTM Standards, ASTM C150, **2015**, 04.02.

[13] Dar, N. A.; Bhalla, D. G. Stabilization of soil using jute fiber and Stone dust. *International Journal of Scientific Development and Research* **2020**, *5*(8), 325-333.

[14] Standard Specification for Woven Wire Test Sieve Cloth and Test Sieves. Annual Book of ASTM Standards, ASTM E11, **2022**, 14.02.

[15] Rattanachu, P.; Toolkasikorn, P.; Tangchirapat, W.; Chindaprasirt, P.; Jaturapitakkul, C. (2020). Performance of recycled aggregate concrete with rice husk ash as cement binder. *Cement and Concrete Composites* **2020**, *108*, 103533.

[16] Standard Test Method for Density of Hydraulic Cement. Annual Book of ASTM Standards, ASTM C188, **2016**, 04.02.

[17] Standard Test Methods for Sampling and Testing Concrete Masonry Units and Related Units. Annual Book of ASTM Standards, ASTM C140, **2005**, 04.02.

[18] Standard Test Method for Linear Drying Shrinkage of Concrete Masonry Units. Annual Book of ASTM Standards, ASTM C426, **2017**, 04.02.

[19] Tonnayopas, D. *Minerals and rocks*, 2nd ed; Faculty of Engineering, Prince of Songkhla University, Songkhla, **2010**. (in Thai)

[20] Mehta, P.K.; Monteiro, P.J.M. *Concrete Microstructure, Properties and Materials*. 3rd ed; McGraw-Hill, **2006**.

[21] Jaturapitakkul, C.; Tangchirapat, W. *Utilization of ash and industrial waste as material in concrete work*, 2nd ed; Department of Civil Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, Bangkok, **2013**. (in Thai)

[22] Standard Specification for Non-Load Bearing Concrete Masonry Units. Annual Book of ASTM Standards, ASTM C129, **2004**, 04.02.
[23] Hollow load-bearing concrete masonry units. Thai Industrial Standards Institute, TIS. 57-2017, Ministry of Industry, Bangkok, **2017**. (In Thai)
[24] Setthabut. C. *Cement and applications*, Thai Cement Industry Co., Ltd., Bangkok, **2009**. (In Thai)
[25] Neville, A. M. *Properties of Concrete*. London: Pearson Education PLC, **2011**.
[26] Akogu, A. E. Effects of sugar on physical properties of ordinary Portland cement paste and concrete. *AU J.T.* **2011**, *14*(3), 225-228.
[27] Khan, B.; Baradan, B. The effect of sugar on setting-time of various types of cements. *Science Vision* **2002**, *8*(1), 71-78.
[28] Suman, R. Effect of Sugar on Setting-Time and Compressive Strength of Ordinary Portland Cement Paste, Proceedings of 3rd World Conference on Applied Sciences. Engineering & Technology, Kathmandu, Nepal, 27-29 September 2014, 127-129.
[29] Suryawanshi, Y. R.; Bhat, P. N.; Shinde, R. R.; Pawar, S.; Mote, N. Experimental Study on Effect of Sugar powder on Strength of cement. *International Journal of Research in Engineering and Technology* **2014**, 249-252.
[30] Hollow non-load-bearing concrete masonry units. Thai Industrial Standards Institute, TIS. 58-2017, Ministry of Industry, Bangkok, **2017**. (In Thai)

ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

A Study on Using Machine Learning to Predict Winner in Multiplayer Online Battle Arena (MOBA) Game

Nattapat Tangniyom[1] and Pruet Boonma[2*]

[1] Data Science Program, Faculty of Engineering, Chiang Mai University, Chiang Mai, 50200, Thailand
[2] Data Science Program, Faculty of Engineering, Chiang Mai University, Chiang Mai, 50200, Thailand
* Correspondence: pruet.b@cmu.ac.th

Citation:
Tangniyom, N.; Boonma, P. A Study using ensemble learning to predict winner in multiplayer online battle Arena (MOBA) game. *ASEAN J. Sci. Tech. Report.* **2024**, *27*(5), e252289. https://doi.org/10.55164/ajstr.vx27i5.252289.

Article history:
Received: January 4, 2024
Revised: September 5, 2024
Accepted: September 6, 2024
Available online: September 7, 2024

Abstract: Realm of Valor (RoV) is a famous multiplayer online battle arena (MOBA) game. An average of 25 million games are played daily in Thailand alone. The game also competes in international events, with millions of U.S. dollars in the prize pool. However, the game is very complex and requires a player to have high experience to win. In particular, the hero selection process that each user has to perform at the beginning of each game because the set of selected heroes can affect the game outcome, but there are many heroes to be selected. This paper compares machine learning techniques to predict the winner's side based on the player's and opponent's selection of heroes and the relationship among the selected heroes. Three traditional machine learning techniques, namely, k-Nearest Neighbor, Logistic Regression, and Decision Tree, are compared against ensemble learning with their optimized parameters. The algorithms are evaluated using k-fold cross-validation, and the accuracy of each algorithm is measured. The results show that winning predictions can be improved by considering the relationship among selected heroes. Also, ensemble learning can compete with traditional learning.

Keywords: Winner prediction; E-Sport; Machine learning; Ensemble learning

1. Introduction

The game industry has expanded rapidly in recent years, especially the MOBA (Multiplayer online battle arena) genre, where multiple players, divided into two teams, compete with each other on an online game battlefield. The expansion leads to professional game competition events, i.e., e-sports, which have been arranged nationally and internationally. In some countries, one of the most popular e-sport games is Realm of Valor (RoV), also known as Arena of Valor (AoV). This game is a variation of Honor of Kings (HoK), a Chinese online game with some internationalized characters. However, they share the same gameplay and mechanism.

RoV was first released to the general public in 2016 in many Asian countries, such as Indonesia, Vietnam, Taiwan, and Thailand. In Thailand alone, within five years after release, there are 45,961,817,910 gameplays, an average of 25 million games played daily. Moreover, many international RoV e-sport events have been organized. For instance, AIC 2021 (Arena of Valor International Championship 2021) has been organized in many countries, e.g., Vietnam, Taiwan, and Thailand, with a prize pool of $1,000,000. In 2022, AIC increased the prize pool to $2,000,000, a 100% increase from the previous year.

In RoV, a player selects a set of five "heroes" to be used throughout a game from a collection of 109 characters, which are added occasionally by the developers. Then, they form a team of five positions: Support, Carry, Mage, Off-line, and Jungle. Each position has its specific function, and the selected hero will affect that function. Therefore, selecting a hero at the beginning of the game is crucial for the game outcome. However, because (1) there are many characters to select from, (2) the rule inhibits players from selecting the same hero in the subsequent game, and (3) the rule that all players ban the hero selection of the opponent; selecting a set of heroes is a complicated process and require a massive amount of information, for instance, the previous selection of opponent, the outcome, the selectable heroes, among the others. Thus, this selection process generally requires the player's initiative and is unreliable.

In this paper, the selection of heroes is used to predict the winning outcome of a game based on records. This paper considers synergy, i.e., a combination of heroes in the same team, and counter, i.e., a combination of heroes from different teams, in the training process to improve prediction accuracy. The game records used in this research are from official international e-sport events such as AIC 2022 and the 31st Southeast Asian Games e-sport event. The training dataset includes the list of heroes from each player, the order of usage, and the result. This research investigates two training strategies; first, individual machine learning algorithms, i.e., k-Nearest Neighbor (KNN), Logistic Regression (LR), and Decision Tree (DT) are studied. They are selected based on their popularity in current literature. Second, this paper proposes to use ensemble learning to improve prediction accuracy. The model is evaluated for accuracy against the actual results. The structure of the paper is as follows: the second section discusses related works. The studied algorithms and datasets are presented in the third section. The following section shows the evaluation results, and the final section concludes and discusses future works.

2. Related Works

A technical report by Kinkade, Yul, and Lim studied the features set that improves win prediction accuracy [1]. The report uses logistic regression to consider *Synergy* and *Counter* in the past game records. This approach can improve the win prediction accuracy based on hero selection alone, with the highest prediction accuracy of 73%.

Semenov et al. proposed using player skills, e.g., normal, high, and very high skills, to partition past game records [2]. Then, the matching algorithm in Dota2, MMR, matches two players with similar skills to balance the game. To predict the winning side, Naïve Bayes, logistic regression, factorization machines, and gradient boosting of decision trees are used to predict the winning players. The study results show that the prediction of winning results of normal-skill players yields higher accuracy than that of higher-skill players. Moreover, factorization machines show the highest prediction accuracy among all studied algorithms.

Chen et al. proposed using the Monte-Carlo tree search (MCTS) and artificial neural network (ANN) to predict winners based on past games [3]. In the paper, the data consists of human and artificial intelligence (A.I.) best-of-N games with the condition that each hero can be used only once. The evaluation result shows that MCTS that use value networks with long-term value can outperform the MCTS without long-term value, using a random approach, the highest winning approach, and the baseline approach to predict a higher winning rate. On the other hand, artificial neural networks show better performance on winning prediction.

Hodge et al. utilize a gradient boosting machine (GBM), logistic regression (LR), and random forest (R.F.) to predict the winning party of Dota 2 [4]. The games used in this research include both public games and professional tournament games. In the dataset from professional tournament games, the first five minutes of the game screen are recorded as time series for live prediction. The evaluation results show that the algorithms can predict the winner with 74.59% accuracy in professional games and 77.51% in hybrid games (public + tournament). Then, the research implements live predictions, which predict the winner while the game is still in play; the accuracy of this approach is 85% after the game is played for the first five minutes.

Tian et al. proposed a Hero featured Network (HFN) that uses past games from Honor of Kings and considers the *Synergy* and *Restraint* features of the game [5]. HFN considers three important properties, gold, kills, and fortress, to predict the winning side. The research focuses on these two features and when to use

them to predict the game's winning. By considering when to use these features, HFN can outperform LSTM, TSSTN, Transformer, LR, and SVM.S.

Song et al. proposed training logistic regression with past games of Dota2 [6]. This research focuses on combo heroes with an effectiveness of at least 50 combos. Then, the research uses Stepwise Regression for heroes with small past game data. K-fold cross-validation was used to evaluate the model. The result shows that at least 60 features are required to have low prediction error.

Yang et al. studied real-time win prediction of Honor of Kings with a two-stage spatial-temporal network (TSSTN) model [7]. The model considers six features: gold, kill, tower, wild resource, soldier, and heroes. The prediction accuracy shows that TSSTN can outperform the heuristic approach but performs slightly worse than Fully Connected Network (FCN) and LSTM.

Yang et al. used LTSM and Transformer to predict the winning side, who can kill the boss (Tyrant), the hero who will kill next, and the hero who will be killed next [8]. The results show that the prediction accuracy keeps increasing when the game progresses; however, the prediction accuracy reduces at the end of the game. The results also show that, in general, Transformer can outperform LSTM.

McGuire et al. proposed using an artificial neural network (ANN) with three, five, and seven fully connected layers to predict the DotA2 winning side [9]. The ANN is configured with a feed-forward configuration with a non-linear sigmoid function in each layer. Two features are considered: hero selection and hero selection/time. Comparing three configurations, five fully connected layers show the best performance at 0.73, while using just hero selection shows better performance than hero selection/time.

Conley and Perry proposed using logistic regression (LR) and K-nearest neighbors (KNN) to predict the winning side by using past games from DotA2 [10]. This work considers both *Synergy* and *Counter* features. The prediction accuracy of Logistic Regression is about 69.8% when trained with 18,000 games, while K-Nearest Neighbors can archive 70% when trained with 50,000 games.

Kalyanaraman proposed using logistic regression, genetic algorithm, and augmented regression to predict the winning side of a DotA2 match from past games [11]. This work also considers both *synergy* and *countering* features for hero selection. The paper observes that the accuracy of augmented regression is improved when augmented with a genetic algorithm. Comparing the three algorithms, Logistic Regression shows the best performance with 75.2% accuracy, while Genetic Algorithm and Augmented Regression can archive 74.1% and 68.4%, respectively. However, augmented regression has the highest recall performance, at 90.9%.

Do et al. proposed using player-champion experience to predict the winning side in League of Legends (LoL) [12]. The champion in this game is similar to the hero in RoV. The paper employs Support Vector Machine, k-Nearest Neighbors, Random Forrest, Deep Neural Network, and Gradient Boosting as prediction models, and the results show that Gradient Boosting has the highest accuracy at 75.4%, followed closely by Deep Neural Network at 75.1%. However, Gradient Boosting has the highest standard error compared with the others, indicating the algorithm is unstable. Hence, a Deep Neural Network is the best choice because it has high accuracy and is stable.

Sena and Emanuel investigated winning prediction in Mobile Legends by considering average gold, average level, total kills, first blood, first turtle, and first lord, besides the hero selections [13]. The results show that the paper can achieve 82% and 80% accuracy using Artificial Neural Network and Random Forrest, respectively. The data set in this work comes from 600 competition events.

Costa et al. proposed to use banned champions, picked champions, player statistics, picked champions + players statistics, and banned champions + picked champions + player statistics as features in winning prediction [14]. The research uses a Decision Tree, Naïve Bayes, k-Nearest Neighbor, Support Vector Machine, Random Forest, and Linear Regression. The results show that Random Forrest and Linear Regression can achieve 97% accuracy when using player statistics.

Tuzcu et al. investigated the impact of feature selections on the winning prediction accuracy of League of Legends [15]. The research employed a Gini score-based algorithm to select the top ten features with Random Forrest, Decision Tree, Naïve Bayes, Logistic Regression, Gradient Boosting, LightGBM, and AdaBoost. The result shows that by selecting the top ten features, the algorithm's accuracy can be improved;

for example, Logistic Regression is improved from 89% to 98%, while Gradient Boosting is improved from 96% to 98%.

Outside of e-sports, machine learning is also used to predict winners in traditional sports. Ishi et al. studied two machine learning algorithms, Logistic Regression and Support Vector Machine, to predict the winning side of the One Day International Cricket match [16]. The two algorithms are studied as individual and as ensemble approaches. The features used in this study are the force to hit the ball, the scoring pattern, and the team's overall strength. The results show that both Logistic Regression and Support Vector Machine, as individual algorithms, can archive up to 96.3%. However, when combining Logistic Regression and Support Vector Machine into an ensemble, the accuracy can increase to 96.07%.

Table 1. Related Works Summary

Related Works	Game	Dataset Criteria			Feature Criteria		
		Pre-Game	Tournament	Global Ban Pick	Hero Selection	Synergy	Counter
[1, 10]	Dota 2	Y	N	N	Y	Y	Y
[2]	Dota 2	Y	N	N	Y	N	N
[3]	HoK	Y	N	Y	Y	N	N
[4]	Dota 2	Y	Y	N	Y	N	N
[5]	HoK	N	N	N	Y	Y	Y
[6, 9, 11]	Dota 2	Y	N	N	N	N	N
[7]	HoK	N	N	N	N	N	N
[8]	HoK	Y	N	N	N	N	N
[12]	LoL	Y	N	N	Y	N	N
[13]	LoL	Y	Y	N	Y	N	N
[14]	LoL	Y	N	Y	Y	Y	N
[15]	LoL	Y	N	N	Y	N	N
This paper	**RoV**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**

Table 1 shows a summarization of related works. Two criteria were considered: dataset criteria and feature criteria. The dataset criteria consider three sets of parameters: pre-game parameters, tournament parameters, and global ban pick parameters. The feature criteria include *Hero Selection*, *Synergy*, and *Counter*. These are the features that are presented in the game. Dota 2 and League of Legends (LoL) are popular MOBA games similar to RoV. Compared with the works in the literature, this paper proposes considering *Hero Selection, Synergy,* and *Counter* features. Also, this paper predicts the game result in the pre-game period and considers the global band pick rule. Finally, this paper uses data from international tournament events where professional, top-of-the-class players play.

3. Materials and Methods

This research considers two machine learning approaches. First, a set of individual algorithms, K-Nearest Neighbor (KNN), Logistic Regression (LR), and Decision Tree (DT), will be trained with the best parameters as a baseline. Then, ensemble learning is proposed with voting and stacking decision-making. The proposed algorithm is validated using 5-fold cross-validation, and the f-score is measured to show the accuracy of the proposed algorithm.

3.1 Individual Learning Models

In this research, three individual machine learning algorithms are selected to predict the winning side. They are selected based on popularity in literature; thus, they can present the current state-of-the-art research in this area. The high-level concepts of each algorithm are as follows.

K-Nearest Neighbors (KNN) is a non-parametric supervised learning method that can be used for classification and regression. In classification, the output is the maximum count of a class of nearest neighbors. In regression, the output is the property value, which is the average of the values of k nearest neighbors.

Logistic Regression (LR) models the probability of an event by having the log odds for the event to be a linear combination of independent variables. Logistic regression estimates the probability of an event occurring based on the linear combination of independent variables of the training dataset.

Decision Tree (DT) is a predictive model to conclude a set of observations. Expressly, in classification, leaves of decision trees represent class labels, and branches represent conjunctions of features that lead to a particular leaf.

3.2 Ensemble Learning Models

Ensemble learning uses multiple machine-learning techniques to solve the same problem. Then, the outputs of the techniques are combined to create a final result. In particular, an ensemble function combines multiple outputs to form a better output than an individual one. Many ensemble functions exist, such as Bayes' optimal classifier, boosting, bucket of models, and bagging. This research utilizes two ensembles, namely, stacking and voting.

Stacking or stacked generalization trains the model by combining the prediction of several other learning algorithms. The process starts by training all the other learning algorithms, called based estimators, with the available data. The prediction performance of all algorithms is recorded. Then, a final estimator, in this paper, either KNN, LR, or DT, is trained to make a final prediction using all the predictions of the base estimators as additional inputs.

Voting combines prediction results from many machine learning algorithms with the same data. There is no assumption on the accuracy or performance of the algorithms. When the problem is classified, the result will be the class with the majority among all algorithms. On the other hand, when the problem is prediction, the result will be the average of all results. Furthermore, the voting process can be either soft or hard voting. The class with the highest vote count will be the final result in hard voting. Soft voting summarizes the average probability of each class and then declares the winner as having the highest weighted probability.

3.3 Datasets

This research uses past game data from professional RoV tournaments from the AIC 2022. The tournament employs the global band pick rule and consists of 1,017 games. Following are the rules of the tournament that are applied to the dataset:

Global ban pick: In a best-of-three or best-of-five game, the global band pick rule disallows players to choose the same heroes from the previous round. However, this rule will be applied to the first six games on a best-of-seven game. However, in the seventh game, any heroes can be used.

Right to pick the playing side: The team on the left-hand side has the right to pick the playing side in the first game. In the following games, the loser can pick the playing side.

Ban pick order: in the first ban period, the blue team will pick two heroes to be banned from the Red team, and vice versa. Then, the blue team will pick the first hero. Next, the red team will pick their first two heroes. This selection of two heroes will be switched for the blue and red teams. Finally, the red team will pick their last hero. Then, the second ban period will be started, repeating the same process.

Table 2. Example Data Used in This Research.

ROfflane	RJungle	RMage	RCarry	RSupport	BSupport	BCarry	BMage	BJungle	BOfflane	Result
Tachi	Keera	Liliana	Tel'Annas	Roxie	Xeniel	Laville	Lorion	Kriknak	Veres	0
Airi	Skud	Dirak	Slimz	Payna	Arum	Elsu	Krixi	Tulen	Yena	0
Qi	Errol	Yue	Laville	Lumburr	Baldum	TheJoker	Krixi	Aoi	Florentino	1

Table 2 shows example data from three RoV games. Each row represents data from a game, while each column shows a feature. The first five columns are selected heroes of the Red Team, with the prefix R. The following five columns are heroes from the Blue Team, with the prefix B. Each hero has their position, e.g., offlane or jungle. The names of selected heroes are in each row's first column to the tenth column. The final column is the end game result; 0 is the Red Team win, while 1 is the Blue Team win. The field type of the 1st – 10th columns is a string, while the last is an integer. However, the data must be pre-processed before applying to the model, as shown in the next section.

3.3 Feature Set

Because the number of heroes in each game is limited to only 5 per team from the 109 heroes, many heroes will not be selected. To represent the *Hero Selection*, one-hot encoding is used to represent the red team and blue team hero selection, as shown in Equations 1 and 2, respectively. In the equations, X_i represents the i^{th} feature. Hence, for the hero's selection, there will be 218 features.

$$X_{0+i} = \begin{cases} 1, if\ hero\ i\ is\ on\ red\ team \\ 0, otherwise \end{cases} \quad (1)$$

$$X_{109+i} = \begin{cases} 1, if\ hero\ i\ is\ on\ blue\ team \\ 0, otherwise \end{cases} \quad (1)$$

Then, the combination of heroes in the same team will be considered to represent the *Synergy* property among heroes. Let S_{ij} be the winning ratio when hero i and hero j are on the same team. Equations 3 and 4 show how to measure the synergy property of the red team, S_R, and the blue team, S_B, respectively. Then, the synergy feature is measured from the difference in the total synergy of the red and blue teams, as shown in Equation 5 as the 219th feature.

$$S_R = \sum_{i \in R} \sum_{j \in R, i \neq j} S_{ij} \quad (2)$$

$$S_B = \sum_{i \in B} \sum_{j \in B, i \neq j} S_{ij} \quad (3)$$

$$X_{218} = S_R - S_B \quad (4)$$

On the other hand, *Counter* property considers how a hero on one team impacts a hero on the other. Let C_{ij} be a winning rate when hero i is used against hero j of the other team. Then, the counter value of the red team is calculated as in Equation 6, where R is the set of heroes of the red team and B is the set of heroes of the blue team. Next, the counter value of the blue team can be calculated from $C_B = 1 - C_R$. Finally, the counter property of the red team can be assigned as the 220th feature, i.e., $\boldsymbol{X_{219} = C_R}$. There is no need to include C_B feature.

$$C_R = \sum_{i \in R} \sum_{j \in B} C_{ij} \quad (5)$$

There will be 220 features used in this paper to train machine learning models.

3.4 Methodology

This research investigates eight machine learning prediction models: three traditional algorithms and five ensemble configurations. The models are evaluated using the following process.

1. The AIC2022 dataset is encoded using one-hot encoding, as mentioned in section 3.3, to create feature set.

2. Each model's optimized parameter is indicated, as discussed in section 4.2-4.7. For ensemble learning with voting, there is no need to find an optimized parameter.

3. The encoded data in step 1 is used to evaluate individual models using k-fold-validation, as mentioned in section 4.1

4. The evaluation results are compared as shown in section 4.8 and 4.9

5. Unseen data from the 31st Southeast Asian Games e-sport event is tested against the models to evaluate the generality of the models.

4. Results and Discussion

This paper evaluates three individual machine learning algorithms, namely, K-Nearest Neighbor (KNN), Logistic Regression (LR), and Decision Tree (DT), against five ensemble learning methods using either voting or stacking. The algorithms will be tested on several scenarios with the combination of the following properties: *Hero Selection*, *Synergy,* and *Counter*. The dataset contains 1,017 games collected during AIC 2022. The algorithms will be tested with accuracy to predict the winning side. For each algorithm, the optimized parameter is identified first and then used in the comparison.

4.1 Validation Method

To validate the prediction outcome from the machine learning algorithms, this paper uses k-Fold Cross-Validation with k is five. K-Fold Cross Validation creates a set of data subsets, five subsets in this paper, then uses one as testing data while the rest is used for training. Then, in the next round, another subset is selected as testing data, while the rest is used for training. This process is repeated until all subsets are selected as testing data once. The output of all subset tests will be averaged and used as the result of the test. The k-fold cross-validation is used in this paper because of the limited dataset size.

4.2 Optimized Parameter for k-Nearest Neighbors

To properly use k-Nearest Neighbors, a value of k that is suitable for the dataset and application must be identified.

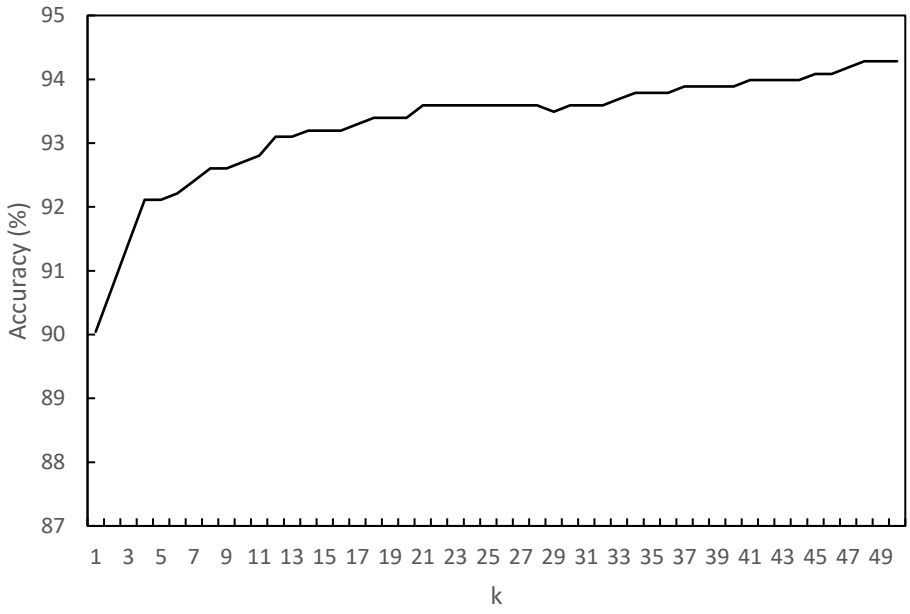


Figure 1. Impact of k Value on Prediction Accuracy

Figure 1 shows the impact of the k value on k-Nearest Neighbors accuracy. The feature set used in this result combines *Hero Selection, Synergy,* and *Counter.* The k value is evaluated in the range of 1 to 50. K-fold cross-validation with five folds is used to measure the algorithm's accuracy. The result shows that the best value of k is 49, which will be used in this paper.

4.3 Optimized Parameter for Logistic Regression

In order to properly use Logistic Regression, a value of C, i.e., regularization strength, that is suitable for the dataset and application needs to be identified.

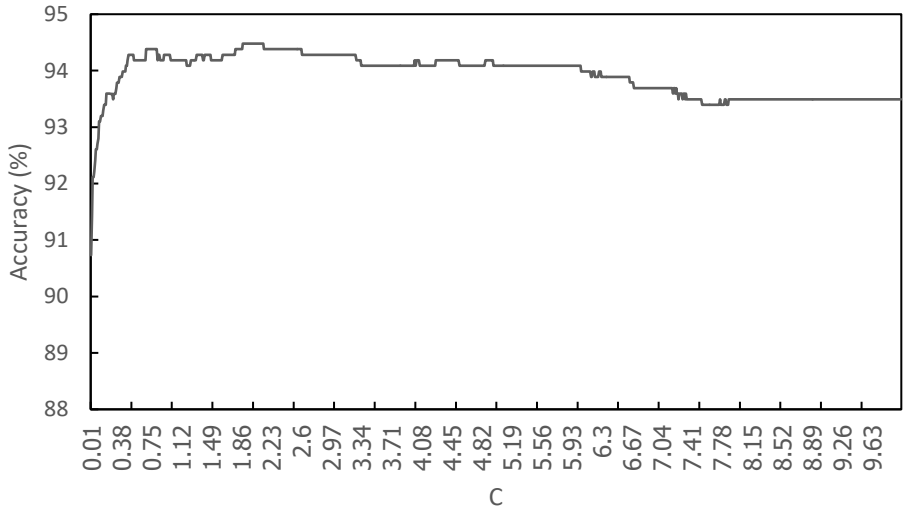


Figure 2. Impact of C Value on Prediction Accuracy

Figure 2 shows the impact of the C value on Logistic Regression accuracy. The feature set used in this result combines *Hero Selection, Synergy,* and *Counter.* The value C is evaluated in the range of 0.01 to 10.0. K-fold cross-validation with five folds is used to measure the algorithm's accuracy. The result shows that the best value of C is 1.89, which will be used in this paper.

4.4 Optimized Parameter for Decision Tree

In order to properly use a decision tree, a tree max depth value suitable for the dataset and application needs to be identified.

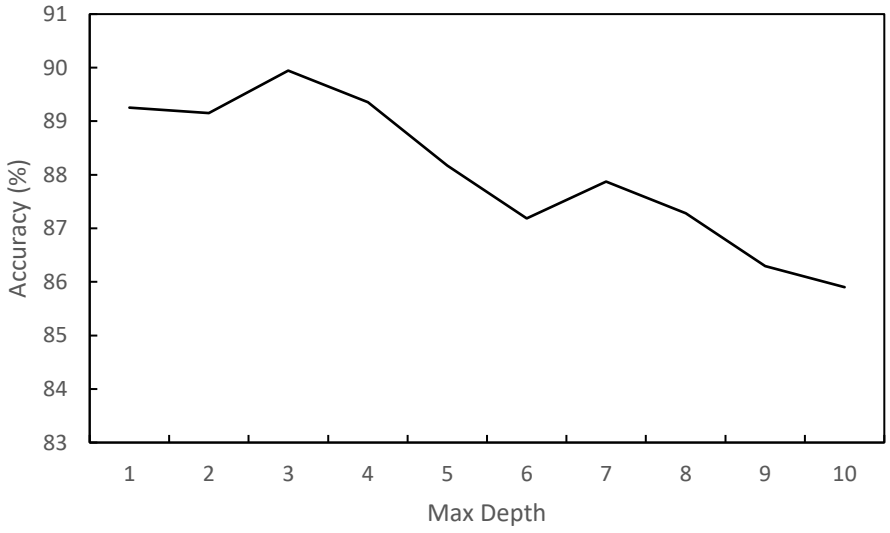


Figure 3. Impact of Max Depth on Prediction Accuracy

Figure 3 shows the impact of max depth value on Decision Tree accuracy. The feature set used in this result combines *Hero Selection, Synergy,* and *Counter.* The depth value is evaluated in the range of 1 to 10. K-

fold cross-validation with five folds is used to measure the algorithm's accuracy. The result shows that the best value of max depth is 3, which will be used in this paper.

4.5 Optimized Parameter for Stack Ensemble Learning with KNN

In order to properly use k-Nearest Neighbors as the final estimator in ensemble learning with stack, a value of k that is suitable for the data set and application needs to be identified.

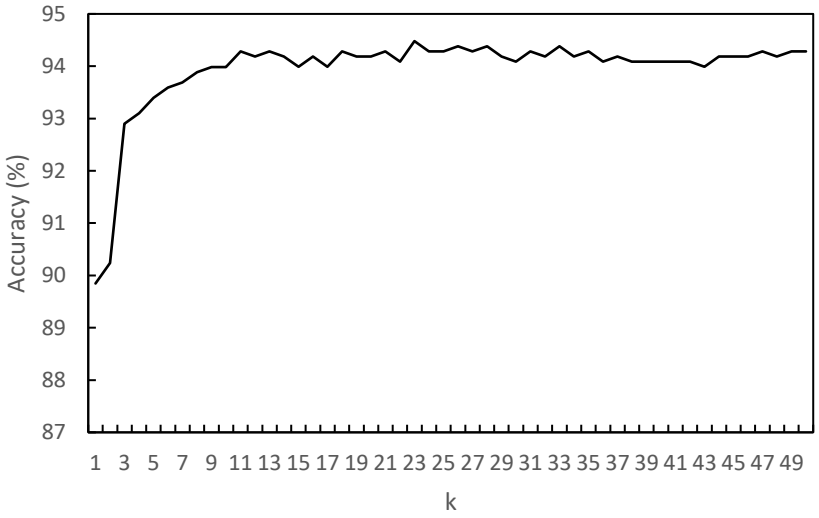


Figure 4. Impact of k Value on Prediction Accuracy

Figure 4 shows the impact of the k value on the accuracy of ensemble learning with stack and k-nearest Neighbors as the final estimator. The feature set used in this result combines *Hero Selection*, *Synergy*, and *Counter*. The k value is evaluated in the range of 1 to 50. The k-fold cross-validation with five folds measures the algorithm's accuracy. The result shows that the best value of k is 23, which will be used in this paper.

4.6 Optimized Parameter for Stack Ensemble Learning with LR

In order to properly use logistic regression as a final estimator in ensemble learning with stack, a value of C, i.e., regularization strength, that is suitable for the data set and application, needs to be identified.

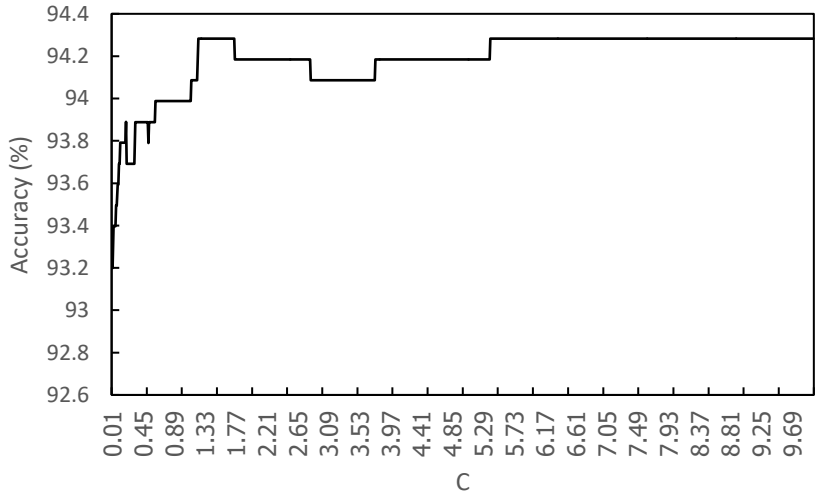


Figure 5. Impact of C Value on Prediction Accuracy

Figure 5 shows the impact of the C value on the accuracy of ensemble learning with stack and Logistic Regression as the final estimator. The feature set used in this result combines *Hero Selection, Synergy,* and *Counter.* The value C is evaluated in the range of 0.01 to 10.0. K-fold cross-validation with five folds measures the algorithm's accuracy. The result shows that the best value of C is 1.24, which will be used in this paper.

4.7 Optimized Parameter for Stack Ensemble Learning with DT

In order to properly use the Decision Tree as the final estimator in ensemble learning with stack, a value of tree max depth suitable for the data set and application needs to be identified.

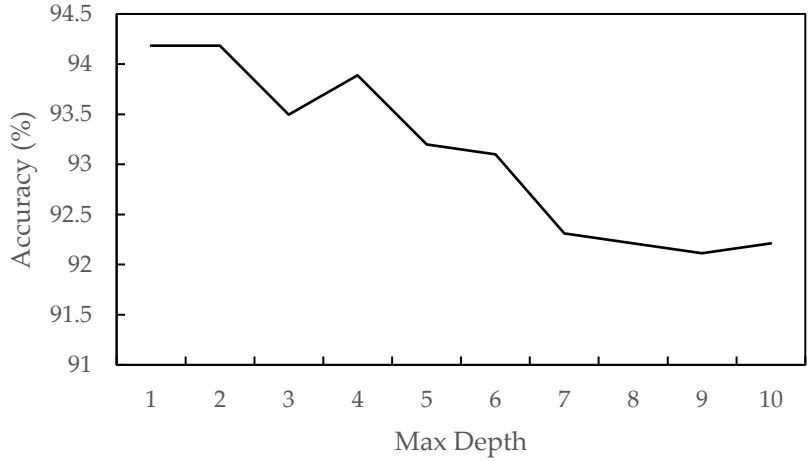


Figure 6. Impact of Max Depth on Prediction Accuracy

Figure 6 shows the impact of max depth value on the accuracy of ensemble learning with stack and Decision Tree. The feature set used in this result combines *Hero selection*, *Synergy,* and *Counter*. The depth value is evaluated in the range of 1 to 10. The k-fold cross-validation with five folds measures the algorithm's accuracy. Interestingly, the result shows that the best max depth values are 1 and 2. When the value of max depth increases, the accuracy is decreased. Thus, the value of max depth will be used in this paper.

The optimized parameter from Section 4.2 – 4.7 will be used in the subsequent evaluation.

Table 3. Comparing Prediction Accuracy with K-fold Validation.

	Method	Hero Selection	Synergy	Counter	Synergy + Counter	Hero Selection + Synergy	Hero Selection + Counter	Hero Selection + Synergy + Counter
Individual	**KNN**	0.525457	0.872995	0.839987	0.**883625**	0.857321	0.844456	0.878027
	LR	0.515958	0.876351	0.845019	0.892579	0.874669	0.878587	**0.924457***
	DT	0.523787	0.875239	0.842783	**0.881942**	0.872447	0.843340	**0.881942**
Ensemble	**Vote (Soft)**	0.531059	0.876353	0.843907	0.887537	0.879713	0.876354	**0.919438**
	Vote (Hard)	0.539458	0.875232	0.842226	0.887542	0.870760	0.857892	**0.899289**
	Stacking (KNN)	0.498029	0.868521	0.836632	0.884734	0.873552	0.869633	**0.916631**
	Stacking (LR)	0.522660	0.877472	0.842228	0.888659	0.878032	0.877468	**0.922787**
	Stacking (DT)	0.514822	0.876913	0.839987	0.885859	0.870749	0.874672	**0.923898**

4.8 Prediction Accuracy with K-fold Validation

After indicating all proper parameters for each algorithm, the algorithms are evaluated against each other, as shown in Table 3. The accuracy is measured using k-fold cross-validation with five folds. The prediction accuracy is measured as the average of the prediction on each testing subset. The value with a bold number is the best in each algorithm, while the value with underline is the best when considering a combination of properties, i.e., *Hero Selection, Synergy,* and *Counter.* The value with an asterisk symbol (*) is the best overall.

The result shows that by considering the combination of *Hero Selection, Synergy,* and *Counter* to train the machine learning model, the winner prediction accuracy can be improved in most cases, as shown in the bold number in each row. Considering traditional algorithms, from the result of section 4.2, the value of k for KNN is very high, i.e., 49, which indicates that many clusters are compared with the data set size. Thus, the prediction accuracy of KNN will be sub-optimal, i.e., too specific. On the other hand, the result of section 4.4 shows that the depth of DT is very shallow, which will lead to sub-optimal prediction, i.e., too generic. Therefore, the result of LR is the best among the three traditional algorithms, as shown by the underscored number in each column. Furthermore, ensemble learning shows promising performance in many cases but cannot outperform individual algorithms, i.e., LR, as the best algorithm. However, it can achieve the second-best prediction accuracy with stacking (DT) configuration and utilize all three features very close, i.e., less than 0.000559, which is different from the best.

The result shows that the prediction can improve by using the two additional features to the heroes' selection, and ensemble learning can improve prediction accuracy compared with the individual algorithms.

Table 4. Comparing Prediction Accuracy with F-Score

	Method	Hero Selection	Synergy	Counter	Synergy + Counter	Hero Selection + Synergy	Hero Selection + Counter	Hero Selection + Synergy + Counter
Individual	**KNN**	0.520192	0.872196	0.836569	**0.882000**	0.855409	0.841983	0.877146
	LR	0.502259	0.875116	0.843147	0.891407	0.872802	0.877524	**0.923809***
	DT	0.537429	0.875998	0.838133	**0.880860**	0.873353	0.836794	**0.880860**
Ensemble	**Vote (Soft)**	0.522808	0.876238	0.839832	0.887369	0.877740	0.875259	**0.918888**
	Vote (Hard)	0.553788	0.874884	0.838370	0.886313	0.870047	0.853910	**0.898619**
	Stacking (KNN)	0.475083	0.865899	0.833573	0.882672	0.869444	0.869265	**0.915175**
	Stacking (LR)	0.472897	0.877402	0.838911	0.887390	0.875785	0.875863	**0.921995**
	Stacking (DT)	0.374027	0.876617	0.836780	0.887983	0.874842	0.869067	**0.923089***

4.9 Prediction Accuracy with F-Score

F-score measures the prediction accuracy from precision and recall of the test. This paper measures a balanced F-score (i.e., F1 score), as shown in Equation 7.

$$F_1 = 2\ \frac{precision\ x\ recall}{precision + recall} \quad (6)$$

Table 4 shows the results of the F-score of each model used in this paper. The results are very similar to the previous section, confirming that using *Hero Selection*, *Synergy*, and *Counter* features in the training model helps achieve the best prediction accuracy in most cases, indicated by the bold number in each row. Also, linear regression shows the best balance on precision and recall; however, stacking ensemble learning with DT performs similarly. As mentioned in the previous section, LR is more suitable for this application than KNN and DT. Moreover, ensemble learning generally can outperform KNN and DT, and ensemble learning with DT can compete with LR

The result shows that the prediction can improve by using the two additional features to the heroes' selection, and ensemble learning can improve prediction accuracy compared with the individual algorithms.

Table 5. Comparing Prediction Accuracy with Unseen Data

	Method	Hero Selection	Synergy	Counter	Synergy + Counter	Hero Selection + Synergy	Hero Selection + Counter	Hero Selection + Synergy + Counter
Individual	KNN	0.555556	0.611111	0.638889	**0.694444**	0.555556	0.583333	0.666667
	LR	0.555556	0.611111	**0.722222***	0.638889	0.527778	0.583333	0.555556
	DT	0.500000	0.611111	0.638889	**0.694444**	0.583333	0.611111	**0.694444**
Ensemble	Vote (Soft)	0.500000	0.611111	**0.638889**	**0.638889**	0.555556	0.583333	**0.638889**
	Vote (Hard)	0.555556	0.611111	0.638889	**0.666667**	0.583333	0.583333	**0.666667**
	Stacking (KNN)	0.638889	0.638889	**0.666667**	0.638889	0.583333	0.555556	0.527778
	Stacking (LR)	0.555556	**0.638889**	**0.638889**	**0.638889**	0.555556	0.583333	0.583333
	Stacking (DT)	0.527778	**0.638889**	**0.638889**	**0.638889**	0.500000	0.583333	0.527778

4.10 Prediction Accuracy with Unseen Data

The models are used to predict unseen data to investigate the generality of the prediction models. The models were trained with the AIC datasets; then, they were tested against the data set from the 31st Southeast Asian Games e-sport event. The testing dataset includes 36 competition games. As shown in Table 5, the result shows that linear regression that considers the *Counter* property shows the highest prediction accuracy, shown in bold number, with 0.722222 accuracy. The result indicates that individual algorithms, such as LR, can be more generalized than ensemble learning. However, the dataset from the 31st Southeast Asian Games is minimal and has only 36 competition games; therefore, this will be left for further study in future works.

5. Conclusions

Realm of Valor (RoV) is a popular multiplayer online battle arena game. This paper proposes to use machine learning to predict the winner's side based on the selection of heroes used in the battle by the player and opponent. Four machine learning techniques, namely, k-nearest neighbor, logistic regression, decision tree, and ensemble learning, are compared with their optimized parameters. The evaluation result shows that by considering three features, namely *Hero Selection*, *Synergy*, and *Counter*, which are features available only in RoV, the winning prediction accuracy can be improved. Also, ensemble learning can outperform individual

algorithms as the best prediction algorithm. Future work includes further investigation of the generality of the prediction models and additional features that can be used to improve the models.

6. Acknowledgements

Author Contributions: Conceptualization, N.T. and P.B..; methodology, N.T. and P.B..; software, N.T.; validation, N.T. and P.B..; formal analysis, N.T. and P.B..; data curation, N.T.; writing–original draft preparation, N.T.; writing–review and editing, P.B..;. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Kinkade, N.; Jolla, L.; Lim, K. Dota 2 win prediction. Technical Report, University California San Diego, La Jolla, CA, USA, **2015**

[2] Semenov, A.; Romov, P.; Korolev, S.; Yashkov, D.; Neklyudov, K.. Performance of machine learning algorithms in predicting game outcome from drafts in Dota 2, In Proceeding of 5th International Conference Analysis of Images, Social Networks and Texts (AIST 2016), **2016**, pp. 26-37.

[3] Chen, S.; Zhu, M.; Ye, D.; Zhang, W.; Fu, Q.; Yang, W. Which heroes to pick? Learning to draft in MOBA games with neural networks and Tree Search, *IEEE T GAMES*, **2021**, *13*(4), 410–421.

[4] Hodge, V. J.; Devlin, S.; Sephton, N.; Block, F.; Cowling, P. I.; Drachen, A. Win prediction in multiplayer esports: Live professional match prediction, *IEEE T GAMES*, **2021**, *13*(4), 368–379.

[5] Tian, P.; Lan, W.; Zhang, X. Hero featured learning algorithm for winning rate prediction of honor of kings, In Proceeding of 24th IEEE Conference on Games (CoG), **2022**, pp. 322-329.

[6] Song, K.; Zhang, T.; Ma, C. Predicting the winning side of DotA2, Technical Report, Stanford University, Stanford, CA, USA, **2015**.

[7] Yang, Z.; Pan, Z.; Wang, Y.; Cai, D.; Shi, S.; Huang, S.-L.; Bi, W.; Liu, X. Interpretable real-time win prediction for honor of kings—a popular mobile MOBA Esport, *IEEE T GAMES*, **2022**, Volume *14*(4), 589–597.

[8] Yang, Z.; Wang, Y.; Li, P.; Lin, S.; Shi, S.; Huang, S.-L.; Bi, W. Predicting events in MOBA games: Prediction, attribution, and evaluation, *IEEE T GAMES*, **2023**, *15*(2), 193-201.

[9] McGuire, S.; Epifano, J. Dota 2 Game Prediction, Technical Report, Rowan University, Glassboro, NJ, USA, **2018**.

[10] Conley, K.; Perry, D. How Does He Saw Me? A Recommendation Engine for Picking Heroes in Dota 2, Technical Report, Stanford University, Stanford, CA, USA, **2013**.

[11] Kalyanaraman, K. To win or not to win? A prediction model to determine the outcome of a DotA2 match, Technical Report, University of California San Diego, La Jolla, CA, USA, 2015.

[12] Do, T. D.; Wang, S. I.; Yu, D. S.; McMilian, M. G.; McMahan, R., P. Using Machine Learning to Predict Game Outcomes Based on Player-Champion Experience in League of Legends, In Proceeding of the 16th International Conference on the Foundations of Digital Games, **2023**, pp. 1 – 5.

[13] Sena, I. G. W.; Emanuel, A. W. R. Mobile Legend Game Prediction Using Machine Learning Regression Method, *Jurnal Teknologi dan Sistem Informasi*, **2023**, *9*(2).

[14] Costa, L. M.; Mantovani, R. G.; Souza, F. C. M.; Xexéo, G. Feature Analysis to League of Legends Victory Prediction on the Picks and Bans Phase, In Proceeding of 23rd IEEE Conference on Games (CoG), **2021**, pp. 1-5.

[15] Tuzcu, A.; Ay, E. G.; Uçar, A. U.; Kılınç, D. A Machine Learning Based Predictive Analysis Use Case For eSports Games, *Artificial Intelligence Theory and Application*, **2023**, *3*(1), 25-35.

[16] Ishi, M.; Patil, D. J.; Patil, D. N.; Patil, D. V. Winner prediction in one-day international cricket matches using Machine Learning Framework: An ensemble approach, *Indian J of Computer Science and Engineering*, **2022**, *13*(3), 628–641.

ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

Research article

Phytochemical Analysis and Biological Activities of Propolis from *Geniotrigona thoracica*: Evaluating its Therapeutic Applications

Kannika Bunkaew[1], Petcharat Ponpichai[2], Monthon Lertworapreecha[1,3], Jakkrawut Maitip[4], and Wankuson Chanasit[1,5*]

[1] Microbial Technology for Agriculture, Food and Environmental Research Centre, Faculty of Science and Digital Innovation, Thaksin University Phatthalung Campus, 93210
[2] Faculty of Science and Digital Innovation, Thaksin University, Phatthalung, 93210, Thailand
[3] Faculty of Science and Digital Innovation, Thaksin University, Phatthalung, 93210, Thailand
[4] Faculty of Science, Energy and Environment, King Mongkut's University of Technology North Bangkok, Rayong 21120, Thailand
[5] Faculty of Science and Digital Innovation, Thaksin University, Phatthalung, 93210, Thailand
* Correspondence: wankuson.c@tsu.ac.th

Citation:
Bunkaew, K.; Ponpichai, P.; Lertworapreecha, M.; Maitip, J.; Chanasit, W. Phytochemical analysis and biological activities of propolis from *Geniotrigona thoracica*: Evaluation its therapeutic applications. ASEAN J. Sci. Tech. Report. **2024**, 27(5), e253219. https://doi.org/10.55164/ajstr.v27i5.253219

Article history:
Received: March 15, 2024
Revised: August 30, 2024
Accepted: September 5, 2024
Available online: September 7, 2024

Abstract: Propolis is a substance that safeguards the bee hive against physical and microbiological threats. This research assessed the phytochemical properties and biological effects of propolis produced by stingless bees *Geniotrigona thoracica* collected from Phatthalung, Southern Thailand. The findings revealed that the ethanolic extract propolis (EEP) from *G. thoracica* exhibited antibacterial properties against certain foodborne pathogens (*Bacillus cereus*, *Staphylococcus aureus*, *Escherichia coli*, and *Salmonella* Typhimurium) with moderate to strong zone of inhibition (ZOI) in the range of 10 mm ≤ ZOI ≤ 15 mm. Additionally, the extract propolis demonstrated antioxidant activity, achieving up to 80% DPPH radical scavenging when 50 mg/mL EEP was tested. Furthermore, the crude propolis extract showed anti-inflammatory effects on macrophage cells, resulting in a 72.9% reduction in nitric oxide (NO) levels in LPS-activated RAW 264.7 cells exposed to 100 mg/mL of EEP. The GC-MS chromatogram identified the phytochemical compositions of the EEP, with Lup-20(29)-en-3-ol or lupeol (25.42%) and β-amyrone (22.66%) as the major compounds, both triterpenoid derivatives. Other notable constituents included alkane hydrocarbon pentacosane (6.63%), fatty alcohol cis-9-eicosenol (4.23%), and phenolic compound 3-pentadecylphenol (3.86%). Therefore, the EEP derived from *G. thoracica*, possessing such diverse biological activities, holds promise for medicinal and functional food applications.

Keywords: Stingless bees; *Geniotrigona thoracica*; Propolis extract; Biological effect

1. Introduction

Stingless bees belong to the Meliponini group, comprising more than 600 species, and are widely distributed in tropical and subtropical regions [1, 2]. Typically, propolis consists of resins collected by bees from plant sources, mixed with saliva and beeswax within the hive. Its primary functions include sealing cracks and protecting against threats [3-5]. The primary component of propolis, comprising over 45%, is lipids, which aid in inhibiting microbial growth. Additionally, propolis contains over 150 compounds, including phenolic

compounds, terpenoids, steroids, and aromatic acids, contributing to its antimicrobial properties. Notably, phenolic compounds such as flavonoids are well-known for their antioxidant properties [6-9]. The bioactive compounds in propolis, especially polyphenols, flavonoids, and terpenes, have led to its application in biomedicine, natural product cosmetics, and as an ingredient in health foods [3, 4, 6, 10-12]. The chemical composition of propolis varies depending on stingless bee species, the timing of collection, the botanical environment, and geographical location [3, 8, 9, 11, 13]**.** Despite the differences in the chemical composition of propolis worldwide, they all show pharmacological activity, making it an attractive natural product. To date, few studies have investigated the chemical composition of propolis from the Thai stingless bee *Geniotrigona thoracica*. Most research has been conducted in Malaysia and Brunei. For example, Nazir et al. [6] identified up to 30 new compounds for the first time from the ethanolic extract of Malaysian *G. thoracica* propolis, with phenolic and terpenoid compounds being the major components as determined by GC-MS analysis [6]. In the propolis of Brunei stingless bees, *G. thoracica* contained lipids as a major component (45.60-47.86%), with minimal carbohydrate and protein content but rich in minerals. Additionally, analysis of functional groups indicates the presence of phenolic and flavonoid compounds. These bioactive compounds contribute to the antioxidant and antibacterial properties of the propolis extract. Ethanol extraction of propolis from *G. thoracica* exhibits antimicrobial effects against both Gram-positive bacteria, e.g., *Bacillus subtilis, Staphylococcus aureus*, and Gram-negative bacteria, e.g., *Escherichia coli, Pseudomonas aeruginosa* [8]. However, the phytochemical compounds, including flavonoid, coumarin, saponin, terpenoid, steroid, and cardiac glycoside, were detected in the ethanol extract of propolis from *G.thoracica*, collected in Pattani province, Thailand, and showing the highest antioxidant activity with the IC_{50} at 262.43 μg/mL and total phenolic content at 60.13 mgGAE/g extract [14]. Therefore, this study aims to investigate the anti-bacterial and antioxidant activity of the ethanolic propolis extract from stingless bees *Geniotrigona thoracica* harvested in Phatthalung Province, Southern Thailand. In addition, the anti-inflammatory potential of its extract is determined by measuring nitric oxide (NO) production. Finally, the bioactive compositions in the propolis extract is identified by GC-MS analysis.

2. Materials and Methods

2.1 Propolis extraction

The propolis sample from *G. thoracica* was collected from Pantae community enterprise, Khuan Khanun district, Phatthalung province, Thailand (locality; 7.806259568725061, 100.01692785193286). The propolis was collected from the beehives between July and September 2023. The collected propolis was rinsed with distilled water and dried using a dehumidifier at room temperature for 2 weeks. After drying, the propolis was ground into small pieces less than 1 millimeter in size. Subsequently, 100 g of dried propolis was extracted in 1 liter of 70% ethanol (the ratio of raw propolis to solvent was 1:10) with an ultrasonicator (DR-MH40, Ultrasonic Cleaner, Derui) for 60 min at 40°C and the sample was then left for maceration for 7 days at 25°C after the treatment. The resulting supernatant was later subjected to rotary evaporation (4001, Heidolph, Schwabach, Germany) until the solvent volume was reduced by approximately half, followed by drying under vacuum at 40°C [7, 8, 12]. The yield content of the propolis extract was calculated by Yield (%) = [Weight of extracted propolis after solvent evaporation(g)/Weight of the initial dried propolis] × 100.

2.2 Antibacterial analysis

The antibacterial activities of the ethanolic extract propolis (EEP) from *G. thoracica* were assessed using the agar disc diffusion assay. The bacterial strains tested included two Gram-positive strains (*Staphylococcus aureus* ATCC-29213 and *Bacillus cereus* ATCC-14579) and two Gram-negative strains (*Escherichia coli* ATCC-11775 and *Salmonella* Typhimurium ATCC 14028). Briefly, the bacterial culture suspension was adjusted in 0.85% (w/v) NaCl to approximately ~10^8 CFU/mL. It was then swabbed on Mueller-Hinton agar, MHA (Oxoid™), while the EEP was prepared to final concentrations ranging from 50 to 1,600 mg/mL. The sterile filter paper discs (6 mm in diameter) containing the extractions were subsequently placed on those MHA. They were

incubated at 37°C for 16-18 hours before assessing bacterial growth inhibition by measuring the diameter of the inhibition zone (mm). Grading of the zone of inhibition (ZOI) followed the description by Bhaigybati et al. (2020) [15], of which 6-8 mm: No antimicrobial activity, 8.1-9 mm: Slight antimicrobial activity, 9.1-12 mm: Moderate antimicrobial activity, 12.1-15 mm: clear antimicrobial activity, and >15 mm: Strong antimicrobial activity. Ampicillin (10 µg) and Ciprofloxacin (5 µg) served as positive controls for Gram-positive and Gram-negative bacteria, respectively, while 5% DMSO was a negative control [13].

2.3 Antioxidant assays

The antioxidant activities of EEP from *G. thoracica* were determined using the 2,2-diphenyl-1-picrylhydrazyl (DPPH) assay [7, 8, 14] by measuring its ability to convert DPPH° into DPPH-H. In brief, 0.1 mL of the extract (ranging from 50 to 200 mg/mL) was combined with 2 mL of DPPH° solution (0.2 mM) in ethanol, and the mixture was left to incubate for 1 hour in darkness at room temperature. The absorbance was then recorded at 517 nm, with ascorbic acid as a positive control. The free radical scavenging activity of the propolis extract was calculated as $\%\text{Scavenging of DPPH}^\circ = [(A_{\text{Initial Absorbance}} - A_{\text{Final Absorbance}})/A_{\text{Initial Absorbance}}] \times 100$.

2.4 Anti-inflammatory

The level of nitric oxide (NO), a signaling molecule that plays a key role in the pathogenesis of inflammation, of the crude propolis extract from *G. thoracica* was determined using the Griess reagent (1% sulphanilamide, 0.1% N-(1-naphthyl) ethylenediamine, each in 2.5% H_3PO_4) as described by Mendez-Encinas et al.(2023) [16]. The RAW 264.7 macrophage cell line (ATCC Number: TIB-71™ ,Lot Number: 7006149 ,Species: Nouse (*Mus musculys*))was purchased from Thermo Fisher Scientific Inc., USA. The cells (1×10^5 cells/well, 100 µL) were plated in a 96-well plate and incubated for 24 hours at 37°C and 5% CO_2. Following incubation, the cells were treated with 50 µL of propolis extract ranging from 50 to 800 mg/mL in DMEM and stimulated with 50 µL of 10 µg/mL lipopolysaccharide (LPS), Sigma-Aldrich, USA, in DMEM, then further incubated for 24 hours. Control groups included cells treated with DMSO (negative control), cells treated with LPS (positive control), and cells treated with neither DMSO nor LPS. After incubation, aliquots (50 µL) of cell supernatants were collected, mixed with an equal volume of Griess reagent, and incubated in the dark at room temperature for another 10 min before measuring the absorbance change at 540 nm. Results were expressed as reduced sodium nitrite concentration (µM), then converted to nitric oxide (NO) production.

2.5 Identification and quantification of bioactive compounds in the propolis extract

Gas chromatography/mass spectrometry (GC-MS) analysis was conducted using a GC7890B and MSD5977B system (Agilent Technologies, USA) equipped with an HP-5Ms column (15 m × 250 mm × 0.25 µm). Helium served as the carrier gas at a flow rate of 1 mL/min. The GC-MS condition was as follows: the injector temperature was set at 280°C in split-less mode, the oven temperature was initially maintained at 60°C for 4 min, then increased to 150°C at a rate of 10°C/min for 15 min with a scan range of 35–500 Da [17]. The mass spectra were then compared to the National Institute of Standards and Technology (NIST) library data to identify and quantify the bioactive compounds.

2.6 Statistical analysis

All the values were presented as the mean ± standard deviation (SD). Data underwent analysis via one-way ANOVA, and differences between means were assessed using the LSD multiple range test ($p \leq 0.05$) with the SPSS program (IBM SPSS Statistics, version 27.0.1).

3. Results and Discussion

3.1 Propolis extraction and yield content

The extraction of propolis from *G. thoracica* using 70% ethanol resulted in a yield of approximately 27.21% of the raw sample. The process of preparing the crude extract is illustrated in **Figure 1**. Additionally, the literature suggests that ethanolic extracts contain more aromatic compounds than water extracts [7, 14]. Like the propolis extract from *Cerana indica*, the ethanolic extract had stronger antimicrobial activity than the

methanolic extract and aqueous extract [17]. This might be attributed to its less lipophilic behavior, as many of the phytochemicals possess electronegative functional groups, rendering them hydrophilic. The secondary metabolites are also highly soluble in organic solvents [6, 18]. Nevertheless, the maceration method using organic solvents is extensively utilized for propolis extraction. Indeed, maceration with 70% ethanol is the preferred organic solvent for propolis extraction [5, 18]. For instance, Silva et al. (2012) conducted a study showing variations in extraction efficiencies among different organic solvents. They found that the hydroalcoholic-extracted propolis demonstrated the highest propolis and flavonoid content levels of about 280 mg and 140 mg, respectively [19].

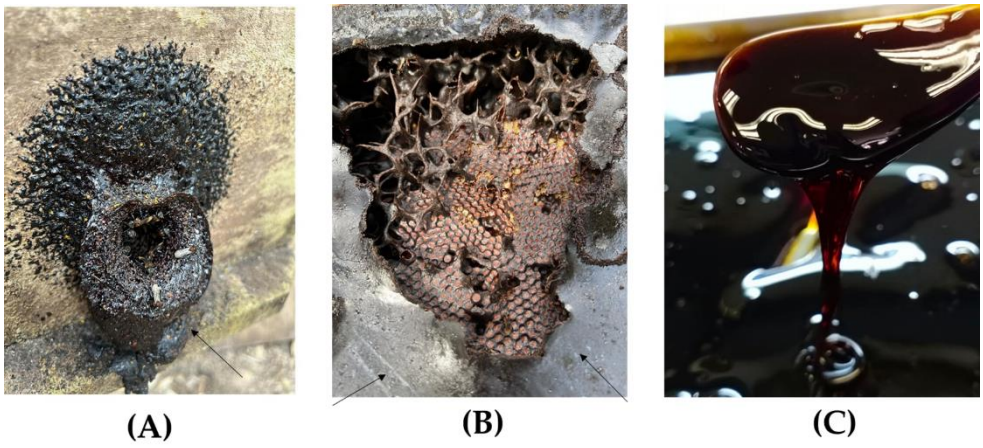


Figure 1. Propolis of *G. thoracica* collected in Pantae community enterprise, Khuan Khanun district, Phatthalung province, Southern Thailand. (A) the entrance of the *G. thoracica* colony and its adults, (B) raw propolis (with the arrows pointing), and (C) the ethanolic extract propolis (EEP) of *G. thoracica*

3.2 Anti-bacterial Activity

The anti-microbial properties of the EEP from *G. thoracica* against certain important foodborne pathogens, including two Gram-positive bacteria (*B. cereus* and *S. aureus)* and two Gram-negative bacteria (*E. coli* and *S.* Typhimurium) was, examined by disc diffusion assay to qualify its inhibitory activities on the tested strains. The inhibition zone of the various crude EEPs was demonstrated in **Table 1**.

Table 1. Antibacterial activity of crude propolis extract of *G. thoracica* against different foodborne pathogens

Antibacterial activity	Concentration	Inhibition zone (mm± SD)			
		B. cereus	*S. aureus*	*E. coli*	*S.* Typhimurium
Ethanolic extract propolis (mg/mL)	50	15.0 ± 0.00	13.5 ± 1.50	10.0 ± 0.01	10.0 ± 0.01
	100	15.5 ± 0.50	12.5 ± 0.50	10.0 ± 0.01	10.0 ± 0.01
	200	15.5 ± 0.50	13.0 ± 0.02	10.0 ± 0.02	11.0 ± 0.04
	400	15.0 ± 0.02	14.0 ± 0.02	-	11.0 ± 0.02
	800	14.5 ± 0.50	12.0 ± 0.05	-	-
	1,600	-	-	-	-
Positive control	Ampicillin[a] (10 µg)	35.0 ± 0.00	35.0 ± 0.02	-	-
	Ciprofloxacin[b] (5 µg)	-	-	30.0 ± 0.01	30.0 ± 0.01
Negative control	5% DMSO	-	-	-	-

(-) refer No inhibition where 6-8 mm: No antimicrobial activity, 8.1-9 mm: Slight antimicrobial activity, 9.1-12 mm: Moderate antimicrobial activity, 12.1-15 mm: clear antimicrobial activity, and >15 mm: Strong antimicrobial activity.
[a] Ampicillina used as a positive control for Gram-positive bacteria
[b] Ciprofloxacin used as a positive control for Gram-negative bacteria
Mean values of three replicates ± standard deviation (SD)

The results revealed that the EEP from *G. thoracica* showed good dose-dependent antibacterial activity. The highest growth inhibition was obtained in the 200 mg/mL extract sample with the largest zone of inhibition (ZOI) of about 15.5 ±0.50 mm against *B. cereus*, while in *S. aureus* needed about 400 mg/mL of the propolis extract to achieve the highest ZOI of 14±0.02 mm. On the other hand, in the tested Gram-negative bacteria (*E. coli* and *S.* Typhimurium), the crude extract showed deficient inhibition (an average ZOI of 10-11 mm). These bacterial inhibition tests of the EEP from *G. thoracica* were graded as moderate to strong antimicrobial activity. It is worth understanding that all these bacteria tests are considered foodborne diseases. In addition, *S. aureus* and *Salmonella* spp. were on the Thai Agricultural Standard Tas 8003-2013 list for honey. Abdullah et al. (2020) evaluated the potential of 2 mg/mL ethanol extract propolis from three different stingless bee species, *Geniotrigona thoracica, Heterotrigona itama,* and *Tetrigona binghami. S*imilarly to this study, *G. thoracica* propolis demonstrated the highest ZOI against *E. coli*, followed by *B. subtilis, P. aeruginosa*, and *S. aureus,* respectively, while in *H. itama* propolis showed the largest clear zone among this three propolis with approximately 13 mm against *B. subtilis* [8]. On the other hand, the investigation of the EEP from *G. thoracica*, *T. binghami*, and *H. itama* carried out by Zullkiflee et al. (2022) revealed that the highest ZOI of 11.7 mm obtained in *G. thoracica* propolis against *S. aureus,* in addition, the overall antibacterial inhibition was higher potent in the ethanolic extracts than in water extract. The MIC was in the range of 2,500-10,000 μg/mL[7]. Similar trends were found in the extract of Malaysian propolis produced by *H. itama* and *G. thoracica,* which inhibited the growth of *S. aureus* better than Gram-negative (*E. coli* and *Salmonella typhi*). The variability in antibacterial activity against different bacterial strains is attributed to the diversity of bioactive compounds in propolis from various species of stingless bees. Literature suggests that the antimicrobial properties of propolis are related to phenolic and flavonoid compounds of varying polarities and their synergistic effects. These polar and lipophilic compounds, containing electronegative functional groups, e.g., carbonyl, amine, thiol, or hydroxyl groups, interacted with the cell wall and membrane of bacterial cells, resulting in the leakage of cellular components and, finally, cell death [3-5, 10, 11, 20]. In general, Gram-negative bacteria are more resistant to antibacterial agents than Gram-positive bacteria due to the presence of the outer membrane, which is comprised mainly of LPS that make Gram-negative bacteria more invulnerable to the antimicrobial agents as well as the function of the efflux pumps [21, 22]. In addition, the extracts from the propolis efficiently inhibited fungal species such as *Candida albicans* and *C. neoformans* with a MIC of 1.56 mg/mL while in stingless bee *Melipona beecheii* was able to inhibit the growth of *C. albicans* and induced dramatic changes in the structure and integrity of the cell wall [23]. A similar observation also found in propolis from *Tetragonisca fiebrigi* that was able to suppress the growth of *C. albicans* and *C. glabrata* [12]. This antifungal activity may also be provided by its phenolic and flavonoid compounds [24]. Similarly, results of antibacterial effect were observed in Malaysian *G. thoracica* honey, with the inhibition zones about 9 -12 mm for the tested population against *S. aureus* and *E. coli*, while the *G. thoracica* honey from Borneo (Sarawak) showed antibacterial against Gram negative bacteria which was in the range of ZOI about 12.3±0.21- 30±0.10 mm [25]. The antimicrobial efficacy observed in stingless bee propolis may be credited to the actions of flavonoids and other chemical constituents. Moreover, factors such as the extraction method, osmotic effect, or the properties of phytochemicals may also contribute to the antimicrobial activity of the propolis produced by stingless bee [3, 26].

3.3 antioxidant Activity

The scavenging potential of the propolis extracts against DPPH° is presented in **Figure 2**. DPPH° undergoes a conversion to DPPH-H upon accepting a hydrogen atom from phenolic compounds. The phenolic compound increases directly with the intensity of DPPH° [14,26,27]. The assessment demonstrated that the % DPPH scavenging was increased by increasing the propolis extract concentration. However, the more EEP added, the more the antioxidant activity maintained at approximately 80% radical scavenging activity, accounting for about 0.8 times of ascorbic acid. Comparing to a study carried out by Akhir et al. (2018) found that the propolis extract from *H. itama* in ethanol-solvent produced a higher antioxidant activity than in n-hexane and also showed a strong positive correlation with total phenolic and flavonoid content [26,28]. The potent antioxidant properties are related to the chemical composition of the propolis [3-5,11]. Furthermore, the antioxidant capacity of propolis produced by *G. thoracica, H. itama*, and *T. binghami* collected in Brunei,

which was extracted in ethanol, revealed varying total antioxidant capacities (TAC) with the highest TAC observed in *H. itama* (317.6 mgAAE/g), followed by *G. thoracica* (42.5 mgAAE/g) and *T. binghami* (12.3 mgAAE/g). In addition, the ethanol extract of propolis from *G. thoracica*, collected in Pattani province, Thailand, exhibited flavonoid, coumarin, saponin, terpenoid, steroid, and cardiac glycoside constituents, demonstrating an IC_{50} value of 262.43 μg/mL and a total phenolic content of 60.13 mg GAE/g extract [11]. Obviously, according to Figure 2 in this study, after the percentages of DPPH radical scavenging were calculated as IC_{50}, the results showed the IC_{50} of approximately 20 mg/mL obtained in EEP from *G. thoracica* while in ascorbic acid reached the IC_{50} less than 12.5 mg/mL. This obtained IC_{50} was almost similar value to *G. thoracica* ethanolic extract propolis that was collected in Serdang, Selangor, Malaysia (N 2° 58′ 45.84″ E 101° 41′ 51.72″), predominantly surrounded by medicinal plants from the *Simaroubaceae*, *Myrsinaceae*, *Primulaceae*, *Zingiberaceae*, *Acanthaceae* and *Lamiaceae* families [9]. Phenolic compounds and flavonoids found in plant constituents are recognized as potent free radical scavengers [27], indicating the significant role of phenolic compounds in the antioxidant activity of the propolis extract. Moreover, Idris et al. (2023) confirmed a strong correlation between total phenolic content (TPC), total flavonoid content (TFC), and IC_{50} of DPPH, indicating that the radical scavenging activity of propolis extract is influenced by the phenolic and flavonoid contents owing to the presence of aromatic hydroxyl groups, which are known for their effective electron accepting abilities [9].

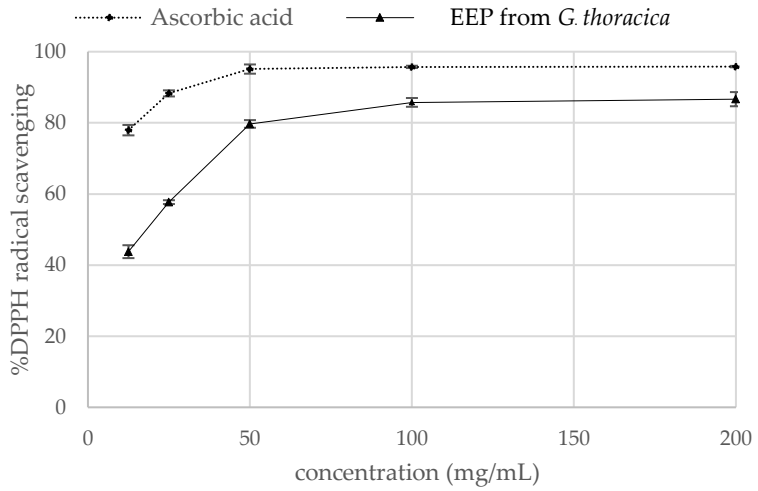


Figure 2. %DPPH° radical scavenging of the ethanolic propolis extract (EEP) of *G. thoracica* (ascorbic acid was used as a positive control) Values represent means ± SD of three independent experiments.

3.4 Anti-inflammatory Activity

NO inhibitory assay was used to test for anti-inflammatory properties of the propolis extract [17, 29]. This study demonstrated the NO production in RAW 264.7 macrophage cells treated with EEP from *G. thoracica* **(Figure 3)**. Interestingly, the anti-inflammatory activity of *G. thoracica* propolis in Thailand has not been investigated much. This study evaluated the ethanolic extract from *G. thoracica* propolis collected in Phatthalung province area by measuring NO production. In this case, LPS is used as an inflammation inducer to induce inflammation in RAW 264.7 cells. It stimulates the cells to produce pro-inflammatory cytokines and mediators, creating an inflammatory environment. Therefore, it is hypothesized that the LPS-induced inflammatory response decreases the levels of inflammatory markers such as NO production, indicating the anti-inflammatory activity of the EEP.

The results showed that at the lowest concentration tested (100 mg/mL), NO secretion in RAW 264.7 cells was reduced to basal levels, decreasing NO production by 72.9% in LPS-activated cells. However, increasing the EEP concentration from 200 to 800 mg/mL appeared to raise nitrite levels, which could potentially trigger an excessive inflammatory response. These may suggest that 100 mg/mL of EEP showed the most substantial reduction in nitrite concentration, indicating a strong anti-inflammatory effect at this

concentration. Still, for a precise IC_{50}, a consistent dose-response curve between 0 and 100 mg/mL is required. However, it can be explained by the non-linear relationship between propolis extract dosage and nitrite production. Lower doses of the extract might exhibit anti-inflammatory effects, while higher doses could potentially induce a pro-inflammatory response, resulting in increased nitrite production [30].

Similarly, the previous study in Brazilian red propolis showed that 50 μg/mL of propolis extract decreased NO production by 78% in LPS-activated RAW 264.7 macrophages cells [31] while in Sonoran propolis at a concentration of 10 μg/mL was able to decrease NO level between 86% and 95% [16]. These results suggested that propolis is supposed to be anti-inflammatory by inhibiting NO production in macrophages [16, 29]. NO is a signaling molecule in the inflammation process and is naturally produced in biological tissues. This may explain why propolis could reduce the levels of specific molecules, such as hydroxyarginine, an intermediate molecule in NO production [31, 32]. The anti-inflammatory mechanism of propolis is associated with its intricate chemical composition. Regardless, the propolis extract's phenolics and flavonoids are considered anti-inflammatory agents. For example, the flavonoids can inhibit the enzyme-inducible NO synthase (iNOS) by binding to the PPAR-γ receptor on macrophage cells [16, 30-33]. However, further studies may include a positive control such as Dexamethasone, a known standard for suppressing the expression of pro-inflammatory cytokines, e.g., TNF-α, IL-6, and IL-1β and certain enzymes such as iNOS and COX-2 by inhibiting transcription factors [34], to quantify the exact efficacy of the EEP affect on anti-inflammatory of the treated cells.

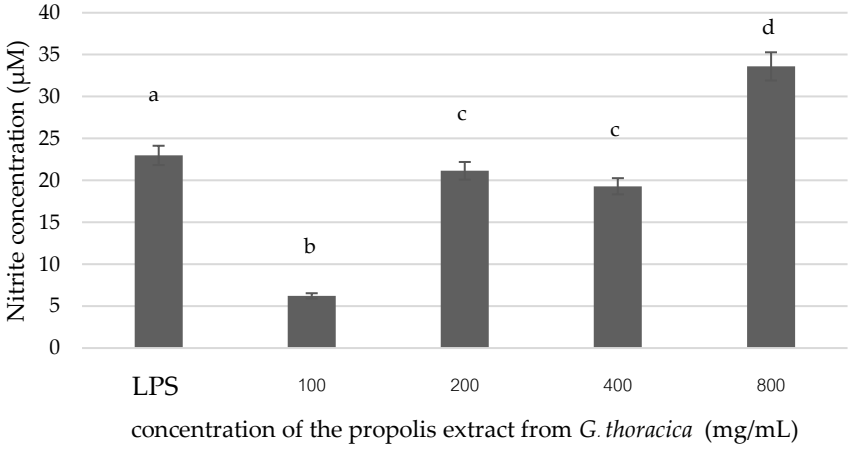


Figure 3. Effect of the ethanolic propolis extract of *G. thoracica* on nitrite levels (LPS was used as an inflammation inducer). Bars with different letters within the same concentration group indicate statistical differences ($p \leq 0.05$).

3.5 Identification and quantification of bioactive compounds in the propolis extract

The chemical constituents found in various types of stingless bee propolis primarily consist of phenolic and flavonoid compounds such as quercetin, vanillic acid, coumaric acid, and benzoic acid [3, 6, 8, 10, 12]. Phenolics are compounds developed by the secondary metabolism of plants [3, 8, 27]. Notably, there were few studies of the chemical composition of the propolis from *G. thoracica*. For example, Nazir et al. [6] studied the chemical constituents of *G. thoracica* propolis in Malaysia and successfully identified 30 new compounds from the ethanolic extract of propolis. While in Brunei, *G. thoracica* propolis contained aromatic acids, terpenes, flavonoids, and phenolic acids with hydroxyl functional groups based on FTIR analysis [6]. The highlights of the phytochemical compounds in the EEP from *G. thoracica* include the triterpenoid derivatives lupeol (25.42%) and β-amyrone (22.66%) as the dominant compounds. These are followed by pentacosane (6.33%), cis-9-eicosenol (4.23%), and the phenolic compound phenol 3-pentadactyl (3.86%), as presented in **Table 2**. Lupeol is a pentacyclic triterpenoid commonly found in the plant. It is used to reduce

inflammatory responses and has immunomodulating properties. Lupeol and its derivatives have a great potential to act as an inflammatory, anti-microbial, and anti-protozoal. Various studies have shown that the anti-inflammatory activity of lupeol through the modulation of p-38 pathways inhibits inflammation[35]. These results validate the importance and role of terpenes in providing bioactive properties to the composition of propolis. The Triterpenoids have been studied for their ability to regulate immune responses, reduce inflammation, and protect the cells from damage caused by oxidative stress [3-5, 16, 19], while the phenolic compounds, including flavonoids and phenolic acids, are known for their antioxidant properties, reducing inflammation, and modulating cellular signaling pathways [8-11, 30]. Although phenols and terpenes are commonly present in many types of propolis, their concentrations, proportions, and types differ among propolis varieties. These variations can be attributed to various extrinsic and intrinsic factors, including botanical sources, geographical location, extraction techniques, climate conditions, bee species, and the foraging preferences of each bee species [3, 5, 6, 12, 17]. Similar results to Nazir et al. [6] reported the EEP from Malaysian *G. thoracica* consisted of 1H-Pyrrole-2-carboxylic acid and 1-(2-hydroxy-2-phenylethyl) as major phenolic compounds followed by terpenoid and its derivatives [6].

Table 2. Phytochemical compounds in the ethanolic extract propolis of *G. thoracica* detected by GC-MS analysis

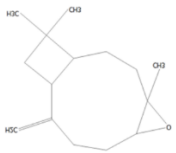
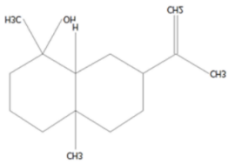
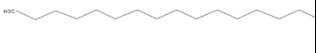
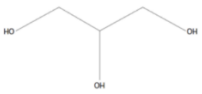
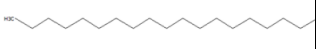
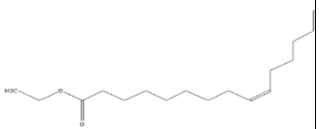
Retention time (min)	**Compound Names**	**Molecular structure**	**Classification of Phytochemicals**	**% of total component area***
17.4939	(-)-5-Oxatricyclo[8.2.0.0(4,6)]dodecane,12- trimethyl-9-methylene-[1R- (1R*,4R*,6R*,10S*)]-	 [$C_{15}H_{24}O$]	tricyclic diterpenoids	1.04
23.5475	Neointermedeol	 [$C_{15}H_{26}O$]	sesquiterpene alcohol	1.35
23.8899	Hexadecanoic acid, ethyl ester	 [$C_{18}H_{36}O_2$]	ethyl ester	2.67
24.711	Glycerol	 [$C_3H_8O_3$]	polyol compound	1.36
24.8731	Tricosane	 [$C_{23}H_{48}$]	alkane hydrocarbon	1.00
27.9522	9-Octadecenoic acid (Z)-, ethyl ester	 [$C_{20}H_{38}O_2$]	ethyl ester	2.49

Table 2. Phytochemical compounds in the ethanolic extract propolis of *G. thoracica* detected by GC-MS analysis **(Continue)**

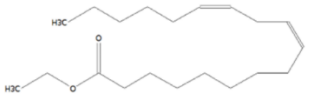
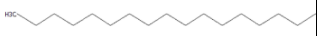
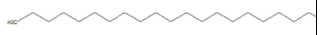
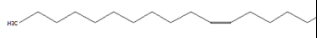
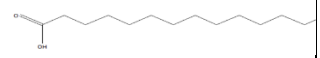
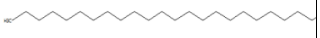
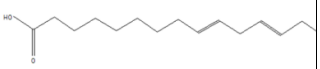
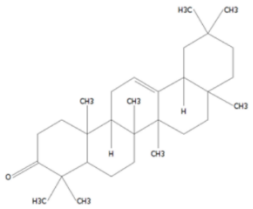
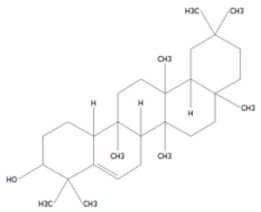
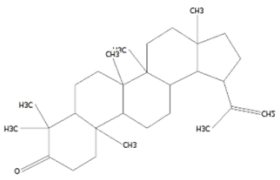
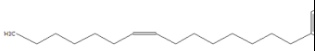
Retention time (min)	Compound Names	Molecular structure	Classification of Phytochemicals	% of total component area*
28.7355	Linoleic acid, ethyl ester	 [$C_{20}H_{36}O_2$]	ethyl ester	2.63
29.6809	1-Octadecanol	 [$C_{18}H_{38}O$]	fatty alcohol	1.47
31.8903	Pentacosane	 [$C_{25}H_{52}$]	alkane hydrocarbon	6.63
34.1700	Eicosen-1-ol, cis-9-	 [$C_{20}H_{40}O$]	fatty alcohol	4.23
36.5174	Palmitic acid	 [$C_{16}H_{32}O_2$]	fatty acid	1.63
36.8494	Nonacosane	 [$C_{29}H_{60}$]	alkane hydrocarbon	1.92
44.4285	Linoleic acid	 [$C_{18}H_{32}O_2$]	fatty acid	1.77
48.0478	.beta.-Amyrone	 [$C_{30}H_{48}O$]	triterpenoid derivative	22.66
51.2512	Glutinol	 [$C_{30}H_{50}O$]	triterpene alcohol	1.37
54.7193	Phenol, 3-pentadecyl-	 $C_{21}H_{36}O$	phenolic compound	3.86

Table 2. Phytochemical compounds in the ethanolic extract propolis of *G. thoracica* detected by GC-MS analysis (Continue)

Retention time (min)	Compound Names	Molecular structure	Classification of Phytochemicals	% of total component area*
55.4216	Lup-20(29)-en-3-one	 $C_{30}H_{48}O$	triterpenoid derivative	25.42
56.8207	(Z)-3-(pentadec-8-en-1-yl)phenol	 $C_{21}H_{34}O$	phenolic compound	2.35

* The GC-MS analysis data here presented only the compound content ≥1 % of the total component area

In addition, according to an observation of major plants and pollens in the propolis collection site, e.g., Pantae community enterprise, Khuan Khanun district, Phatthalung province, Thailand, it was found that the oil palm tree *(Elaeis guineensis)* and the rubber tree *(Hevea brasiliensis)* were dominant plant species. These results corresponded with the phytochemical compounds described above, especially the triterpene derivatives derived from isoprene units obtained from the latex, bark, leaves, or other parts of the rubber tree plant. Although the stingless bee species are similar, the plant sources surrounding them may differ, resulting in the bioactives' variation [3, 5]. In addition, a small amount of fatty acid compounds, e.g., linoleic acid and palmitic acid, were also identified in this study. In summary, the chemical composition of stingless bee propolis comprises aromatic acids, phenolic compounds, alcohols, terpenes, and sugar as the dominant compounds [3, 5, 13 ,28]. According to the literature reviews, up to 16 species of stingless bee-producing propolis that were harvested in Malaysia, Brazil, Mexico, Thailand, the Philippines, Vietnam, and India reveal the largest amount of phenolic compounds (e.g., *p*-coumaric acid and gallic acid). However, the chemical compositions vary significantly when comparing propolis from the same species of stingless bees. These differences are due to variations in the identification methods, collection locations, and collection periods for the propolis [3, 5]. The hypothesis suggests that the bioactive compounds present in propolis may interact synergistically, combining their complementary mechanisms of action to enhance the beneficial biotherapeutic potential of this valuable bee product [3-5,9]. The synergistic effects between the various bioactive components in propolis often occur. For example, antibacterial synergy by different compounds in propolis may target different aspects of bacterial physiology, such as lupeol. It can integrate into the bacterial cell membrane, causing disruption of its integrity and has been shown to inhibit certain bacterial enzymes, in addition, can interfere the biofilm formation [35, 36], while β-amyrone may inhibit the synthesis of bacterial cell walls and also can affect key metabolic pathways in bacteria including the inhibition of the enzymes lipase, α-glucosidase, and α-amylase, disrupting their growth and replication [37]. Therefore, both Lupeol and β-anyone may have a synergistic effect as their different mechanisms of action can target multiple bacterial processes simultaneously. Additionally, both compounds in EEP contribute antioxidant activity by neutralizing free radicals and reducing oxidative stress. Notably, β-Amyrone may influence endogenous antioxidant enzyme activity to further enhance its antioxidant potential [38]. Finally, these triterpenoid derivatives revealed capability of interacting with multiple molecular targets, affecting and modulating the inflammation process, carcinogenesis and cellular stress response by inhibit the production of pro-inflammatory cytokines such as TNF-α, IL-1β, IL-6,and may suppress the activation of NF-κB signaling pathway, which is a central regulator of inflammation, resulted in lower expression of inflammation-related genes [30, 36, 37]. Therefore, the extracted propolis acts through a combination of

mechanisms to exert its antibacterial, antioxidant, and anti-inflammatory effects, primarily due to its rich content of bioactive compounds such as triterpenoid derivatives as mentioned above.

4. Conclusions

This study unveils the phytochemical composition of ethanolic extract propolis (EEP) sourced from *G. thoracica*, harvested in Phatthalung province, Southern Thailand, revealing its significant influence on biological activities such as antibacterial, antioxidant, and anti-inflammatory capabilities. The primary constituents found in the EEP were triterpenoids and phenolic compounds, recognized for their potential as bioactive. Remarkably, the EEP exhibited remarkable outcomes, including up to 80% DPPH radical scavenging activity and a notable 73% reduction in NO production in LPS-activated RAW 264.7 macrophage cells, accompanied by reduced anti-inflammatory effects and robust antibacterial activity with moderate to strong inhibition. These findings position the propolis of *G. thoracica* as a promising therapeutic agent and a safe food supplement. Additionally, this study offers novel insights into the phytochemical and biological activities of extracted propolis from the Phatthalung region of Thailand.

5. Acknowledgements

We would like to sincerely thank Pantae Community Enterprise, Khuan Khanun district, Phatthalung province, Thailand, for supporting the raw propolis of *G. thoracica* through the experiment and thank Microbial Technology for Agriculture, Food and Environment Research Center, Faculty of Science and Digital Innovation, Thaksin University for supporting the laboratory facilities and equipment.

Author Contributions: Conceptualization, W.C. and J.M.; methodology, W.C., K.B., M.L.; investigation, K.B., and P.P.; data curation, W.C., K.B., P.P.; writing—original draft preparation, W.C.; writing—review and editing; W.C., K.B., M.L., J.M.; supervision, W.C., M.L., J.M. All authors have read and agreed to the published version of the manuscript." Authorship must be limited to those who have contributed substantially to the work reported.

Funding: This work was financially supported by the National Higher Education, Science, Research and Innovation Policy Council, Thaksin University (Research project grant no. FF-2567A10512038), Fiscal Year 2024.

Conflicts of Interest: The authors declare no conflict of interest

References

[1] Rasmussen, C.; Cameron, S. A. A molecular phylogeny of the Old World stingless bees (Hymenoptera: Apidae: Meliponini) and the non-monophyly of the large genus *Trigona*. *Systematic Entomology,* **2007**, *32*(1), 26-39.

[2] Engel, M. S.; Rasmussen, C.; Ayala, R.; de Oliveira, F. F. Stingless bee classification and biology (Hymenoptera, Apidae): A review, with an updated key to genera and subgenera. *ZooKeys,* **2023**, *1172*, 239.

[3] Rozman, A. S.; Hashim, N.; Maringgal, B.; Abdan, K. A comprehensive review of stingless bee products: Phytochemical composition and beneficial properties of honey, propolis, and pollen. *Applied Sciences,* **2022**, *12*(13), 6370.

[4] Rocha, V. M.; Portela, R. D.; Dos Anjos, J. P.; De Souza, C. O.; Umsza-Guez, M. A. Stingless bee propolis: Composition, biological activities and its applications in the food industry. *Food Production, Processing and Nutrition,* **2023**, *5*(1), 29.

[5] Chuttong, B.; Lim, K.; Praphawilai, P.; Danmek, K.; Maitip, J.; Vit, P.; Wu, M.-C.; Ghosh, S.; Jung, C.; Burgett, M. Exploring the functional properties of propolis, geopropolis, and cerumen, with a special emphasis on their antimicrobial effects. *Foods,* **2023**, *12*(21), 3909.

[6] Nazir, H.; Shahidan, W. N. S.; Ibrahim, H. A.; Tuan Ismail, T. N. N. Chemical constituents of Malaysian *Geniotrigona thoracica* propolis. *Pertanika Journal of Tropical Agricultural Science,* **2018**, *41*(3).

[7] Zullkiflee, N.; Taha, H.; Abdullah, N. A.; Hashim, F.; Usman, A. Antibacterial and antioxidant activities of ethanolic and water extracts of stingless bees *Tetrigona binghami*, *Heterotrigona itama*, and *Geniotrigona thoracica* propolis found in Brunei. *Philippine Journal of Science,* **2022**, *151*, 1455-1462.

[8] Abdullah, N. A.; Zullkiflee, N.; Zaini, S. N. Z.; Taha, H.; Hashim, F.; Usman, A. Phytochemicals, mineral contents, antioxidants, and antimicrobial activities of propolis produced by Brunei stingless bees *Geniotrigona thoracica*, *Heterotrigona itama*, and *Tetrigona binghami*. *Saudi Journal of Biological Sciences,* **2020**, *27*(11), 2902-2911.

[9] Idris, L.; Adli, M. A.; Yaacop, N. N.; Zohdi, R. M. Phytochemical screening and antioxidant activities of *Geniotrigona thoracica* propolis extracts derived from different locations in Malaysia. *Malaysian Journal of Fundamental and Applied Sciences,* **2023**, *19*(6), 1023-1032.

[10] Šuran, J.; Cepanec, I.; Mašek, T.; Radić, B.; Radić, S.; Tlak Gajger, I.; Vlainić, J. Propolis extract and its bioactive compounds—From traditional to modern extraction technologies. *Molecules,* **2021**, *26*(10), 2930.

[11] Bankova, V.; Popova, M. Propolis of stingless bees: A promising source of biologically active compounds. *Pharmacognosy Reviews,* **2007**, *1*(1).

[12] Campos, J. F.; Santos, U. P. d.; Rocha, P. d. S. d.; Damião, M. J.; Balestieri, J. B. P.; Cardoso, C. A. L.; Paredes-Gamero, E. J.; Estevinho, L. M.; de Picoli Souza, K.; Santos, E. L. d. Antimicrobial, antioxidant, anti-inflammatory, and cytotoxic activities of propolis from the stingless bee *Tetragonisca fiebrigi* (Jatai). *Evidence-Based Complementary and Alternative Medicine,* **2015**, *2015*(1), 296186.

[13] Bees, M. S. (n.d.). Antibacterial and phenolic content of propolis produced by two Malaysian stingless bees, *Heterotrigona itama* and *Geniotrigona thoracica*.

[14] Meechai, I.; Chelong, I. Total phenolic content and anti-radical activity of stingless bee honey at different harvesting times. *Progress in Applied Science and Technology,* **2018**, *8*(2), 65-72.

[15] Bhaigybati, T.; Sanasam, S.; Gurumayum, J.; Bag, G.; Singh, L. R.; Devi, P. G. Phytochemical profiling, antioxidant activity, antimicrobial activity, and GC-MS analysis of *Ipomoea aquatica* Forsk collected from EMA market, Manipur. *Journal of Pharmacognosy and Phytochemistry,* **2020**, *9*(1), 2335-2342.

[16] Mendez-Encinas, M. A.; Valencia, D.; Ortega-García, J.; Carvajal-Millan, E.; Díaz-Ríos, J. C.; Mendez-Pfeiffer, P.; Soto-Bracamontes, C. M.; Garibay-Escobar, A.; Alday, E.; Velazquez, C. Anti-inflammatory potential of seasonal Sonoran propolis extracts and some of their main constituents. *Molecules,* **2023**, *28*(11), 4496.

[17] Mohiuddin, I.; Kumar, T. R.; Zargar, M. I.; Wani, S. U. D.; Mahdi, W. A.; Alshehri, S.; Alam, P.; Shakeel, F. GC-MS analysis, phytochemical screening, and antibacterial activity of *Cerana indica* propolis from Kashmir region. *Separations,* **2022**, *9*(11), 363.

[18] Gupta, A.; Naraniwal, M.; Kothari, V. Modern extraction methods for the preparation of bioactive plant extracts. *International Journal of Applied and Natural Sciences,* **2012**, *1*(1), 8-26.

[19] Silva, J. C.; Rodrigues, S.; Feás, X.; Estevinho, L. M. Antimicrobial activity, phenolic profile, and role in the inflammation of propolis. *Food and Chemical Toxicology,* **2012**, *50*(5), 1790-1795.

[20] Cushnie, T. T.; Hamilton, V. E.; Lamb, A. J. Assessment of the antibacterial activity of selected flavonoids and consideration of discrepancies between previous reports. *Microbiological Research,* **2003**, *158*(4), 281-289.

[21] Epand, R. M.; Walker, C.; Epand, R. F.; Magarvey, N. A. Molecular mechanisms of membrane targeting antibiotics. *Biochimica et Biophysica Acta (BBA)-Biomembranes,* **2016**, *1858*(5), 980-987.

[22] Breijyeh, Z.; Jubeh, B.; Karaman, R. Resistance of gram-negative bacteria to current antibacterial agents and approaches to resolve it. *Molecules,* **2020**, *25*(6), 1340.

[23] Hau-Yama, N. E.; Magaña-Ortiz, D.; Oliva, A.; Ortiz-Vázquez, E. Antifungal activity of honey from stingless bee *Melipona beecheii* against *Candida albicans*. *Journal of Apicultural Research,* **2020**, *59*(1), 12-18.

[24] Shehu, A.; Ismail, S.; Rohin, M. A. K.; Harun, A.; Abd Aziz, A.; Haque, M. Antifungal properties of Malaysian Tualang honey and stingless bee propolis against *Candida albicans* and *Cryptococcus neoformans*. *Journal of Applied Pharmaceutical Science,* **2016**, *6*(2), 044-050.

[25] Tuksitha, L.; Chen, Y.-L. S.; Chen, Y.-L.; Wong, K.-Y.; Peng, C.-C. Antioxidant and antibacterial capacity of stingless bee honey from Borneo (Sarawak). *Journal of Asia-Pacific Entomology,* **2018**, *21*(2), 563-570.

[26] Akhir, R. A. M.; Bakar, M. F. A.; Sanusi, S. B. Antioxidant and antimicrobial potential of stingless bee (*Heterotrigona itama*) by-products. *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences,* **2018**, *42*(1), 72-79.

[27] Abbas, A.; Naqvi, S. A. R.; Rasool, M. H.; Noureen, A.; Mubarik, M. S.; Tareen, R. B. Phytochemical analysis, antioxidant and antimicrobial screening of *Seriphidium oliverianum* plant extracts. *Dose-Response,* **2021**, *19*(1), 15593258211004739.

[28] Akhir, R. A. M.; Bakar, M. F. A.; Sanusi, S. B. Antioxidant and antimicrobial activity of stingless bee bread and propolis extracts. In *AIP Conference Proceedings*. AIP Publishing. **2017**

[29] Parichatikanond, W.; Mangmool, S.; Chewchinda, S.; Hirunpanich, V.; Vongsak, B. Anti-inflammatory activity of propolis extract from the stingless bee, *Tetragonula pagdeni*, in mangosteen orchard. *The Thai Journal of Pharmaceutical Sciences,* **2024**, *47*(3), 6.

[30] Zulhendri, F.; Lesmana, R.; Tandean, S.; Christoper, A.; Chandrasekaran, K.; Irsyam, I.; Suwantika, A. A.; Abdulah, R.; Wathoni, N. Recent update on the anti-inflammatory activities of propolis. *Molecules,* **2022**, *27*(23), 8473.

[31] Bueno-Silva, B.; Kawamoto, D.; Ando-Suguimoto, E. S.; Alencar, S. M.; Rosalen, P. L.; Mayer, M. P. Brazilian red propolis attenuates inflammatory signaling cascade in LPS-activated macrophages. *PLOS ONE,* **2015**, *10*(12), e0144954.

[32] Alanazi, S.; Alenzi, N.; Fearnley, J.; Harnett, W.; Watson, D. G. Temperate propolis has anti-inflammatory effects and is a potent inhibitor of nitric oxide formation in macrophages. *Metabolites,* **2020**, *10*(10), 413.

[33] Liang, Y.-C.; Tsai, S.-H.; Tsai, D.-C.; Lin-Shiau, S.-Y.; Lin, J.-K. Suppression of inducible cyclooxygenase and nitric oxide synthase through activation of peroxisome proliferator-activated receptor-γ by flavonoids in mouse macrophages. *FEBS Letters,* **2001**, *496*(1), 12-18.

[34] Walker, G.; Pfeilschifter, J.; Kunz, D. Mechanisms of suppression of inducible nitric-oxide synthase (iNOS) expression in interferon (IFN)-γ-stimulated RAW 264.7 cells by dexamethasone: Evidence for glucocorticoid-induced degradation of iNOS protein by calpain as a key step in post-transcriptional regulation. *Journal of Biological Chemistry,* **1997**, *272*(26), 16679-16687.

[35] Sharma, N.; Palia, P.; Chaudhary, A.; Verma, K.; Kumar, I. A review on pharmacological activities of lupeol and its triterpene derivatives. *Journal of Drug Delivery and Therapeutics,* **2020**, *10*(5), 325-332.

[36] Gallo, M. B.; Sarachine, M. J. Biological activities of lupeol. *International Journal of Biomedical and Pharmaceutical Sciences,* **2009**, *3*(1), 46-66.

[37] Ferreira, R. G.; Silva Junior, W. F.; Veiga Junior, V. F.; Lima, Á. A.; Lima, E. S. Physicochemical characterization and biological activities of the triterpenic mixture α, β-amyrenone. *Molecules,* **2017**, *22*(2), 298.

[38] Karen Cardoso, B.; Line Marko de Oliveira, H.; Zonta Melo, U.; Mariano Fernandez, C. M.; Franco de Araújo Almeida Campo, C.; Gonçalves, J. E.; Laverde Jr, A.; Barion Romagnolo, M.; Andrea Linde, G.; Cristiani Gazim, Z. Antioxidant activity of α and β-amyrin isolated from *Myrcianthes pungens* leaves. *Natural Product Research,* **2020**, *34*(12), 1777-1781.

ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

ASEAN
Journal of Scientific and Technological Reports
Online ISSN: 2773-8752

Article type

Type of the Paper (Article, Review, Communication, etc.) *about 8,000 words maximum*

Title (Palatino Linotype 18 pt, bold)

Firstname Lastname[1], Firstname Lastname[2] and Firstname Lastname[2*]

[1] Affiliation 1; e-mail@e-mail.com
[2] Affiliation 2; e-mail@e-mail.com
* Correspondence: e-mail@e-mail.com; (one corresponding authors, add author initials)

Citation:
Lastname, F.; Lastname, F.; Lastname, F. Title. *ASEAN J. Sci. Tech. Report.* **2023**, *26*(X), xx-xx. https://doi.org/10.55164/ajstr.vxxix.xxxxxx

Article history:
Received: date
Revised: date
Accepted: date
Available online: date

Abstract: A single paragraph of about 400 words maximum. Self-contained and concisely describe the reason for the work, methodology, results, and conclusions. Uncommon abbreviations should be spelled out at first use. We strongly encourage authors to use the following style of structured abstracts, but without headings: (1) Background: Place the question addressed in a broad context and highlight the purpose of the study; (2) Methods: briefly describe the main methods or treatments applied; (3) Results: summarize the article's main findings; (4) Conclusions: indicate the main conclusions or interpretations.

Keywords: keyword 1; keyword 2; keyword 3 (List three to ten pertinent keywords specific to the article yet reasonably common within the subject discipline.)

1. Introduction

The introduction should briefly place the study in a broad context and highlight why it is crucial. It should define the purpose of the work and its significance. The current state of the research field should be carefully reviewed and critical publications cited. Please highlight controversial and diverging hypotheses when necessary. Finally, briefly mention the main aim of the work. References should be numbered in order of appearance and indicated by a numeral or numerals in square brackets—e.g., [1] or [2, 3], or [4–6]. See the end of the document for further details on references.

2. Materials and Methods

The materials and methods should be described with sufficient details to allow others to replicate and build on the published results. Please note that your manuscript's publication implicates that you must make all materials, data, computer code, and protocols associated with the publication available to readers. Please disclose at the submission stage any restrictions on the availability of materials or information. New methods and protocols should be described in detail, while well-established methods can be briefly described and appropriately cited.

Interventional studies involving animals or humans, and other studies that require ethical approval, must list the authority that provided approval and the corresponding ethical approval code.

2.1 Subsection

2.1.1. Subsubsection

3. Results and Discussion

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.Authors should discuss the results and how they can be interpreted from previous studies and the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

3.1. Subsection

3.1.1. Subsubsection

3.2. Figures, Tables, and Schemes

All figures and tables should be cited in the main text as Figure 1, Table 1, etc.

Figure 1. This is a figure. Schemes follow the same formatting.

Table 1. This is a table. Tables should be placed in the main text near the first time they are cited.

Title 1	Title 2	Title 3
entry 1	data	data
entry 2	data	data [1]

[1] Table may have a footer.

3.3. Formatting of Mathematical Components

This is example 1 of an equation:

$$a = 1, \tag{1}$$

The text following an equation need not be a new paragraph. Please punctuate equations as regular text. This is example 2 of an equation:

$$a = b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t + u \tag{2}$$

The text following an equation need not be a new paragraph. Please punctuate equations as regular text. The text continues here.

4. Conclusions

Concisely restate the hypothesis and most important findings. Summarize the significant findings, contributions to existing knowledge, and limitations. What are the future directions? Conclusions MUST be well stated, linked to original research question & limited to supporting results.

5. Acknowledgements

Should not be used to acknowledge funders – funding will be entered as a separate. As a matter of courtesy, we suggest you inform anyone whom you acknowledge.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript." Please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: Please add: "This research received no external funding" or "This research was funded by NAME OF FUNDER, grant number XXX" and "The APC was funded by XXX". Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding. Any errors may affect your future funding.

Conflicts of Interest: Declare conflicts of interest or state "The authors declare no conflict of interest." Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript, or in the decision to publish the results must be declared in this section. If there is no role, please state "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results".

References

References must be numbered in order of appearance in the text (including citations in tables and legends) and listed individually at the end of the manuscript. We recommend preparing the references with a bibliography software package, such as EndNote, ReferenceManager to avoid typing mistakes and duplicated references. Include the digital object identifier (DOI) for all references where available.

Citations and references in the Supplementary Materials are permitted provided that they also appear in the reference list here.

In the text, reference numbers should be placed in square brackets [ ] and placed before the punctuation; for example [1], [1–3] or [1, 3]. For embedded citations in the text with pagination, use both parentheses and brackets to indicate the reference number and page numbers; for example [5] (p. 100), or [6] (pp. 101–105).

Using the American Chemical Society (ACS) referencing style

[1] Author 1, A.B.; Author 2, C.D. Title of the article. *Abbreviated Journal Name* Year, *Volume*, page range.

[2] Author 1, A.; Author 2, B. Title of the chapter. In *Book Title*, 2nd ed.; Editor 1, A., Editor 2, B., Eds.; Publisher: Publisher Location, Country. **2007**, Volume *3*, pp. 154–196.

[3] Author 1, A.; Author 2, B. *Book Title*, 3rd ed.; Publisher: Publisher Location, Country, **2008**, pp. 154–196.
[4] Author 1, A.B.; Author 2, C. Title of Unpublished Work. *Abbreviated Journal Name* stage of publication (under review; accepted; in press).
[5] Author 1, A.B. (University, City, State, Country); Author 2, C. (Institute, City, State, Country). Personal communication, 2012.
[6] Author 1, A.B.; Author 2, C.D.; Author 3, E.F. Title of Presentation. In Title of the Collected Work (if available), Proceedings of the Name of the Conference, Location of Conference, Country, Date of Conference; Editor 1, Editor 2, Eds. (if available); Publisher: City, Country, Year (if available); Abstract Number (optional), Pagination (optional).
[7] Author 1, A.B. Title of Thesis. Level of Thesis, Degree-Granting University, Location of University, Date of Completion.
[8] Title of Site. Available online: URL (accessed on Day Month Year).

Reviewers sugestion

1. Name, Address, e-mail
2. Name, Address, e-mail
3. Name, Address, e-mail
4. Name, Address, e-mail

URL link:

Notes for Authors >>
https://drive.google.com/file/d/1r0zegnIVeQqe4iLQyT1xDEIinNggINPD/view?usp=sharing
https://drive.google.com/file/d/1r0zegnIVeQqe4iLQyT1xDEIinNggINPD/view?usp=sharing

Online Submissions >> https://ph02.tci-thaijo.org/index.php/tsujournal/user/register

Current Issue >> https://ph02.tci-thaijo.org/index.php/tsujournal/issue/view/16516

AJSTR Publication Ethics and Malpractice >> https://ph02.tci-thaijo.org/index.php/tsujournal/ethics

Journal Title Abbreviations >> http://library.caltech.edu/reference/abbreviations

ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

ASEAN
Journal of Scientific and Technological Reports
Online ISSN:2773-8752

ASEAN Journal of Scientific and Technological Reports (AJSTR)
Institute of Research and Innovation, Thaksin University, Phatthalung Campus
222 Moo. 2, Baan Prao Sub-District, Papayom District, Phatthalung, 93210, Thailand
Tel. +668 1540 7304