**Edelweiss Applied Science and Technology**
*ISSN: 2576-8484*
*Vol. 9, No. 6, 1897-1904*
*2025*
*Publisher: Learning Gate*
*DOI: 10.55214/25768484.v9i6.8275*


# Detection and statistic analysis of formal error spelling in Indonesian

Emmy Erwina[1], Tommy[2*], Mayasari[3]
[1,2,3]Universitas Harapan Medan, Indonesia; tomshirakawa@gmail.com (T.)

**Abstract:** Using language with correct pronunciation, especially for teachers and lecturers, can help foster language awareness among the younger generation. The high use of non-standard language and incorrect pronunciation of words in the younger generation is a serious concern, considering that culture and trends in the current era of globalization have caused language standards to fade. This study detects and classifies errors in the pronunciation of standard words in Indonesian, using teachers at the teacher level and lecturers at the Yayasan Pendidikan Harapan Medan as respondents. The data on the classification of pronunciation errors were obtained from the respondents' teaching video recordings. The detection results showed that the most common pronunciation errors made by the respondents were vowel errors, accounting for 59.8%, followed by consonant errors at 25.2%, and diphthong errors at 15%. Statistical analysis indicates that the type of teacher attribute influences the pronunciation errors made. The teacher's age attribute has the least effect on pronunciation errors.

***Keywords:*** *Indonesian, Pronunciation, Statistic analysis, Teachers and lectures.*

## 1. Introduction

Indonesian is the official language used in the Unitary State of the Republic of Indonesia. Over time, the use of Indonesian has experienced a shift caused by various factors such as ethnicity and culture, trends and foreign languages [1]. These influences have led to the use of non-standard words both in formal and non-formal communication. Standard Language is a conventional language based on the agreement of a linguistic community [2]. Standard and non-standard languages can be determined from the words used. Standard words in Indonesian can be defined as words used in accordance with standardized rules or guidelines [3]. Indonesian Non-standard words can be defined by words used that are not in accordance with the rules of the Indonesian language, both in terms of writing, pronunciation and sentence composition [4]. The phenomenon of the use of non-standard words in language can be found in several studies, such as the widespread use of slang by students [5] and communication in public spaces [6].

Language errors among students are considered quite high at this time, at one of the Senior High Schools in Indonesia, namely SMA Negeri 1 Sukasada, the results of language errors were 373 sentences out of 896 sentences where the percentage of errors was 41.63%. Of this number, 16.33% of pronunciation or pronunciation errors were obtained [7]. Language errors at the student level can also be found in several studies such as those conducted by Ariyana and Anggraini [8]; Supriadin [9] and Alber and Hermaliza [10].

The language used by teachers and lecturers can have an influence on the quality of students. Errors in pronunciation of language made by the teacher can set a bad example for students. The practice of mispronouncing the standard language by teachers and lecturers that continues to be allowed will result in the loss of good and correct language habits as can be seen in several studies [11]. The continuous use of non-standard language in students causes students to get used to and prefer teaching using non-standard language compared to teaching using standard language. This phenomenon can be seen in the research conducted by Dewi Purwati which was obtained from the group of students of the English


**History:** Received: 21 April 2025; Revised: 26 May 2025; Accepted: 30 May 2025; Published: 21 June 2025
* Correspondence: tomshirakawa@gmail.com

Study Program of IAIN Bone [12]. In research conducted by Al-Nawrasy [13] it shows the correlation of originality from the teacher to student achievement in the context of EFL learning [13].

Detection of standard language pronunciation errors, especially in Indonesian, can be an early stage in an effort to increase awareness and habits of using standard language, especially for educators such as teachers and lecturers so that they can transmit good language habits to students. Detection of pronunciation errors can be done by analysing the transcript of text characters generated from the sound of pronunciation (phonetics). Pronunciation errors contained in the text resulting from conversational sounds can be detected using several approaches such as statistical approaches [14] phonological and orthographic similarities [15] n-gram frequencies [16] and various other approaches.

Bigram is a technique part of n-gram which is used to perform frequency analysis on text. In some cases, bigrams can be used to distinguish words and non-words [17] text representation and classification [18] machine translation [19] and various other applications. Rana, et al. [20] apply bigrams and trigrams in the detection and correction of word errors in Bangla which gives an accuracy of 96% [20]. Other bigram applications in the detection and correction of word errors give quite good results [21].

The main objective of this study is to obtain statistical information on standard pronunciation errors in Indonesian, especially for educators such as teachers and lecturers. The high number of conversation transcripts obtained from respondents causes the need for a model that can detect standard pronunciation errors in Indonesian so that statistical information on standard pronunciation errors by teachers and lecturers can be obtained efficiently. Seeing the pretty good performance of bigram from some of the previously mentioned studies, this study uses the bigram method in the error detection process. The error words obtained will then be categorized based on the type of error so that statistical information can then be obtained that can describe the phenomenon of using non-standard language by educators such as teachers and lecturers.

## 2. Research Methodology

This study uses a quantitative research approach which is carried out by detecting and analysing standard pronunciation errors in Indonesian for teachers and lecturers. The design of this study broadly adopts previous research conducted by Nurgiyantoro, et al. [22]. This research begins with observation and data collection followed by identification of standard language pronunciation errors that will be used for statistical analysis. This study uses text data obtained from teaching videos of respondents who are teachers from the educational institution Yayasan Pendidikan Harapan Medan consisting of elementary, middle, and high school teachers and lecturers of bachelor degree. The teaching videos used as source data were selected from teachers and lecturers who represented each level of education. The number of teaching videos obtained from respondents can be seen in Table 1. Text transcripts of teaching videos obtained from respondents are then compiled by taking into account the articulatory phonetic aspects [23] where the characters of the words obtained follow how the sound is pronounced.

**Table 1.**
Respondent Types and Number of Teaching Videos.

| No. | Educators Type | Education Category | Total |
|---|---|---|---|
| 1 | Teachers | Elementary | 6 |
| 2 | | Middle | 6 |
| 3 | | High | 3 |
| 4 | Lecturers | Faculty of Technique and Computer | 15 |
| | | Faculty of Language and Communication | 4 |

Bigram analysis used in the process of detecting errors in pronunciation of standard language words in this study adopted the best correction candidate approach [24, 25] using a standard word dictionary obtained from the Big Indonesian Dictionary Edition. 5 [26]. The standard word candidate of a word that is the result of a mispronunciation is obtained using Equation 1.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 9, No. 6: 1897-1904, 2025*
*DOI: 10.55214/25768484.v9i6.8275*

$$w_b = arg \max_{w_i \in C_{(s)}} P(s|w_i)P(w_i)$$

Where P (s | w_i ) is the probability of the string s against candidate w_i and P(w_i ) is the probability of candidate w_i in the dictionary W. After the best correction candidate word is obtained, the type of pronunciation error can be identified based on the character difference between the input word and the word best correction candidate obtained. In this study, word pronunciation errors are classified into three errors, namely vowels, consonants and diphthongs [1].

Each word error will be mapped back into the dataset equipped with additional attributes obtained from the respondent's information which can be seen in table 2. The statistical analysis that will be used in this study is a one-way ANOVA analysis [27] on each respondent's attribute to the error class using SPSS.

**Table 2.**
Dataset Attributes.

| No. | Attribute | Values |
|---|---|---|
| 1 | Gender | Male, Female |
| 2 | Age | 0, .. , 100 |
| 3 | Educators Type | Teacher, Lecturer |
| 4 | Education Category | Elementary, Middle, High, Bachelor |
| 5 | Error Class | Vocal, Diphthong, Consonant |

## 3. Results and Discussion

### *3.1. Result*

The words obtained from the transcript of the teaching video were processed using the previously mentioned error detection model. The words detected as misspellings were then classified into vowel, diphthong and consonant error classes. The number of word errors in each error class can be seen in table 3. The results of the classification process show that vocal errors dominate overall pronunciation errors compared to other errors where 59.8% of identified pronunciation errors are vocal error classes. Meanwhile, the diphthong error is the least, which is 15%, followed by a consonant error of 25.2%.

**Table 3.**
Error Detection Results.

| No. | Attributes Values | Error Class | | |
|---|---|---|---|---|
| | | Vocal | Diphthong | Consonant |
| 1 | Gender | | | |
| | Male | 35 | 12 | 15 |
| | Female | 29 | 4 | 12 |
| 2 | Age | | | |
| | 28 - 34 | 27 | 4 | 11 |
| | 34 - 40 | 12 | 3 | 9 |
| | 40 - 46 | 25 | 9 | 7 |
| 3 | Educators | | | |
| | Teacher | 41 | 10 | 14 |
| | Lecturer | 23 | 6 | 13 |
| | **Error Class** | **Count** | | |
| | **Vocal** | 64 | | |
| | **Diphthong** | 16 | | |
| | **Consonant** | 27 | | |
| | **Total** | 107 | | |

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 9, No. 6: 1897-1904, 2025*
*DOI: 10.55214/25768484.v9i6.8275*

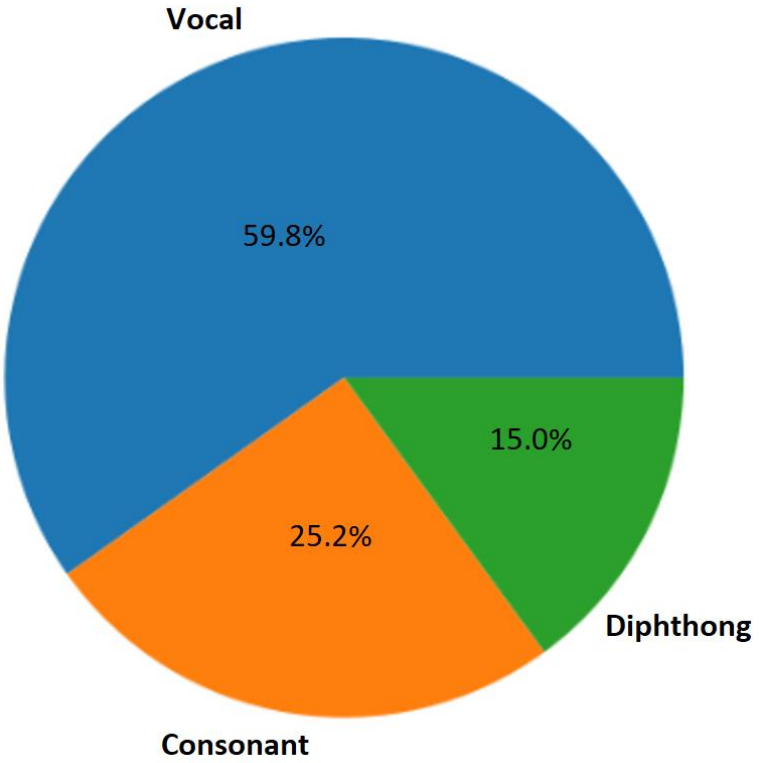


**Figure 1.**
Spelling Error Class Percentage.

Using data on word pronunciation errors, respondent attributes with categorical type will be transformed using numeric coding which can be seen in table 4. Descriptive statistics of respondent attributes for the class of pronunciation errors were presented in Table 5.

**Table 4.**
Categorical Attributes Recode.

| No. | Attributes | Recode |
|---|---|---|
| **1** | Gender | |
| | Male | 1 |
| | Female | 2 |
| **2** | Educators Type | |
| | Lecturer | 1 |
| | Teachers | 2 |
| **3** | Education Category | |
| | Bachelor | 2 |
| | Elementary | 1 |
| | Middle | 3 |
| | High | 4 |

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 9, No. 6: 1897-1904, 2025*
*DOI: 10.55214/25768484.v9i6.8275*

**Table 5.**
Descriptive Analysis.

| Attributes | Error Class | N | Mean | Min. | Max. | Std. Dev. | Std. Error |
|---|---|---|---|---|---|---|---|
| Gender | Diphthong | 16 | 1.25 | 1 | 2 | 0.447 | 0.112 |
| | Consonant | 27 | 1.44 | 1 | 2 | 0.506 | 0.097 |
| | Vocal | 64 | 1.45 | 1 | 2 | 0.502 | 0.063 |
| | Total | 107 | 1.42 | 1 | 2 | 0.496 | 0.048 |
| Age | Diphthong | 16 | 38.75 | 30 | 46 | 6.476 | 1.619 |
| | Consonant | 27 | 36.41 | 30 | 46 | 6.290 | 1.210 |
| | Vocal | 64 | 37.36 | 28 | 46 | 7.000 | 0.875 |
| | Total | 107 | 37.33 | 28 | 46 | 6.729 | 0.651 |
| Educators Type | Diphthong | 16 | 1.63 | 1 | 2 | 0.500 | 0.125 |
| | Consonant | 27 | 1.52 | 1 | 2 | 0.509 | 0.098 |
| | Vocal | 64 | 1.64 | 1 | 2 | 0.484 | 0.060 |
| | Total | 107 | 1.61 | 1 | 2 | 0.491 | 0.047 |
| Education Category | Diphthong | 16 | 1.94 | 1 | 3 | 0.854 | 0.213 |
| | Consonant | 27 | 1.74 | 1 | 3 | 0.813 | 0.156 |
| | Vocal | 64 | 1.80 | 1 | 3 | 0.694 | 0.087 |
| | Total | 107 | 1.80 | 1 | 3 | 0.745 | 0.072 |

Based on the descriptive analysis obtained, it can be seen in the gender attribute, the mean value shows that male gender makes more pronunciation errors than female. In the age attribute, the age of the teacher who made a lot of pronunciation errors was between 36-39 years. In the educators' type attribute, it can be seen that the teacher makes more pronunciation errors than the lecturer. Homogeneity test and ANOVA can be seen in Table 6 and Table 7.

**Table 6.**
Homogeneity of Variances.

| Attributes | Leven Statistic | df1 | df2 | Sig. |
|---|---|---|---|---|
| Gender | 10.503 | 2 | 104 | 0.000 |
| Age | 1.485 | 2 | 104 | 0.231 |
| Educators Type | 1.064 | 2 | 104 | 0.349 |
| Education Category | 1.583 | 2 | 104 | 0.210 |

**Table 7.**
ANOVA.

| | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| Gender | Between Groups | 0.549 | 2 | 0.274 | 1.118 | 0.331 |
| | Within Groups | 25.526 | 104 | 0.245 | | |
| | Total | 26.075 | 106 | | | |
| Age | Between Groups | 55.299 | 2 | 27.649 | 0.606 | 0.547 |
| | Within Groups | 4744.253 | 104 | 45.618 | | |
| | Total | 4799.551 | 106 | | | |
| Educators Type | Between Groups | 0.289 | 2 | 0.144 | 0.596 | 0.553 |
| | Within Groups | 25.225 | 104 | 0.243 | | |
| | Total | 25.514 | 106 | | | |
| Education Category | Between Groups | 0.396 | 2 | 0.198 | 0.353 | 0.704 |
| | Within Groups | 58.482 | 104 | 0.562 | | |
| | Total | 58.879 | 106 | | | |

## 4. Discussion

The error detection research and statistical analysis that have been carried out have shown quite interesting results. Based on the results of the detection of pronunciation errors, the largest pronunciation error is vocal error, which is 59.8% compared to diphthong and consonant errors. Respondents who live in areas that are familiar with the procedures for speaking the Batak culture are more likely to make many mistakes in pronouncing the letter 'e'. This phenomenon is thought

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 9, No. 6: 1897-1904, 2025*
*DOI: 10.55214/25768484.v9i6.8275*

to be influenced by several factors such as family background, culture, environment.

In the educators' type attribute, the lecturer made a vocal error of 54.8%. The tendency of the majority of vowel errors made by lecturers is the emphasis of sounds like 'i' becomes 'e' or 'a' becomes 'e'. Meanwhile, the consonant sound error in lecturer is 31% and ranks second from pronunciation error followed by diphthong error of 14.3%. The majority of pronunciation errors in educators type teachers are vocal errors, which are 63.1% which are mostly caused by regional and cultural elements. However, this initial assessment is still limited to field observations. While the errors of consonant and diphthongs are 24.2% and 19%, respectively.

On the gender attribute, male has a percentage of vowel pronunciation errors of 56.8%, followed by consonant and diphthong errors of 24.2% and 19%, respectively. As for female respondents, the ratio of vowel pronunciation errors is greater than the percentage of male vocal errors, which is 64.4%.

Based on the descriptive analysis value, the mean value obtained for the gender attribute shows that the diphthong error class is dominated by male respondents. While there is no significant difference between the classes of consonant and vowel errors, although from the mean value obtained, male respondents have a greater number of errors than female respondents in the two error classes. In the age attribute, the diphthong error was dominated by teachers aged 38 – 39 years. In the consonant error class, the ages that make a lot of pronunciation errors in this class range from 36 to 37 years. While in the vocal error class, the average age of those who make mistakes in this class is 37-38 years.

In the homogeneity of variance test, it can be seen that the age attribute does not have a significant effect on the class of pronunciation errors made by respondents where the Leven score obtained is quite large which indicates a very heterogeneous level of data in a class. The educator type attribute gave a significantly better score than the other attributes. However, the Leven Statistics value obtained is still much larger than the optimal value. The one-way ANOVA test that has been carried out shows that the lowest sum of squares and mean square values are found in the Educators Type attribute which shows that there is a correlation between teacher and lecturer attributes on the number of pronunciation errors.

## 5. Conclusion

Errors in the pronunciation of standard Indonesian words are still found in educators. This can be seen from the large number of errors identified from the error detection process carried out. The most common class of errors found is vowel errors where the number of errors is quite significant compared to other error classes. This high level of class error can be assumed as a result of language habits that follow the culture, especially Batak and Javanese culture, which greatly affects the language procedures of the respondents. Judging from the type of teacher, the teacher makes more pronunciation errors than the lecturer. This can be a serious concern because teachers are educators who play a very important role in shaping children's mental and language skills in a formal environment, especially at the elementary and middle school levels.

## Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

## Copyright:



*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 9, No. 6: 1897-1904, 2025*
*DOI: 10.55214/25768484.v9i6.8275*

## References


[1] E. Erwina, Study of standard Indonesian language terms. Singapore: Singapore International Press, 2012.

[2] A. Bezuidenhout, "The coherence of contextualism," Mind and Language, vol. 21, no. 1, pp. 1–10, 2006.

[3] E. Kosasih and W. Hermawan, Indonesian language based on scientific writing and journals. Bandung: CV Thursina, 2012.

[4] V. Ningrum, "Use of standard and non-standard words among students of the national development university "VETERAN" YOGYAKARTA," Jurnal Skripta, vol. 5, no. 2, pp. 1-96, 2019.

[5] D. K. Anggraeni, "Slang in Facebook status of Muhammadiyah Kramat Vocational School students and its implications," Jurnal Skripta, vol. 6, no. 1, pp. 44–53, 2020.

[6] Z. Sirait, "The use of Indonesian in public spaces that do not meet standard language," Linguistics: Journal of Language and Literature, vol. 6, no. 1, pp. 1-9, 2021.

[7] N. M. D. Purnamayani, I. N. S. Sudiara, and S. A. P. Sriasih, "Analysis of language errors in discussion in Indonesian language learning of class XI students at SMA Negeri 1 Sukasada," Jurnal Pendidikan Bahasa dan Sastra Indonesia Undiksha, vol. 2, no. 1, pp. 1-11, 2014.

[8] A. Ariyana and N. Anggraini, "Indonesian language errors in discussion of undergraduate students of English study program, FKIP UMT," Lingua Rima: Jurnal Pendidikan Bahasa dan Sastra Indonesia, vol. 10, no. 2, pp. 27-33, 2021.

[9] S. Supriadin, "Analysis of language errors in Indonesian language learning interactions of FPOK IKIP Mataram students," Journal of Social Sciences and Education, vol. 3, no. 3, pp. 88–93, 2019.

[10] A. Alber and H. Hermaliza, "The ability to analyze language errors of students of the Indonesian language and literature education study Program, Riau islamic university," Jurnal Sastra Indonesia, vol. 9, no. 1, pp. 1-10, 2020. https://doi.org/10.15294/jsi.v9i1.36366

[11] S. Supriadin, "Identification of the use of standard vocabulary in Indonesian language discourse in grade VII students at SMP Negeri 1 Wera, Bima Regency, 2013/2014 Academic Year," Jurnal Ilmiah Mandala Education, vol. 2, no. 2, pp. 67-76, 2016.

[12] D. Purwati, "The effects of lecturersâ€™ formal and informal talks on studentsâ€™ understanding of the material in the language learning process," IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature, vol. 8, no. 1, pp. 93-104, 2020. https://doi.org/10.24256/ideas.v8i1.1315

[13] O. Al-Nawrasy, "The effect of native and nonnative English language teachers on secondary students' achievement in speaking skills," Jordan Journal of Educational Sciences, vol. 9, no. 2, pp. 243-254, 2013.

[14] N. Bertoldi, M. Cettolo, and M. Federico, "Statistical machine translation of texts with misspelled words," in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010, pp. 412-419.

[15] T.-H. Chang, H.-C. Chen, Y.-H. Tseng, and J.-L. Zheng, "Automatic detection and correction for Chinese misspelled words using phonological and orthographic similarities," in Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing, 2013, pp. 97-101.

[16] R. Aşliyan, K. Günel, and T. Yakhno, "Detecting misspelled words in Turkish text using syllable n-gram frequencies," in Pattern Recognition and Machine Intelligence: Second International Conference, PReMI 2007, Kolkata, India, December 18-22, 2007. Proceedings 2, 2007: Springer, pp. 553-559.

[17] G. A. Rice and D. O. Robinson, "The role of bigram frequency in the perception of words and nonwords," Memory & Cognition, vol. 3, no. 5, pp. 513-518, 1975. https://doi.org/10.3758/BF03197523

[18] F. Elghannam, "Text representation and classification based on bi-gram alphabet," Journal of King Saud University-Computer and Information Sciences, vol. 33, no. 2, pp. 235-242, 2021. https://doi.org/10.1016/j.jksuci.2019.01.005

[19] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), 2004, pp. 605-612.

[20] M. M. Rana, M. T. Sultan, M. Mridha, M. E. A. Khan, M. M. Ahmed, and M. A. Hamid, "Detection and correction of real-word errors in Bangla language," presented at the 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), 2018.

[21] N. M. Jahan, A. Sarker, S. Tanchangya, and M. Abu Yousuf, "Bangla real-word error detection and correction using bidirectional lstm and bigram hybrid model," in Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020, 2021: Springer, pp. 3-13.

[22] B. Nurgiyantoro, B. Lestyarini, and D. H. Rahayu, "Mapping junior high school students' functional literacy competence," Cakrawala Pendidikan, vol. 39, no. 3, pp. 560-572, 2020.


*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 9, No. 6: 1897-1904, 2025*
*DOI: 10.55214/25768484.v9i6.8275*

[23] H. Alwi and A. Moeliono, Standard grammar of Indonesian language, 3rd ed. Jakarta: Badai Pustaka, 1998.
[24] D. Hládek, J. Staš, and M. Pleva, "Survey of automatic spelling correction," Electronics, vol. 9, no. 10, p. 1670, 2020. https://doi.org/10.3390/electronics9101670
[25] E. Brill and R. C. Moore, "An improved error model for noisy channel spelling correction," in Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, 2000, pp. 286-293.
[26] Language Development and Fostering Agency, Big Indonesian dictionary. Jakarta: Language Development and Fostering Agency, Ministry of Education and Culture, 2018.
[27] N. Fitriyana, A. Wiyarsi, J. Ikhsan, and K. H. Sugiyarto, "Android-based-game and blended learning in chemistry: Effect on students' self-efficacy and achievement," Jurnal Cakrawala Pendidikan, vol. 39, no. 3, pp. 507-521, 2020.

*Edelweiss Applied Science and Technology*
*ISSN: 2576-8484*
*Vol. 9, No. 6: 1897-1904, 2025*
*DOI: 10.55214/25768484.v9i6.8275*