Simultaneous Bismuth and Copper Determination from Full Differential Pulse Voltammogram Using XGBoost and Gradient Boosting Regression

Suprapto Suprapto*, Yessy I. Riwayati , Yatim L. Ni'mah

Department of Chemistry, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Surabaya, Indonesia

Abstract
In this study, the differential pulse voltammetry (DPV) method was used to simultaneously determine bismuth and copper concentrations. A 25 bismuth and copper mixtures at the designed ratio were measured using the DPV technique. However, the overlapping differential pulse voltammograms obtained made it difficult to quantitatively analyze the concentrations based on adaptive peak current selection. To address this issue, the voltammograms were preprocessed using derivatization and peak subtraction. The second derivative voltammogram was found to be highly correlated with the copper-bismuth concentration ratio, resulting in improved fit and prediction accuracy. To further improve the accuracy and precision of the training and prediction results, XGBoost and Gradient Boosting regression models were applied. The XGBoost and Gradient Boosting regression models showed high accuracy and precision with r-squared values of 0.877 and 0.993 for copper, and 0.879 and 0.993 for bismuth, respectively. The mean recoveries of copper were 99.84% and 98.07%, while bismuth recoveries were 93.17% and 90.85% for XGBoost and Gradient Boosting, respectively. Additionally, cross-validation using 10 splits produced a mean score of 45.565 and a mean absolute error of 13.051 for copper, and a mean score of 13.600 and a mean absolute error of 10.920 for bismuth. Overall, the results indicate that the proposed method is an accurate and precise way to simultaneously determine bismuth and copper concentrations.

1.INTRODUCTION
Copper and bismuth are two important elements that have a wide range of applications in industry and technology. Copper is used in electrical wiring, electronics, and plumbing, while bismuth is used in pharmaceuticals, cosmetics, and alloys. However, both of these elements can be toxic at high concentrations, making it important to accurately measure their concentrations in various matrices [1-6].
Overlapping differential pulse voltammetry technique that involves simultaneous measurement of two or more analytes in a single solution can provide significant advantages in terms of time, cost, and sample volume. By using overlapping differential pulse voltammetry, it is possible to determine the concentrations of copper and bismuth in a single measurement, which can be particularly useful in applications where speed and efficiency are critical [7-8].
The research on simultaneously determination of copper and bismuth using overlapping differential pulse voltammogram involves the development of modified analytical methods, the optimization of existing methods, and the application of these methods to various matrices. The goal of this research is to improve the accuracy, precision, and

sensitivity of the analysis of copper and bismuth, while also reducing the cost and time required for the analysis. This research can have important applications in environmental monitoring, industrial process control, and biomedical analysis [9].To address the challenges associated with the quantitative analysis of overlapping voltammograms of bismuth and copper ions, this study proposed a novel approach that involved a full voltammogram, derivatization, and voltammogram subtraction. The processed voltammogram was then optimized using XGBoost and Gradient Boosting regression models to identify significant features in the obtained data [10].
The Gradient Boosting regression model was chosen because of its ability to handle sparse and quantile-weighted data and its effectiveness in identifying nonlinear relationships between model targets and features. It can also effectively deal with missing values, outliers, and categorical values with high collinearity without special handling. The fitting model utilized a differentiated loss function and a gradient descent optimization algorithm. XGBoost has built-in regularization techniques to prevent overfitting, such as L1 and L2 regularization. This can be

* Corresponding author:
Email: suprapto@chem.its.ac.id

particularly useful when dealing with high-dimensional datasets. XGBoost has a built-in mechanism for handling missing data, which can be a common problem in real-world datasets. XGBoost uses a method called "tree pruning" to remove unnecessary splits in decision trees, which can reduce overfitting and improve model performance. XGBoost uses a technique called "weighted quantile sketch" to approximate the gradients of the loss function, which can improve the speed and accuracy of the algorithm [11].

In this study, Differential Pulse Voltammetry (DPV) was used to obtain the full voltammogram of bismuth and copper ions, followed by XGBoost and Gradient Boosting regression to extract voltammogram features that correlated with the bismuth and copper concentration ratios. The training datasets were prepared using bismuth and copper ion solutions at specific concentration ratios, and the regression models were developed. These models were then utilized to predict the concentration ratios of bismuth and copper ions in the test solution (test dataset).

2.EXPERIMENTAL
2.1.Materials and Instrument
In this study, a glassy carbon was applied as working electrode, an Ag/AgCl as reference electrode, a platinum as counter electrode, and an eDAQ potentiostat as electric potential source as well as current measurement apparatus. The chemicals used in the study were bismuth nitrate pentahydrate ($Bi(NO_3)_3.5H_2O$) and copper nitrate trihydrate ($Cu(NO_3)_2.3H_2O$), which were obtained from Merck, a 65% nitric acid solution (HNO_3) and deionized water, which were obtained from a local store in Surabaya. All chemicals were used as received without any prior purification.

2.2.Method
A bismuth stock solution of 2000 mg/L was prepared by dissolving 4.6422 grams of $Bi(NO_3)_3.5H_2O$ in 1 M HNO_3, while 3.8017 grams of $Cu(NO_3)_2.3H_2O$ were dissolved in deionized water. From these stocks, training solutions were prepared by diluting bismuth and copper to concentrations ranging from 600-1000 mg/L and 100-500 mg/L, respectively, with a concentration range of 100 mg/L. The solutions were analyzed using differential pulse voltammetry (DPV) with an eDAQ potentiostat, glassy carbon working electrode, Ag/AgCl reference electrode, and platinum counter electrode. The measurements were repeated three times.

Two test solutions of bismuth and copper were prepared at concentrations of 650:150 and 850:250 mg/L, respectively. These solutions fell within the training data set range, but were not part of the 25 variations of training datasets, and their concentrations were treated as unknown data.

Copper and bismuth voltammograms were subjected to derivatization using the Scipy Signal Savitzky-Golay filter with a 3-window length [12]. The regression analysis was conducted using the sklearn linear models: Gradient Boosting and XGBoost regression [9,13].

3.RESULTS AND DISCUSSION
3.1.Bismuth and copper voltammograms
DPV measurements were conducted on copper and bismuth solution using an eDAQ potentiostat with glassy carbon as the working electrode. The copper DPV measurements at concentrations of 100, 200, 300, 400, and 500 mg/L were presented in Fig. 1(a), while the bismuth DPV measurements at concentrations of 600, 700, 800, 900, and 1000 mg/L were shown in Fig. 1(b). The superimposed voltammograms of single Cu, Bi, and their mixture were illustrated in Fig. 1(c). Fig. 1(c) reveals that the oxidation-reduction currents of the two metal ions coincide, resulting in the voltammogram overlapping.

The copper peak current potential was observed at 0.06 V, while the bismuth peak current potential was at 0.0065 V. Regression analysis was performed on the peak current values as a function of individual copper and bismuth concentrations, resulting in linear regression slopes of 3.8424e-05 and 9.6947e-05, intercepts of 0.00065 and -0.0276, and R-values of 0.999567 and 0.9995, respectively, for bismuth and copper. The regression curves can be seen in Fig 2.

3.2.Bismuth and copper training and test solution measurement
Figure 3(a) displays the voltammograms of bismuth and copper solutions at different concentrations, which constituted the training dataset. A total of 25 concentration variations with three measurement replications were used, resulting in 75 DPV measurement voltammogram curves.

The regression model based on the current peak values at each copper and bismuth peak potential in Fig 1 did not perform well, as indicated in Fig. 4. Fig 3(b) shows that the peaks of each copper and bismuth ion in the combined solution shift to the left or right depending on changes in their concentration. This conforms to the Nernst equation, which states that the electrochemical cell potential is determined by the concentration ratio of oxidized and reduced chemicals. Therefore, it is necessary to optimize alternative methods to compensate for peak shifts, as the current peaks vary as a function of changes in the concentration of copper and bismuth.

The quantitative analysis of copper using full voltammogram was carried out using regression method based on concentration variation. The voltammograms were divided into 90% training sets and 10% test sets. We obtained a linear model with an r-value of 0.9723 and a root mean square error (rmse) of 31.9029 for the 90% training set. Predicting the 10% test set produced a regression with an r-value of 0.6111 and a rmse of 188.8228. For bismuth, the quantitative analysis resulted in an r-value of 0.9985 and a rmse of 7.5504 for the 90% training set and an r-value of 0.98198 with a rmse of 33.6755 for the 10% test set. Fig 5 shows the regression curves for 90% of the training and 10% of the test datasets.
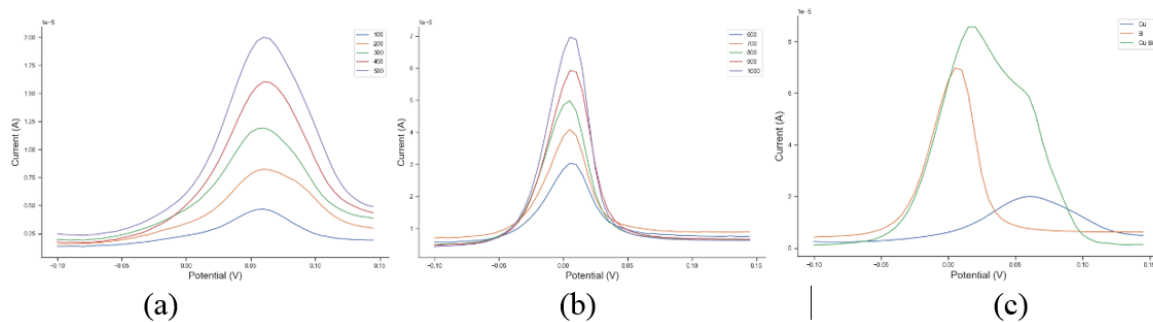
Fig. 1. Voltammogram of the single solution Cu (a), Bi (b), and Bi-Cu overlay (c)
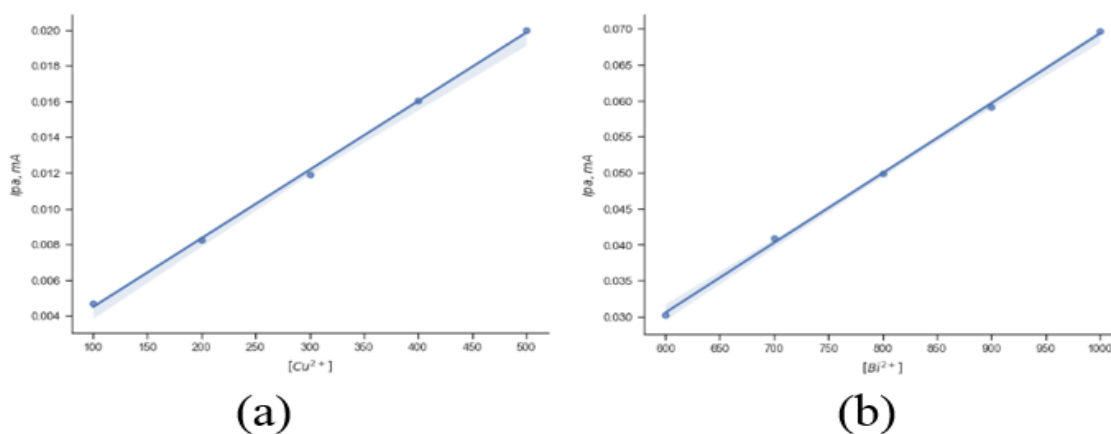
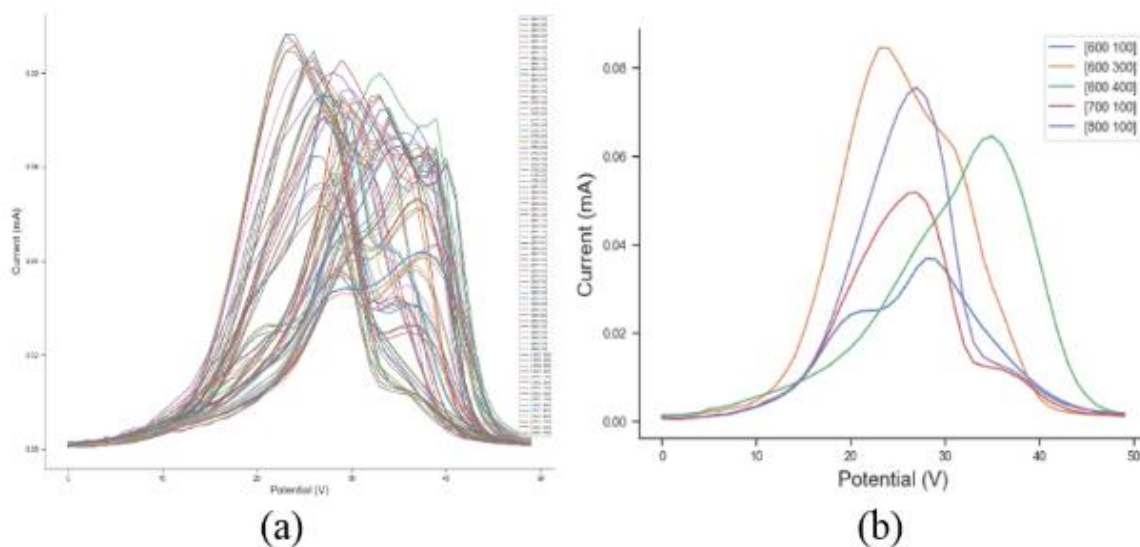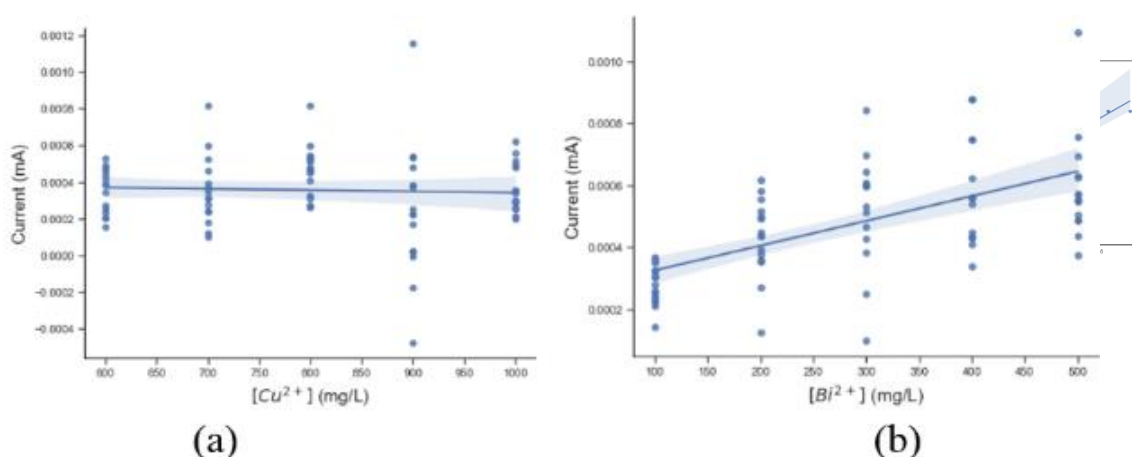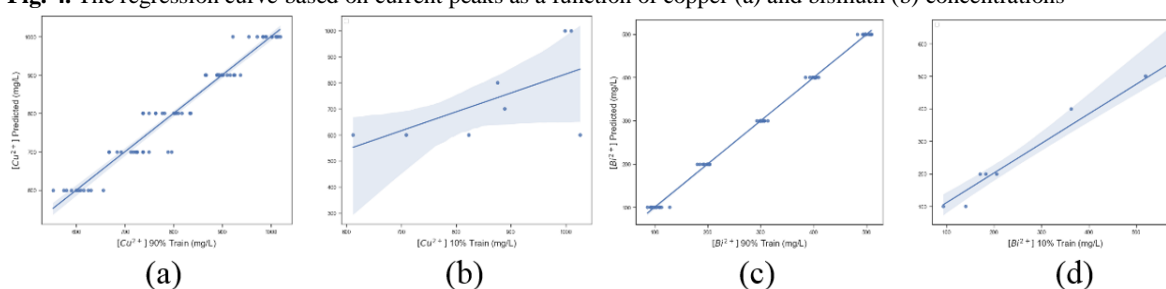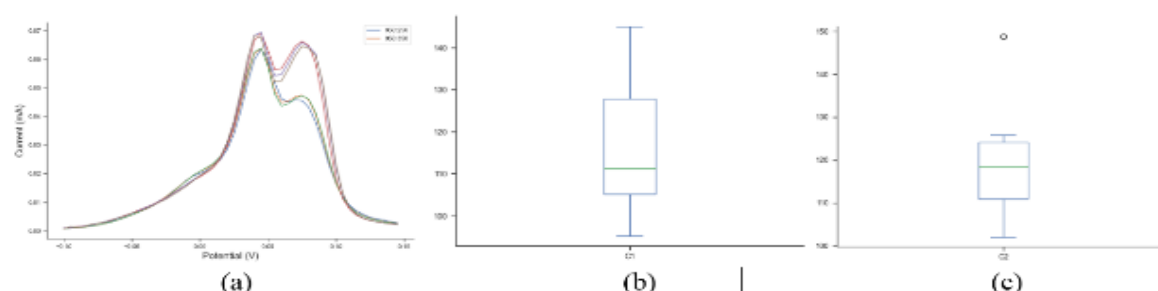**Fig. 2.** The regression curve of the individual (a) bismuth and (b) copper ion

Fig. 3. The 75 Bi and Cu solution voltammogram as the training data set

Fig. 4. The regression curve based on current peaks as a function of copper (a) and bismuth (b) concentrations

**Fig. 5.** The regression curves of 90% train and 10% test datasets for copper (a-b) and bismuth (c-d)

**Fig. 5.** Voltammogram of Bi and Cu test solutions (a) Cu recovery (b), and Bi recovery (c)

The out-of-sample solution with concentration ratios of 850:250 and 950:350 mg/L was used to validate the accuracy of the linear model trained on the datasets. Fig 6 shows the voltammogram of the out-of-sample solution with three replications. The linear regression analysis of the full voltammogram resulted in a recovery of 116.57% for copper and 120.35% for bismuth, indicating poor accuracy. Additionally, the rmse for both predictions was high, leading to the conclusion that the full voltammogram analysis approach was unsuccessful. To improve the quantitative analysis of copper and bismuth using DPV methods, the voltammogram derivatization technique was employed. Two approaches were proposed: first-order and second-order derivatization of the voltammogram. Another approach involved peak subtraction, whereby the copper and bismuth mixture voltammogram was subtracted by each copper or bismuth voltammogram. Fig. 6 and 7 show the recovery rates of copper and bismuth for the full voltammogram, first-order derivatization, second-order derivatization, and voltammogram subtraction methods, as reported in [8].

The results presented in Fig 6 and 7 demonstrate that the second-order voltammogram derivatization method has a higher recovery rate. However, the overall rmse prediction values for copper using the full voltammogram, first-order derivatization, second-order derivatization, and voltammogram subtraction methods were all quite high, with values of 785.2571, 794.6386, 811.8610, and 785.4763, respectively. The rmse prediction values for bismuth were 185.6625, 180.0312, 193.8273, and 185.6625, respectively. Fig 8 displays the first-order and second-order derivatization of the copper and bismuth voltammograms.
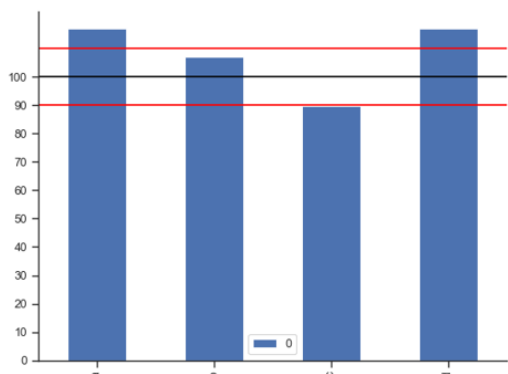
Fig. 6. The copper recovery of full voltammogram (a), first-order derivatization (b), second-order derivatization (c), and voltammogram subtraction (d)
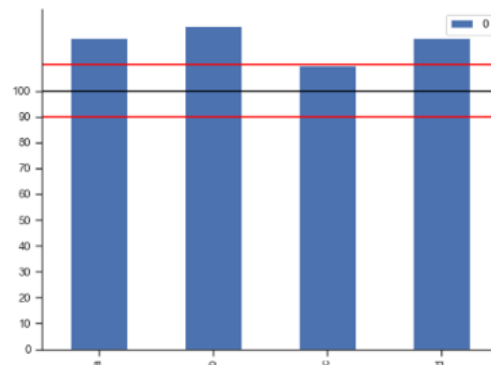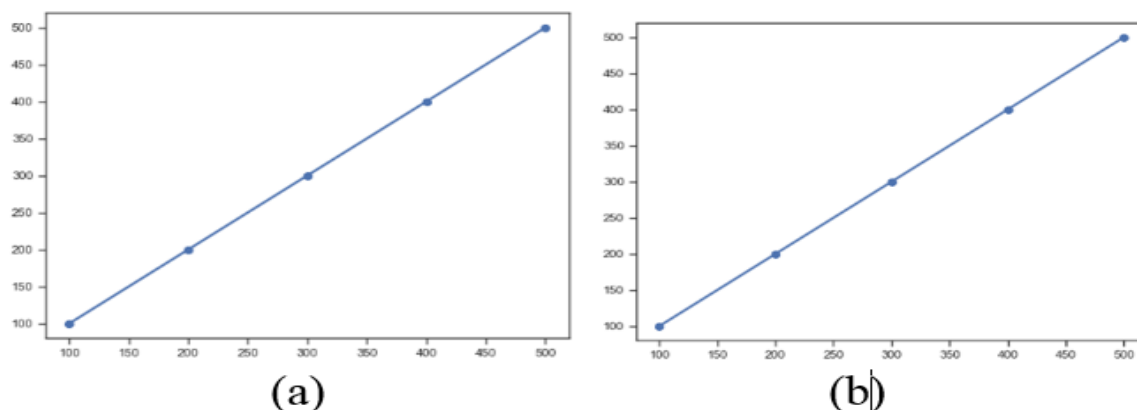

Fig. 7. The bismuth recovery of full voltammogram (a), first-order derivatization (b), second-order derivatization (c), and voltammogram subtraction (d)
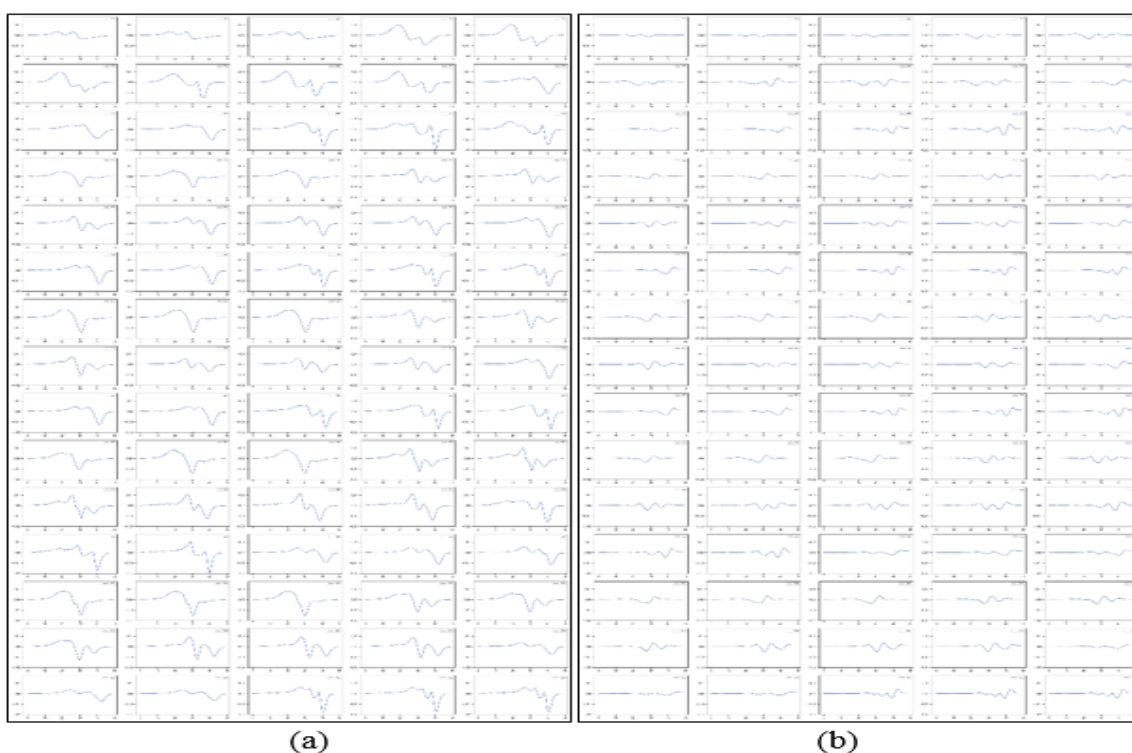
Fig. 8. First derivative (a) and second derivative (b) of copper and bismuth voltammograms
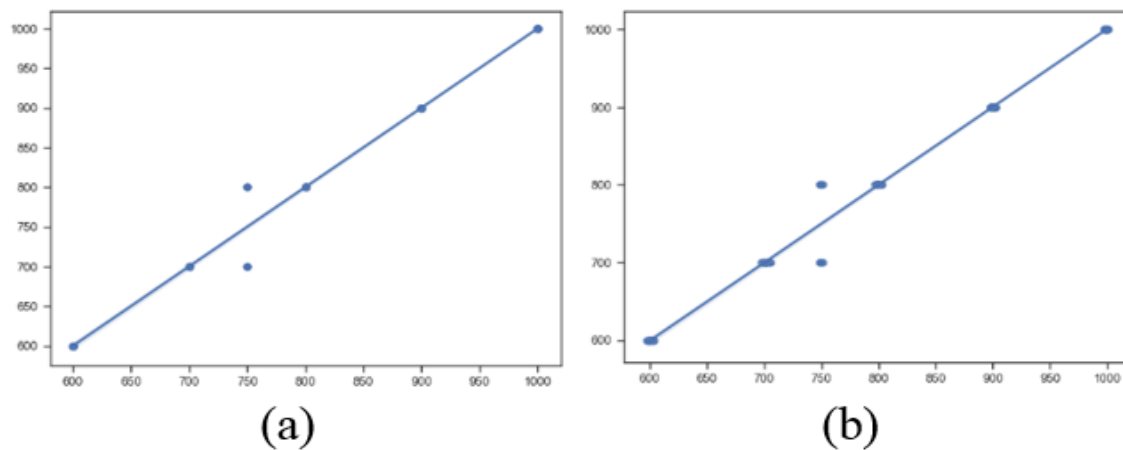
Fig. 9. The copper regression curves using XGBoost (a) and Gradient Boosting (b) for training datasets

**Fig. 10.** The bismuth regression curves using XGBoost (a) and Gradient Boosting (b) for training datasets

After conducting previous optimizations, the focus for quantitative analysis of copper and bismuth was on the second-order derivatization voltammogram. The optimization process involved refining regression models that demonstrated a good fit for training datasets and accurate predictions for out-of-sample test solutions. Based on the optimized regression models, XGBoost and Gradient Boosting regression were found to produce both good-of-fit and accuracy. The regression curves for copper and bismuth using XGBoost and Gradient Boosting training sets are shown in Fig. 9 and 10, respectively.

The XGBoost regression models for copper and bismuth had r-squared values of 0.877 and 0.993, respectively, with adjusted r-squared values of 0.697 and 0.983. The probability (F-statistic) values were 9.53e-06 and 1.24e-23, and the Log-Likelihood values were -398.81 and -290.16, respectively. The regression coefficients are depicted in Fig 11. Cross-validation with 10 splits resulted in a mean of 56.331 and mean absolute error (MAE) of 20.035 for copper, and a mean of 16.890 and MAE of 9.755 for bismuth.
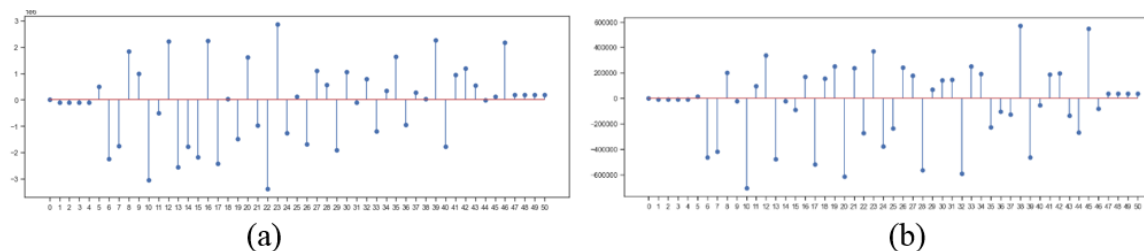
The Gradient Boosting regression models for copper and bismuth had r-squared values of 0.879 and 0.993, respectively, with adjusted r-squared values of 0.702 and 0.983. The probability (F-statistic) values were 7.65e-06 and 1.24e-23, and the Log-Likelihood values were -397.77 and -290.16, respectively. The regression coefficients for Gradient Boosting regression of copper and bismuth are depicted in Fig. 12. Cross-validation with 10 splits resulted in a mean of 45.565 and mean absolute error (MAE) of 13.051 for copper, and a mean of 13.600 and MAE of 10.920 for bismuth

Fig 13 illustrates that the mean recoveries for copper and bismuth using XGBoost were 99.84% and 93.17%, respectively. On the other hand, the mean recoveries using Gradient Boosting were 98.07% and 90.85%, respectively, as shown in Fig 14. Cross-validation using 10 splits resulted in a mean of 45.565 with a mean absolute error of 13.051 for copper and a mean of 13.600 with a mean absolute error of 10.920 for bismuth.
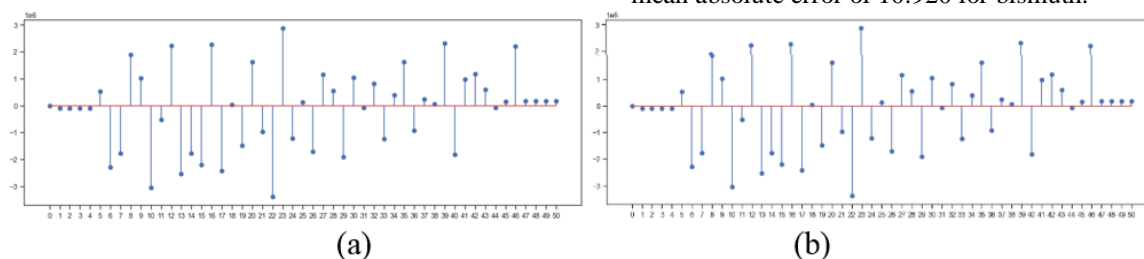
Fig. 11. XGBoost regression coefficient for copper (a) and bismuth (b)

Fig. 12. Gradient Boosting regression coefficient for copper (a) and bismuth (b)
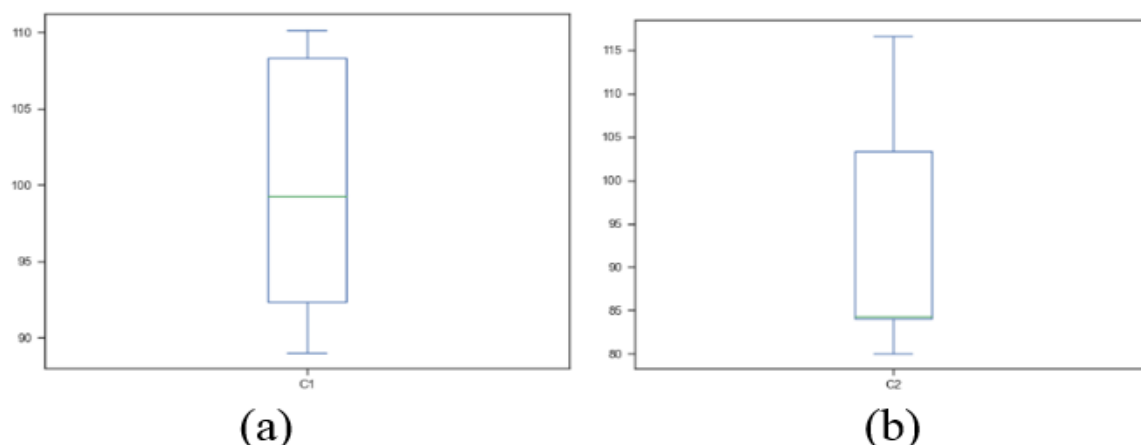
Fig. 13. XGBoost prediction of copper (a) and bismuth (b) out of sample test solution
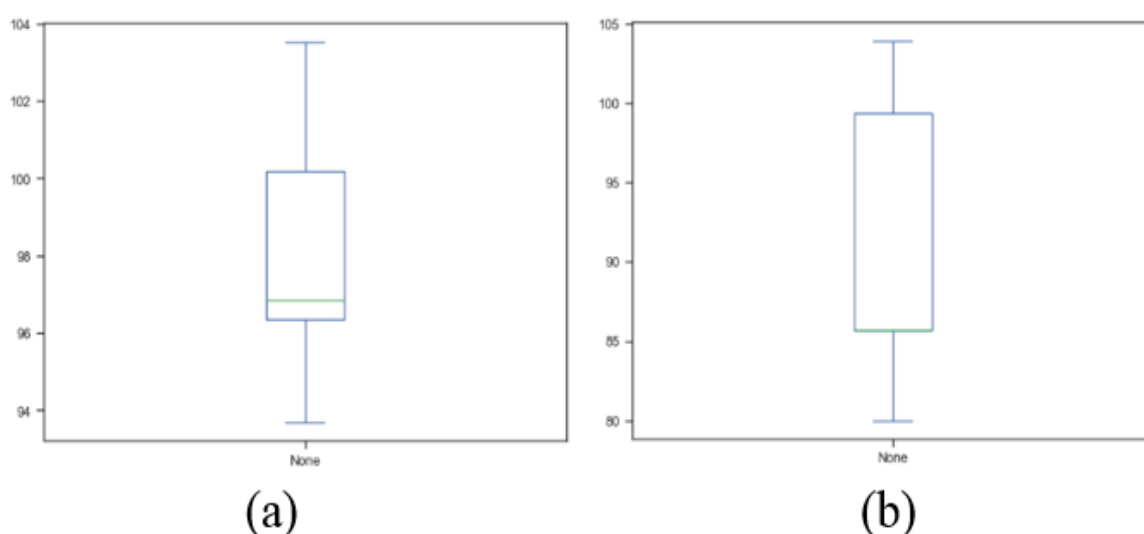

Fig. 14. Gradient Boosting prediction of copper (a) and bismuth (b) out of sample test solution

4. CONCLUSION

The successful quantitative analysis of copper and bismuth using overlapping differential pulse voltammogram required voltammogram derivatization to achieve feasible r-squared scores for training data and prediction accuracy. The second derivative voltammogram was found to produce high r-squared values of 0.877 and 0.993 for copper and bismuth, respectively, with XGBoost regression, and 0.879 and 0.993 for copper and bismuth, respectively, with Gradient Boosting regression. The mean recoveries of copper using XGBoost and Gradient Boosting were 99.84% and 98.07%, respectively, while for bismuth, they were 93.17% and 90.85%, respectively.

Acknowledgments

REFERENCES

[1] H. Liu, R. Cheng, X. Dong, S. Zhu, R. Zhou, L. Yan, C. Zhang, Q. Wang, Z. Gu and Y. Zhao, BiO 2– x Nanosheets as Radiosensitizers with Catalase-Like Activity for Hypoxia Alleviation and Enhancement of the Radiotherapy of Tumors. Inorg, *Chem*. 59 (2020) 3482–3493.

[2] A. Babaei, E. Shams, and A. Samadzadeh, Simultaneous Determination of Copper, Bismuth and Lead by Adsorptive Striing Voltammetry in the Presence of Thymolphthalexone, *Anal. Sci*. 22 (2006) 955–959.

[3] G. M. S. Alves, J. M. C. S. Magalhães, and H. M .V. M. Soares, Simultaneous Determination of Copper(II), Lead(II), and Zinc(II) at Bismuth Film Electrode by Multivariate Calibration, *Electroanalysis* 23 (2011) 1410–1417.

[4] J. Baldenebro-López, J. Castorena-González, N. Flores-Holguín, J. Almaral-Sánchez, and D. Glossman-Mitnik, Computational Molecular Nanoscience Study of the Properties of Copper Complexes for Dye-Sensitized Solar Cells, *Int. J. Mol. Sci*. 13 (2012) 16005–16019.

[5] D. G. Barceloux and D. Barceloux, Copper, J. Toxicol. *Clin. Toxicol*. 37 (1999) 217–230.
[6] J. Vilaplana and C. Romaguera, New developments in jewelry and dental materials, *Contact Dermatitis*. 39 (1998) 55-57.

[7] J. Threeprom, S. Purachaka, and L. Potipan, Simultaneous determination of Cr(III)-EDTA and Cr(VI) by ion interaction chromatography using a C18 column, *J. Chromatogr. A*. 1073 (2005) 291-295.

[8] M. B. Gholivand and A. A. Romiani, Alication of adsorptive striing voltammetry to the simultaneous determination of bismuth and copper in the presence of nuclear fast red, *Anal. Chim. Acta*. 571 (2006) 99-104.

[9] R. Abedi, A. Parchehbaf Jadid, and Z. Azari, Modified Sodium Dodecyl Sulfate Coated γ-Alumina Nanoparticles with (e)n'-(2-Hydroxybenzylidene)Isonicotinohydrazide as New Sorbent for Copper Determination in Real Samples, *Iran. J. Anal. Chem.* 8 (2022).

[10] T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2016) 785-794.

[11] E. Hasdemir and K. Karaboduk, Simultaneous Determination of Bismuth and Copper by Square Wave Voltammetry in The Presence of Ethylenediaminetetraacedic Acid, *Gazy Univ. J. Sci*. 23 (2010) 33-40.

[12] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, and J. VanderPlas, The SciPy 1.0 Contributors SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, *Nat. Methods*. 17 (2020) 261-27.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res*. 12 (2011) 2825-2830.

.

تعیین همزمان بیسموت و مس از ولتاموگرام پالس تفاضلی کامل با استفاده از رگرسیون تقویت ایکس جی و تقویت گرادیان

ساپراپتو ساپراپتو *، سی آی ریوایاتی ، یاتیم ال نای ماه

گروه شیمی، دانشکده علوم و تجزیه و تحلیل داده ها، سپولا، نوامبر، سورابایا، اندونزی

* E-mail: suprapto@chem.its.ac.id

چکیده

در این مطالعه از روش ولتامتری پالس تفاضلی برای تعیین همزمان غلظت بیسموت و مس استفاده شد. ۲۵ مخلوط از بیسموت و مس در نسبت های طراحی شده با استفاده از تکنیک ولتامتری پالس تفاضلی اندازه گیری شد. با این حال، ولتاموگرام های پالس تفاضلی این دو ماده با هم همپوشانی دارند که تجزیه و تحلیل کمی غلظت ها را بر اساس جریان پیک آنها دشوار می کند. برای حل این موضوع، ولتاموگرام ها با استفاده از مشتق سازی و تفریق پیک پیش پردازش شدند. ولتاموگرام مشتق دوم با نسبت غلظت مس به بیسموت همبستگی زیادی داشت که منجر به بهبود صحت پیش‌بینی ها شد. برای بهبود بیشتر دقت و صحت نتایج مجموعه های آموزش و پیش‌بینی، از مدل‌های رگرسیون تقویت ایکس جی و تقویت گرادیان استفاده شد. مدل‌های رگرسیون تقویت ایکس جی و تقویت گرادیان دقت و صحت بالایی را با مقادیرضریب همبستگی به ترتیب ۰/۸۷۷ و ۰/۹۹۳ برای مس و ۰/۸۷۹ و ۰/۹۹۳ برای بیسموت نشان دادند. برای تقویت ایکس جی و تقویت گرادیان میانگین بازیابی مس به ترتیب ۹۹/۸۴٪ و ۹۸/۰۷٪ بود، در حالی که بازیابی بیسموت به ترتیب ۹۳/۱۷٪ و ۹۰/۸۵٪ بود. علاوه بر این، اعتبارسنجی متقاطع با استفاده از ۱۰ نمونه، میانگین امتیاز ۴۵/۵۶۵ و میانگین خطای مطلق ۱۳/۰۵۱ برای مس، و میانگین نمره ۱۳/۶۰۰ و میانگین خطای مطلق ۱۰/۹۲۰ را برای بیسموت نشان داد. به طور کلی، نتایج نشان می دهد که روش پیشنهادی روشی دقیق و صحیحی برای تعیین همزمان غلظت بیسموت و مس است.

کلید واژه ها

بیسموت مس؛ ولتامتری پالس تفاضلی؛ تقویت ایکس جی؛ تقویت گرادیان؛ رگرسیون