

Trademark Logo Infringement Detection: A Threshold Determination Approach Using Bayesian Optimization and Siamese Neural Networks

Muhammad Rifzal Alief Ramadhan^{1*}, Stevano Wahyu Al'fandi², Egi Dio Bagus Sudewo³,
Suyahman⁴

¹Department of Business Law, Universitas Sugeng Hartono, Sukoharjo, Indonesia

^{2,4}Department of Computer Science, Universitas Sugeng Hartono, Sukoharjo, Indonesia

³Department of Information System, Universitas Royal, Asahan, Indonesia

Abstract. Trademark infringement detection is essential to prevent consumer confusion, particularly in the digital era where visually similar logos are widespread. This study proposes a similarity threshold determination framework using Siamese Neural Networks (SNN) combined with Bayesian Optimization to improve the accuracy of trademark similarity assessment. Logo images were collected from the Indonesian intellectual property database, preprocessed into a uniform format, and trained using a triplet loss approach. Bayesian Optimization was applied to determine the optimal similarity threshold, minimizing false positive and false negative classifications. The proposed model achieved an accuracy of 92.23%, with precision of 93.34%, recall of 91.44%, and an F1-score of 92.39%. The optimal threshold (0.200149) effectively balanced sensitivity and specificity, resulting in low misclassification rates. These findings demonstrate that integrating SNN with Bayesian Optimization provides a robust and legally relevant framework for trademark infringement detection, offering practical implications for strengthening intellectual property protection.

Keywords: Trademark, Bayesian Optimization, Siamese Neural Network, Triplet Loss

Received January 2025 / **Revised** February 2026 / **Accepted** February 2026

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).



INTRODUCTION

The importance of establishing a threshold for measuring similarity in trademark infringement detection is underscored by the complexities involved in trademark law and consumer perception. Trademark infringement claims hinge on the likelihood of confusion among consumers, which is a central tenet in legal disputes surrounding trademarks. The U.S. Supreme Court's decision in *McLean v. Fleming* (1878) highlights that these claims are fundamentally consumer-oriented, emphasizing the need for a clear understanding of what constitutes a confusingly similar trademark [1]. This legal framework necessitates the development of robust methodologies for assessing similarity, particularly in an era where digital commerce has proliferated, leading to an increase in potential infringements.

Siamese Neural Networks (SNNs) have emerged as a promising tool for evaluating trademark similarity due to their ability to effectively learn from limited datasets and identify nuanced differences between logos. The architecture of SNNs allows for the comparison of image pairs, making them particularly adept at detecting subtle variations that may not be immediately apparent to human observers [2]. However, the effectiveness of SNNs in trademark infringement detection is heavily reliant on the calibration of the similarity threshold. This threshold serves as a critical boundary that determines whether two trademarks are considered similar enough to warrant further legal scrutiny. A well-defined threshold can significantly reduce both false positives, where legitimate trademarks are incorrectly flagged as infringing, and false negatives, where actual infringements are overlooked [3].

The challenge of establishing an appropriate threshold is compounded by the varying standards across different industries regarding what constitutes a confusingly similar logo. Research indicates that the perception of similarity can be influenced by numerous factors, including the distinctiveness of the

¹*Corresponding author.

Email addresses: rifzalalief@gmail.com (Ramadhan)

trademarks involved and the nature of the goods or services they represent [4]. For instance, Horen and Pieters [5] discuss how varying degrees of similarity can impact consumer perceptions and brand reputation, suggesting that moderate-similarity copycats may sometimes benefit at the expense of the original brand. This highlights the necessity of a nuanced approach to threshold determination, one that considers the specific context of the brands involved.

Bayesian optimization has been proposed as a method for refining the threshold values used in SNN outputs. This probabilistic model-based optimization technique is particularly effective for optimizing complex functions, such as those encountered in machine learning applications. By iteratively adjusting the threshold based on observed performance metrics, Bayesian optimization can help identify the most effective threshold for distinguishing between similar and dissimilar logos [6]. This approach not only enhances the accuracy of trademark infringement detection systems but also contributes to a more systematic understanding of the factors influencing trademark similarity assessments.

Moreover, the integration of semantic similarity assessments into trademark evaluation processes represents a significant advancement in the field. Research by Anuar et al. emphasizes the importance of considering semantic dimensions in trademark similarity assessments, suggesting that a multi-faceted approach can provide a more comprehensive understanding of potential infringement issues [7]. This perspective aligns with the legal requirement for a thorough analysis of the likelihood of consumer confusion, which necessitates a consideration of various interdependent factors, including the similarity of goods and the distinctive elements of the trademarks in question.

The establishment of a threshold for measuring similarity in trademark infringement detection is a complex but essential task. The interplay between legal standards, consumer perceptions, and technological advancements necessitates a comprehensive approach that incorporates both traditional legal frameworks and modern machine learning techniques. By leveraging SNNs and Bayesian optimization, researchers can develop more effective systems for detecting trademark infringements, ultimately contributing to the protection of brand identities in an increasingly competitive marketplace.

METHODS

This research aims to detect trademark infringement using a threshold determination approach based on Bayesian Optimization and Siamese Neural Network. This research method consists of several main stages, as shown in Figure 1.

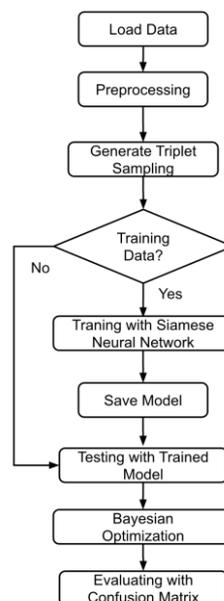


Figure 1. Research Method

These stages include data collection, preprocessing, triplet sampling, training using Siamese Neural Network with triplet loss, optimal threshold determination using Bayesian Optimization, and model

evaluation using confusion matrix. Bayesian Optimization is a principled and efficient technique for the global optimization of functions that are expensive to evaluate.

Dataset

The dataset used in this study comes from the intellectual property database of the Directorate General of Intellectual Property of Indonesia. Five registered brands are used as the main dataset, namely Gojek, Grab, Uniqlo, Miniso, and CircleCI. Each brand was reproduced by taking images from Google Images of 50 variations per brand, resulting in a dataset with a wide variety of logos. Images of the trademarks used in the study can be seen in Table 1.

No.	Trademark	Logo	Number of Data Training	Number of Data Testing
1.	Uniqlo		40 Logo	10 Logo
2.	Miniso		40 Logo	10 Logo
3.	Grab		40 Logo	10 Logo
4.	Gojek		40 Logo	10 Logo
5.	CircleCi		40 Logo	10 Logo

The preprocessing process involved converting the images to RGB format and size adjustment to 224x224 pixels. The dataset was then divided into training and test data with a ratio of 80:20 [8]-[9]. Triplet sampling was performed to generate 8000 data pairs of five classes, which were used as inputs in model training [10]. Siamese Neural Network

The training process is conducted using the Siamese Neural Network, which is designed to compare the similarity between two inputs [11]-[13]. Triplet loss is used as the loss function to minimize the distance between anchor and positive, and maximize the distance between anchor and negative, based on the semantic relationship between the data [14]-[15]. With this approach, the model can effectively distinguish genuine logos from visually resembling logos [16].

Bayesian Optimization

Once the model is trained, the Bayesian Optimization process is used to determine the optimal threshold [17]-[18]. This threshold is important to set a tolerance limit in determining whether two logos are considered similar or different.

Bayesian Optimization is a principled and efficient technique for the global optimization of functions that are expensive to evaluate [19]. The core idea behind Bayesian Optimization is to place a prior distribution

over the target function and update that prior with a set of “true” observations of the function by expensively evaluating it. This process produces a posterior predictive distribution, which is then used to inform where to make the next observation of the target function. The optimization process is guided by an acquisition function, which balances the exploitation of regions known to have good performance with the exploration of regions where there is limited information about the function's behavior. This approach allows for efficient optimization with minimal evaluations of the costly objective function [20].

One commonly used acquisition function in Bayesian Optimization is the Expected Improvement (EI), which is defined as:

$$\alpha(\theta) = E[\max(0, f(\theta) - f(\theta^+))] \quad (1)$$

where θ represents the proposed next point to evaluate, $f(\theta)$ is the predicted value of the objective function at θ , and $f(\theta^+)$ is the best observed value of the objective function. The acquisition function $\alpha(\theta)$ is used to propose the next point, guiding the optimization toward areas that are expected to yield the greatest improvement [21].

Algorithm 1: Pseudocode Bayesian Optimization

Input:

- $f(\text{threshold})$: Objective function to optimize (accuracy based on threshold)
- Bounds: [0.0, 1.0] (threshold range)
- X_{init} : Initial set of threshold values
- Acquisition function: Criteria to propose new thresholds based on minimizing the surrogate model (Gaussian Process) of the objective function.

Method:

1. Initialize surrogate model with X_{init} and $f(X_{\text{init}})$.
2. For $t = 1$ to max_iterations do:
 3. Propose next threshold (threshold_next) using acquisition function:
 $\text{threshold_next} \leftarrow \text{argmax}(\text{acquisition_function}(\text{threshold} \mid \text{surrogate_model}))$
 4. Evaluate objective function at threshold_next :
 $\text{accuracy_next} \leftarrow f(\text{threshold_next})$
 5. Update dataset with (threshold_next , accuracy_next).
 6. Update surrogate model with the updated dataset.
 7. If stopping criteria met (e.g., maximum iterations or convergence), break.
3. End for
4. Return threshold_best where $f(\text{threshold_best}) = \max(f(X))$ from the observed data.

The process of Bayesian Optimization proceeds iteratively. Initially, a set of observations is used to train a surrogate model, typically a Gaussian Process, which approximates the target function. In each iteration, the acquisition function is employed to propose the next point to evaluate based on the surrogate model's predictions [22]. The objective function is evaluated at this proposed point, and the new data is added to the dataset. The surrogate model is updated accordingly, and the process continues until a stopping criterion is met, such as reaching a maximum number of iterations or converging to an optimal solution. Ultimately, the optimal threshold or parameter is identified as the one that maximizes the objective function based on the observed data [23].

Confusion Matrix

The last stage involves evaluating the model using the confusion matrix, which includes metrics such as accuracy, precision, recall, and F1-score [24]-[25]. This evaluation provides an overall picture of the model's performance in detecting trademark infringement based on the dataset used. The confusion matrix consists of four main elements, namely True Similar (TS), True Different (TD), False Similar (FS), and False Different (FD). True Similar (TS) refers to the amount of data that is truly similar and correctly predicted as similar by the model. True Different (TD) includes the number of correct predictions for different data, where the predicted results correspond to different actual conditions. In contrast, False Similar (FS) occurs when the model incorrectly predicts data that is actually different as similar.

Meanwhile, False Different (FD) occurs when data that is actually similar is predicted as different. Based on the data from the confusion matrix, various evaluation metrics can be calculated, such as accuracy, precision, recall, and F1-score. Calculation of accuracy, precision, recall, and F1-score using the following equations.

$$Accuracy = \frac{TS}{Total\ Data} \quad (2)$$

$$Precision = \frac{TS}{TS+FS} \quad (3)$$

$$Recall = \frac{TS}{TS+FD} \quad (4)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

Accuracy measures the proportion of correct predictions made by the model out of the total data, reflecting its overall performance in predicting both similar and different classes, although it may be less reliable in cases of class imbalance [26]. Precision, on the other hand, focuses on the accuracy of the model in predicting the similar class, emphasizing the proportion of correct similar predictions, which is critical in minimizing false similar outcomes, particularly for nearly identical trademarks [27]. Recall complements this by evaluating the model's ability to identify all data in the similar class, highlighting its sensitivity in detecting true similarities, which is crucial in contexts like trademark plagiarism detection where missing similar data (false different) can have serious consequences [28]. To balance these metrics, the F1-score harmonizes precision and recall into a single comprehensive measure, offering a holistic view of the model's effectiveness, especially when both false similar and false different predictions carry significant implications [29].

RESULT AND DISCUSSION

The training process was conducted over a span of 10 epochs, with the results depicted in the accompanying graphs. The training loss, shown in Figure 2, exhibited a significant decline from the initial epoch, starting at approximately 0.20 and steadily decreasing to stabilize around 0.01 by the end of the training period.

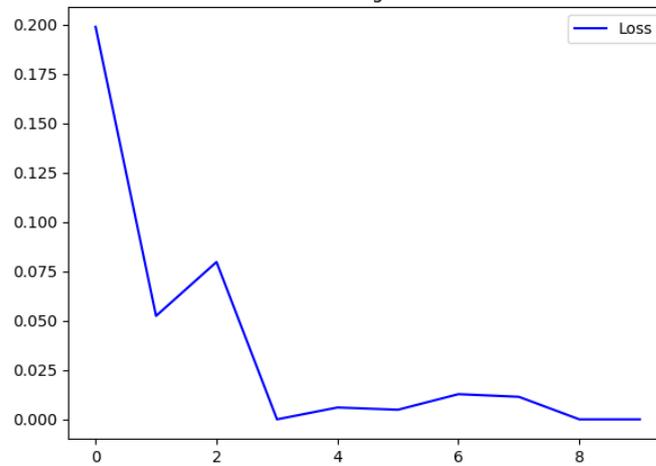


Figure 2. Training Loss

This trend indicates that the model effectively minimized prediction errors throughout the training process, as shown in Figure 3. On the right, the graph for testing accuracy shows fluctuations during the epochs. While the accuracy peaked at 1.00 in some epochs, the overall average was around 92.22%, suggesting that the model correctly classified test data at times, though performance varied across the epochs.

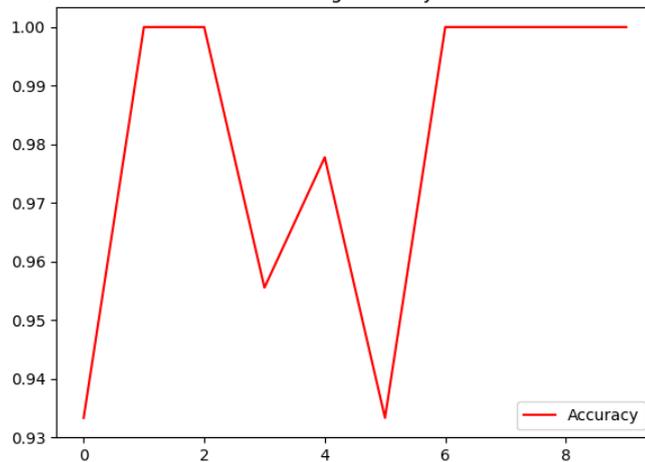


Figure 3. Training Accuracy

The Bayesian optimization process was crucial in identifying the optimal threshold for similarity assessment. The analysis revealed that the best threshold was approximately 0.200149, which effectively optimized the balance between sensitivity and specificity in the model's predictions. The corresponding accuracy at this threshold was around 92.22%. Figure 4 illustrates the relationship between various thresholds and their associated accuracy values over multiple iterations, demonstrating how accuracy fluctuated with changes in the threshold. As the threshold increased, a notable decline in accuracy was observed, emphasizing the importance of selecting an appropriate threshold to maintain high classification performance.

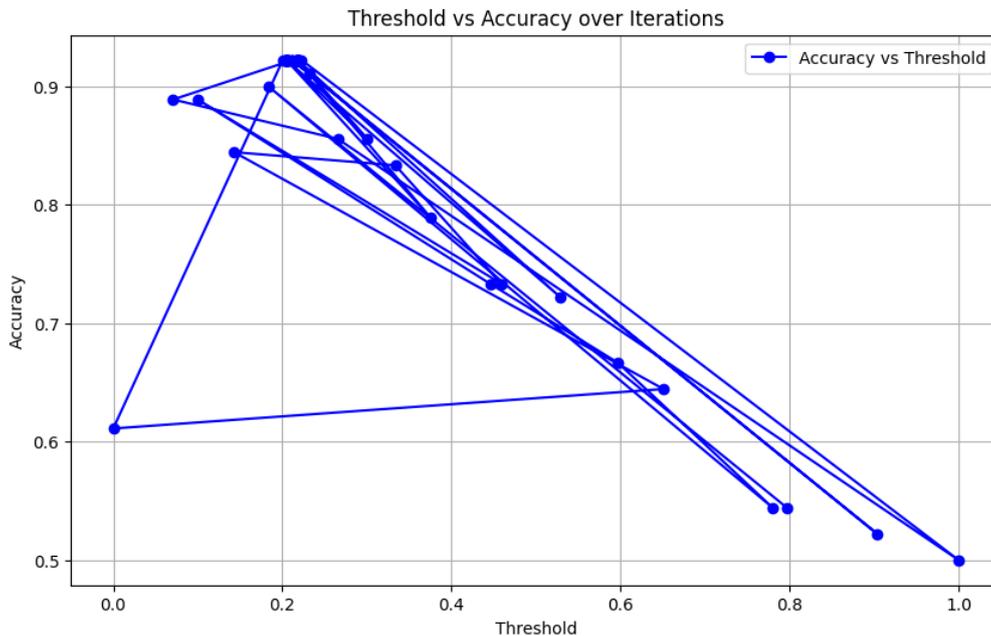


Figure 4. Threshold vs Accuracy over Iterations

The confusion matrix offers a more detailed breakdown of the model's performance. It reveals that 46.67% of predictions were true similarities, meaning the model correctly identified logos that were similar. The false similar rate was only 3.33%, indicating that the model rarely misclassified different logos as similar. The false different rate stood at 4.44%, suggesting there is still room for improvement in detecting all similar logos. The true different rate was 45.56%, demonstrating that the model was also effective at correctly identifying logos that were different, as shown in Figure 5.

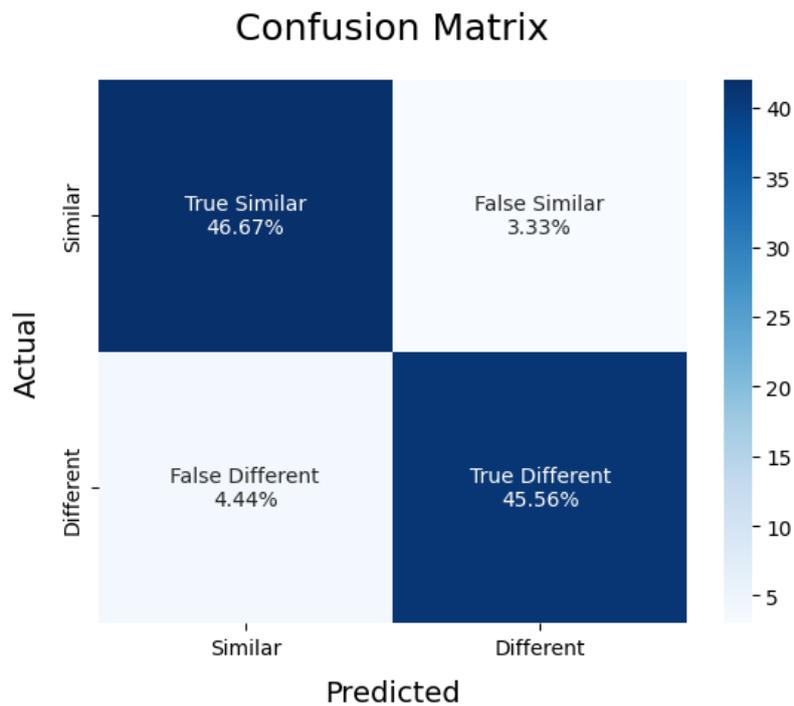


Figure 5. Confusion Matrix

From the confusion matrix, several performance metrics were calculated. The accuracy of the model was 92.23%, reflecting the proportion of correct predictions out of the total data. Precision, which measures the accuracy of the model in predicting the similar class, was found to be 93.34%. Recall, evaluating the model's ability to identify all instances in the similar class, was recorded at 91.44%. The F1-score, harmonizing precision and recall into a single comprehensive measure, was calculated at 92.39%.

The results indicate that the proposed approach is effective in detecting trademark logo infringements. The model's performance metrics suggest that it can serve as a valuable tool in trademark management, aiding in the protection of brand identity and reducing the risk of consumer confusion.

CONCLUSION

This research successfully developed a robust framework for detecting trademark logo infringements by employing a combination of Siamese Neural Networks and Bayesian optimization. The findings indicate that the Siamese Neural Network effectively learned to distinguish between similar and different logos, achieving an average testing accuracy of 92.22% over 10 epochs, with a significant reduction in training loss. The Bayesian optimization process established an optimal threshold of approximately 0.200149, which balances sensitivity and specificity, thereby minimizing false positives and false negatives in trademark assessments. The confusion matrix revealed that the model correctly identified 46.67% of true similarities and 45.56% of true differences, with low rates of false similarities (3.33%) and false differences (4.44%), demonstrating its reliability. The calculated performance metrics—accuracy (92.23%), precision (93.34%), recall (91.44%), and F1-score (92.39%)—further underscore the model's effectiveness in identifying trademark infringements. These findings highlight the potential of advanced machine learning techniques to enhance trademark management practices, serving as a valuable tool for brand owners in monitoring and protecting their intellectual property rights. Future research should consider applying this framework to larger and more diverse datasets to validate its effectiveness further, as well as integrating other machine learning techniques to enhance robustness and adaptability in various trademark scenarios. Overall, this study addresses the critical issue of trademark infringement detection, contributing valuable insights to the field of intellectual property protection.

REFERENCES

[1] S. Cho and A. Moorman, "An examination of the psychometrical comparability of survey evidence in sport trademark litigation," *J. Legal Aspects Sport*, vol. 24, no. 1, pp. 3-23, 2014. doi: 10.1123/jlas.2013-0010.

- [2] J. Kim, M. Ricci, and T. Serre, "Not-So-CLEVR: learning same-different relations strains feedforward neural networks," *Interface Focus*, vol. 8, no. 4, p. 20180011, 2018.
- [3] B. Hanberry and H. He, "Prevalence, statistical thresholds, and accuracy assessment for species distribution models," *Web Ecol.*, vol. 13, no. 1, pp. 13-19, 2013. doi: 10.5194/we-13-13-2013.
- [4] S. Uchida, C. Shimada, N. Sakuma, F. Kagitani, A. Kan, and S. Awata, "Olfactory function and discrimination ability in the elderly: a pilot study," *J. Physiological Sci.*, vol. 72, no. 1, 2022. doi: 10.1186/s12576-022-00832-6.
- [5] F. Horen and R. Pieters, "When high-similarity copycats lose and moderate-similarity copycats gain: the impact of comparative evaluation," *J. Mark. Res.*, vol. 49, no. 1, pp. 83-91, 2012. doi: 10.1509/jmr.08.0405.
- [6] C. Colliander and P. Ahlgren, "Experimental comparison of first and second-order similarities in a scientometric context," *Scientometrics*, vol. 90, no. 2, pp. 675-685, 2011. doi: 10.1007/s11192-011-0491-x.
- [7] F. Anuar, R. Setchi, and Y. Lai, "Semantic retrieval of trademarks based on conceptual similarity," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 46, no. 2, pp. 220-233, 2016. doi: 10.1109/TSMC.2015.2421878.
- [8] Q. H. Nguyen et al., "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil," *Math. Problems Eng.*, vol. 2021, no. 1, p. 4832864, 2021.
- [9] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Comput. Sci.*, vol. 10, e2245, 2024.
- [10] A. H. Efat, S. M. Hasan, M. P. Uddin, and M. A. Mamun, "A multi-level ensemble approach for skin lesion classification using customized transfer learning with triple attention," *PLoS One*, vol. 19, no. 10, e0309430, 2024.
- [11] D. Chicco, "Siamese neural networks: An overview," in *Artificial Neural Networks*, pp. 73-94, 2021.
- [12] M. Shorfuzzaman and M. S. Hossain, "MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients," *Pattern Recognit.*, vol. 113, p. 107700, 2021.
- [13] K. Martin, N. Wiratunga, S. Sani, S. Massie, and J. Clos, "A convolutional Siamese network for developing similarity knowledge in the SelfBACK dataset," in *CEUR Workshop Proceedings*, June 2017.
- [14] M. Zhang, Q. Cheng, F. Luo, and L. Ye, "A triplet nonlocal neural network with dual-anchor triplet loss for high-resolution remote sensing image retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2711-2723, 2021.
- [15] Y. Qiu, T. Misu, and C. Busso, "Use of triplet-loss function to improve driving anomaly detection using conditional generative adversarial network," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1-7, Sep. 2020.
- [16] Suyahman, Sunardi, Murinto, and Khusna, A. N. 2024. "Siamese Neural Networks with Chi-Square Distance for Trademark Image Similarity Detection." *Scientific Journal of Informatics* 11 (2): 429-438.
- [17] P. I. Frazier, "A tutorial on Bayesian optimization," arXiv preprint arXiv:1807.02811, 2018.
- [18] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proc. IEEE*, vol. 104, no. 1, pp. 148-175, 2015.
- [19] P. I. Frazier, "Bayesian optimization," in *Recent Advances in Optimization and Modeling of Contemporary Problems*, pp. 255-278, INFORMS, 2018.
- [20] P. Feliot, J. Bect, and E. Vazquez, "A Bayesian approach to constrained single-and multi-objective optimization," *J. Global Optim.*, vol. 67, no. 1, pp. 97-133, 2017.
- [21] J. Berk, V. Nguyen, S. Gupta, S. Rana, and S. Venkatesh, "Exploration enhanced expected improvement for Bayesian optimization," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part II*, vol. 18, pp. 621-637, Springer International Publishing, 2019.
- [22] R. B. Gramacy, *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. Chapman and Hall/CRC, 2020.
- [23] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proc. IEEE*, vol. 104, no. 1, pp. 148-175, 2015.
- [24] G. Naidu, T. Zuva, and E. M. Sibanda, "A review of evaluation metrics in machine learning algorithms," in *Computer Science On-line Conference*, pp. 15-25, Cham: Springer International Publishing, Apr. 2023.
- [25] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," arXiv preprint arXiv:2008.05756, 2020.

- [26] R. Barandela, J. S. Sánchez, V. García, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognit.*, vol. 36, no. 3, pp. 849-851, 2003.
- [27] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *J. Mach. Learn. Res.*, vol. 20, no. 177, pp. 1-81, 2019.
- [28] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognit. Lett.*, vol. 30, no. 1, pp. 27-38, 2009.
- [29] S. K. Papia, M. A. Khan, T. Habib, M. Rahman, and M. N. Islam, "DistilRoBiLSTMFuse: an efficient hybrid deep learning approach for sentiment analysis," *PeerJ Comput. Sci.*, vol. 10, e2349, 2024.