JOIN (Jurnal Online Informatika)
Volume 7 No. 1 | June 2022: 62-69
DOI: 10.15575/join.v7i1.839

p-ISSN: 2527-1682
e-ISSN: 2527-9165

Analysis of the Combination of Naïve Bayes and MHR (Mean of Horner's Rule) for Classification of Keystroke Dynamic Authentication

Zamah Sari[1], Didih Rizki Chandranegara[2], Rahayu Nurul Khasanah[3], Hardianto Wibowo[4], Wildan Suharso[5]
[1,2,3,4,5]Department of Informatics, University of Muhammadiyah Malang

Article Info

Article history:

Received December 01, 2021
Revised March 05, 2022
Accepted March 30, 2022
Published June 30, 2022

Keywords:

Keystroke Dynamic
Authentication
Classification
Naïve Bayes
Mean of Horner's Rule

ABSTRACT

Keystroke Dynamics Authentication (KDA) is a technique used to recognize somebody dependent on typing pattern or typing rhythm in a system. Everyone's typing behavior is considered unique. One of the numerous approaches to secure private information is by utilizing a password. The development of technology is trailed by the human requirement for security concerning information and protection since hacker ability of information burglary has gotten further developed (hack the password). So that hackers can use this information for their benefit and can disadvantage others. Hence, for better security, for example, fingerprint, retina scan, et cetera are enthusiastically suggested. But these techniques are considered costly. The advantage of KDA is the user would not realize that the system is using KDA. Accordingly, we proposed the combination of Naïve Bayes and MHR (Mean of Horner's Rule) to classify the individual as an attacker or a non-attacker. We use Naïve Bayes because it is better for classification and simple to implement than another. Furthermore, MHR is better for KDA if combined with the classification method which is based on previous research. This research showed that False Acceptance Rate (FAR) and Accuracy are improving than the previous research.

Corresponding Author:

Didih Rizki Chandranegara
Department of Informatics
University of Muhammadiyah Malang
Jl. Raya Tlogomas No.246 Malang, Jawa Timur
didihrizki@umm.ac.id

1. INTRODUCTION

Keystroke Dynamic Authentication (KDA) is a technique used to recognize somebody on his or her typing pattern [1]. KDA also can be defined as an automatic method to identify or to confirm the identity of someone based on typing way and typing rhythm on the keyboard. Every user's typing behavior is considered unique. Along these lines, this uniqueness of typing rhythm can be utilized as an establishment for password security.

The development of technology is trailed by the human requirement for security concerning information and protection since hacker ability of information burglary has gotten further developed. One of the numerous approaches to secure private information is by utilizing a password. Hence, for better security, for example, fingerprint scan, retina scan, et cetera are enthusiastically suggested. But these techniques are considered costly. Regarding the problems which have been mentioned before, we use KDA because the user would not realize that the system is using KDA.

In this research, we propose research to classify users using the combination of Naïve Bayes and MHR (Mean of Horner's Rule) to determine whether someone is an attacker or non-attacker based on the typing rhythm. The KDA will only extract the feature of typing characteristics from every user. It allows the authentication system to be more affordable since it does not require any extra sensor, which will ease the user [2].

Naïve Bayes (NB) is one of many methods of classification and simple to implement. Furthermore, NB will calculate various probabilities by calculating the frequency and value combination within the available dataset [3]. NB also has been used as the implementation of KDA. Previous research by Hoobi [3] shows the researcher did user authentication to avoid unknown access which is not recognized by the system. The result from the research [3] shows that feature combination (DT (Down Time) – DDT (Down-Down Time)) gave a better result with low mistakes. Thus, this result is better than only using DT or only using DDT with 94.1% accuracy. On the other hand, the Mean of Horner's Rule (MHR) is one method to determine the mean. MHR will provide a better result in calculating stream data. This research about the combination of NB and MHR is based on the results of research by Chandranegara and Sumadi [4]. In this research [4], researchers made the application regarding the algorithm combination of MHR and Standard Deviation. The research [4] shows a positive result with 90% accuracy and decreases the False Acceptance Rate (FAR) from 0.114 to 0.113. Thus, it shows that the combination of the classification method and MHR can increase the accuracy of previous research.

Based on Zainab Mahmood Fadhil's [5] research, focuses on how the incorporation of K-Means clustering and the Naïve Bayes (NB) classification increase the exactness of employee performance predicting. The result from the research shows that the feature combination K-Means and Naïve gave a better result with 80%, 91.29%, 98.56%, and 92.24% accuracy. Thus, it shows that the combination of Naïve Bayes classification can increase the accuracy.

In V.Vanita and D.Akila [6] research focuses on using tools and techniques to manipulate image processing results, pattern recognition results, and classification methods and then validating the image classification result against medical expert expertise with hybrid combining NBC. The result from the research shows the feature combination hybrid (Random Forest + FCM) segmentation and Naive Bayes classification gave a better result with 89.4% accuracy.

On the other hand, In 2019 this paper by Mochamad Wahyudi and Anik Andriani [7] the objective of this thesis is to create classification from diarrhea outbreaks data to obtain the data patterns in the form of classification rule that can be applied to detect Case Fatality Rate of diarrhea using C4.5 algorithm and Naïve Bayes algorithm. The research shows a result of 82.82% accuracy used C4.5 algorithm and 79.85% accuracy used Naïve Bayes algorithm.

In this research, we proposed a combination of Naïve Bayes and MHR to classify the individual as an attacker or a non-attacker. The combination will do the labeling process to assume attacker and non-attacker which will be explained in the next section. This combination is based on previous research [4]. We will do the False Acceptance Rate (FAR), False Rejected Rate (FRR), and accuracy to determine our method is better than the previous research. The reason for using FAR, FRR, and accuracy is because previous research [4] is used to determine it.

2. MATERIAL & METHOD

2.1. Dataset

This research uses the Keystroke Dynamics Authentication dataset that can be accessed from Kevin S. Killourhy and Roy A. Maxion's research [8]. This dataset consists of data such as password typing time from 51 users. These 51 users comprised 30 male users and 21 female users. Every user will be instructed to type ".tie5Roanl" as the password 400 times within eight sessions. This dataset comprises 31 features such as hold time, latency time, and flight time. The second will be used as the unit of time [4][8].

2.2. Keystroke Dynamic Authentication (KDA)

Keystroke Dynamic Authentication (KDA) is a method to identify someone based on his or her typing pattern [1]. This typing pattern is uniquely similar to handwriting since the setting has a neurophysiologic mechanism. Keystroke Dynamic Authentication consists of information that works as a biometric identifier and can recognize certain keyboard users [8][9]. Different people have different typing behavior; hence the input of someone's typing behavior will work as second-factor authentication to strengthen password security. Thus, every person's typing behavior is considered unique. The hacker should know precisely the particular typing behavior of someone to hack.

Keystroke Dynamic Authentication (KDA) uses some important features such as Latency time and Hold Time [4][8]. Latency time is the duration to release a character until pressing the next character, and hold time is the duration to press and release a character. Additionally, besides latency time and hold time, there is flight time which is the duration of a certain character is pressed to the next character is pressed. A further illustration can be seen in Figure 1.

e-ISSN 2527-1682
p-ISSN 2527-9165

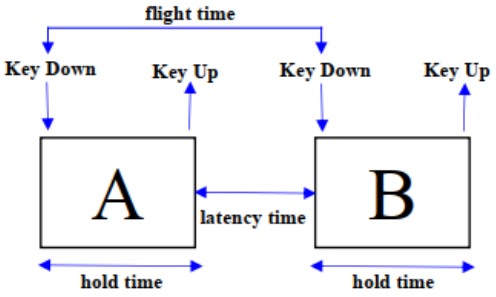

Figure 1. Hold Time and Latency Time illustration [4]

2.3. MHR (Mean of Horner's Rule)

Mean of Horner's Rule (MHR) is one of the methods to determine the mean. The MHR in this research will be combined with the Naïve Bayes method. The following is the MHR formula [4]:

$$MHR = \frac{\left(\frac{\left(\frac{\left(\frac{x_1+x_2}{2}\right)+x_3}{2}\right)+x_4}{\dots}\right)+x_n}{2} \quad (1)$$

2.4. Naïve Bayes Classifier (NBC)

Naïve Bayes Classifier (NBC) is a simple statistical classification of probability based on the application of the Bayes probability theorem [3][10]. Naïve Bayes is a simple probabilistic classifier based on the application of the Bayes theorem [11]. Naïve Bayes is also known to perform very well with simple calculations [12]. The advantage of Naïve Bayes is that it only requires a small amount of training data to estimate the parameters required for classification [13]. NB can calculate a set of probabilities by calculating the frequency and combination of values in a given dataset [14][15]. Naïve Bayes theory can be described as follows:

$$P(X) = \frac{P(C)\,P(C)}{P(X)} \quad (2)$$

Where,
C = Hypothesis Data which is a specific class
P(C|X) = Posterior probability
P(X|C) = Probability based on the condition of hypothesis
P(C) = Hypothesis probability
P(X) = Probability A

2.5. Combination Naïve Bayes and Mean of Horner Rules (MHR)

In this research, we propose a combination of data labeling that determines the attacker or non-attacker then classified using Naïve Bayes. The following is a labeling process that still uses an average based on previous research [4] (where Ø = threshold and we will discuss in the next section):

1. If the Average value > Ø, therefore assumed as non-attacker.
2. If the Average value < Ø, therefore assumed as the attacker.

The labeling process that we will use is as follows:

1. If the MHR value>Ø, therefore assumed as non-attacker.
2. If the MHR value<Ø, therefore assumed as the attacker.

2.6. Test Scenario and Evaluation

The test is conducted using a programming language of Python and divides the dataset into testing data and training data. The scenario can be seen as follow:

1. Training data are acquired from 1st to 350th data.
2. Testing data are acquired from 351st to 400th data.

3. Every user will be tested toward the testing data.

After the clarification result is acquired, the evaluation will determine how efficient the method is and compare which method is better to recognize attackers and non-attacker.

Tests were carried out using several test size and labeling scenarios—the first scenario using test data of 10% and training data of 90%. The second scenario uses 20% test data and 80% training data. The third scenario uses 30% test data and 70% training data. Each of these scenarios uses a different threshold (Ø) starting with 0 until 1.

For evaluation, we use Accuracy, False Acceptance Rate (FAR), and False Rejected Rate (FRR). The accuracy measurement is using TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative), which get based on Table 1. To get this value, we can use the following formula [4]:

$$accuracy = \left(\frac{TP+TN}{TP+FP+TN+FN}\right) x\ 100\% \quad (3)$$

Table 1. Confusion Matrix

	User (Prediction)	**Attacker (Prediction)**
User (Actual)	True Positive (TP)	False Negative (FN)
Attacker (Actual)	False Positive (FP)	True Negative (TN)

Then, FAR in this research is the value of the classification mistake in determining the attacker as the user. The smaller the FAR value means the system is better to avoid attackers. To get this value, we can use the following formula [4]:

$$FAR = \left(\frac{FP}{TN+FP}\right) \quad (4)$$

FFR in this research is the value of the classification mistake in determining the user as the attacker. The smaller FFR value means the system is better able to accept users/non-attacker. To get this value, we can use the following formula [4]:

$$FRR = \left(\frac{FN}{TP+FN}\right) \quad (5)$$

Where,
TP = Correct prediction towards a user
TN = Correct prediction towards an attacker
FP = Wrong prediction towards attacker assumed as user
FN = Wrong prediction towards user assumed as an attacker

3. RESULTS AND DISCUSSION

In this research, the test is done by using two processes which are Naïve Bayes classification and Naïve Bayes classification combined with Mean of Horner's Rule (MHR). Each of these scenarios was tested using a different threshold (Ø), and the best threshold results are 0.20, 0.21, 0.22, 0.25, 0.26, 0.27, and 0.3. The reason for using these numbers i.e.:

1. If using less than 0.20, then the labeling results show more attackers than non-attackers. This means that the method is more likely to accept the attacker as a user or non-attacker.
2. If using 0.23, 0.24, 0.28, 0.29 and above 0.30 it also shows the same labeling results as below 0.20.'

So, the threshold number that we use is a stable threshold based on the evaluation results which will be discussed in the next section.

3.1. Accuracy

From Figures 2, 3, and 4, it can be seen that the Naïve Bayes method of combination MHR is better than the Naïve Bayes Method. This is because the accuracy value of the Naïve Bayes combination of MHR (with the mean results of scenarios 1, 2, and 3 is 94%) is higher than the accuracy value of Naïve Bayes (with an average result of scenario 1,2,3, is 86 %). From the results of the accuracy of scenario 3 (because it has the highest accuracy value), it will be compared with previous research. The results of the accuracy of previous research are shown in table 2.

In scenarios one, two, and three, the results of the Naïve Bayes combination MHR method have the best accuracy value, where the accuracy value is 94\% compared to the accuracy value of the Naïve Bayes method. This is consistent with research [4] which explains that the use of the Classification Method (Naïve

e-ISSN 2527-1682
p-ISSN 2527-9165

Bayes) combination MHR is more suitable for data streams such as KDA, which has the characteristics of changing rapidly over time.

However, there are two scenarios where the Naïve Bayes results get higher accuracy than the Naïve Bayes combination MHR, namely in scenarios one and three, where the labeling is 0.20. This is because the number of attackers at the labeling value 0.20 is more than the number of attackers at the other labeling values with fewer attackers. So, it can be concluded that the labeling value of 0.20 is more suitable for use with the Naïve Bayes method. For better security in the system, this labeling value (0.20) is suitable to implement in the system using KDA (using our proposed method). Table 2 shows the results of the accuracy of previous research.

Table 2. Confusion Matrix

Method	Mean of Accuracy (%)
Joyce & Gupta [16]	75.388
Yang & Fang [17]	75.156
Didih & Fauzi [4]	93.872
Naïve Bayes	86%
Naïve Bayes & MHR	94%

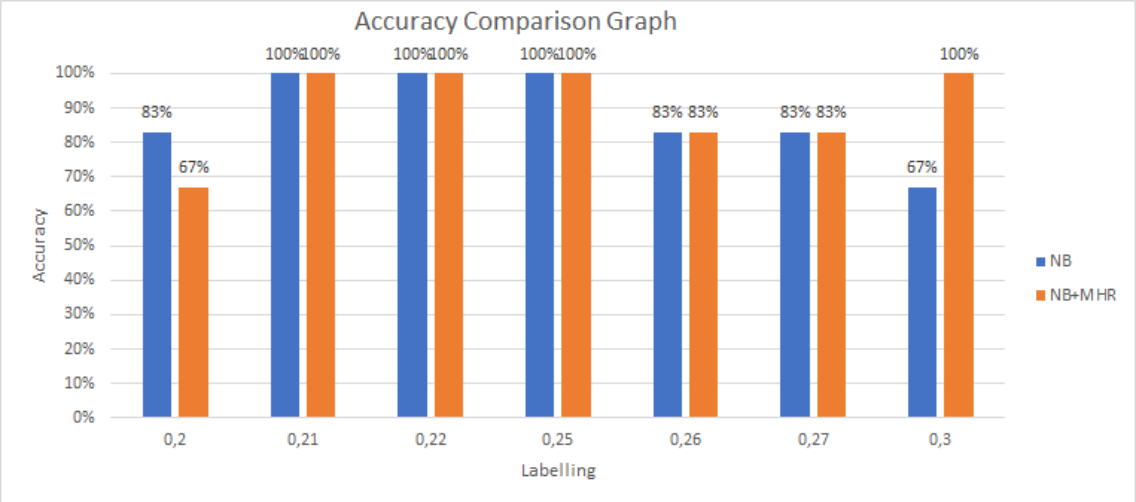

Figure 2. Scenario 1 of Accuracy Comparison Graph between Naïve Bayes and Naïve Bayes Combined MHR

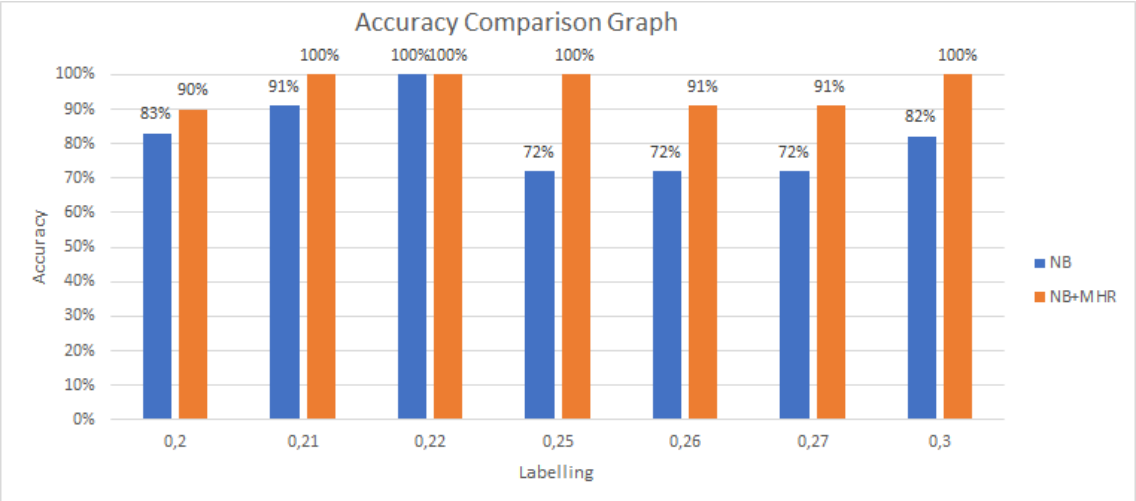

Figure 3. Scenario 2 of Comparison Graph between Naïve Bayes and Naïve Bayes Combined MHR

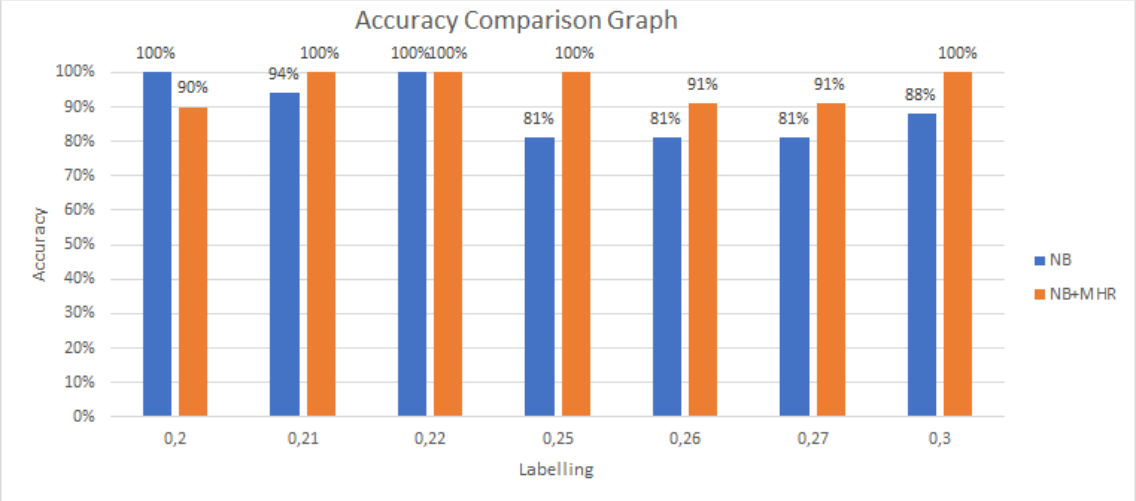

Figure 4. Scenario 3 of Comparison Graph between Naïve Bayes and Naïve Bayes Combined MHR

3.2. FAR (False Acceptance Rate)

Tables 3 and 4 show that there are some False Acceptance Rate (FAR) values that get a value of 0, but our proposed method, namely Naïve Bayes and MHR, shows the number of 0 values generated. This indicates that our proposed method is perfect for rejecting attackers who will try to log into the system as a user.

Table 5 is the result of a comparison with several previous research. From table 5, it can be seen that using the Naïve Bayes method and the Naïve Bayes method, the combination of Mean of Horner's Rule (MHR) can reduce the FAR value from previous studies [4]. However, from the two methods, using the Naïve Bayes method, the combination of the Mean of Horner's Rule (MHR), which is our proposed method, shows that the results are better at reducing the FAR value than the Naïve Bayes method. This result is also shown in the accuracy, where the labeling value 0.20 is the best labeling for the proposed method we use.

Table 3. FAR Naïve Bayes from All Scenario

Scenario	Labeling Value						
	0.2	0.21	0.22	0.25	0.26	0.27	0.3
1	0.25	0	0	0	0.2	0.2	0.33
2	0.25	0	0	0.272	0.272	0.272	0.181
3	0	0	0	0.187	0.187	0.187	0.125

Table 4. FAR Naïve Bayes Combination MHR from All Scenario

Scenario	Labeling Value						
	0.2	0.21	0.22	0.25	0.26	0.27	0.3
1	0	0	0	0	0.166	0.166	0
2	0	0	0	0	0.09	0.09	0
3	0	0	0	0	0.09	0.09	0

Table 5. FAR Result Comparison from All Scenario using Mean of Accuracy

Method	Mean of Accuracy (%)
Joyce & Gupta [16]	0.432
Yang & Fang [17]	0.141
Didih & Fauzi [4]	0.113
Naïve Bayes	0.14
Naïve Bayes & MHR (our research)	0.03

3.3. FRR (False Rejected Rate)

Table 6 and 7 shows many 0 values of False Rejected Rate (FRR). As well as FAR value, Naïve Bayes and researcher's suggested method show a better FRR value than previous research. However, research [16], our method cannot approach 0 value. On the other hand, this is not something to be worried about since, fundamentally, KDA is a method to reject attackers to log into the system as a user. Furthermore, research [16]

e-ISSN 2527-1682
p-ISSN 2527-9165

has a disadvantage of high FAR value while our research has a smaller FAR value and almost similar FFR value. In addition, Table 8 compares with several previous research and our research-based on FRR using the mean of all scenario results. The result of this evaluation is different from accuracy and FAR because the result does not show the best from many labeling values. But it is not a problem, because the most significant aspect of KDA is how to reject attackers from login in the system.

Table 6. FRR Naïve Bayes from All Scenario

Scenario	Labeling Value						
	0.2	0.21	0.22	0.25	0.26	0.27	0.3
1	0	0	0	0	0	0	0
2	0	0.333	0	0	0	0	0
3	0	0.25	0	0	0	0	0

Table 7. FRR Naïve Bayes Combination MHR from All Scenario

Scenario	Labeling Value						
	0.2	0.21	0.22	0.25	0.26	0.27	0.3
1	0.5	0	0	0	0	0	0
2	0.25	0	0	0	0	0	0
3	0.25	0	0	0	0	0	0

Table 8. FRR Result Comparison from All Scenario using Mean of Accuracy

Method	Mean of Accuracy (%)
Joyce & Gupta [16]	0
Yang & Fang [17]	0.004
Didih & Fauzi [4]	0.009
Naïve Bayes	0.03
Naïve Bayes & MHR	0.05

4. CONCLUSION

Based on the result of this research, the combination of Naïve Bayes classification and Means of Horner's Rule (MHR) has a better accuracy result than previous research. The False Acceptance Rate (FAR) value is the smallest value if compared to the previous research. Additionally, the False Rejected Rate (FRR) shows a relatively small value even though not as small as previous research. However, this is not a problem since KDA is the way to reject attackers to log into the system. Attackers cannot take a benefit because all typing activities are recorded using KDA and users will feel safe using the most frequently used passwords. So, it can improve the security of the system using KDA and can reduce hacking issues on user accounts in the system. The last, labeling value 0.20 on this research gives the best results using our proposed method than the previous method. For future research, it is expected to add more datasets from other KDA research so it can be the comparison regarding the accuracy with this research and also other research.

5. REFERENCES

[1] B. S. Saini, N. Kaur, and K. S. Bhatia, "Keystroke dynamics based user authentication using numeric keypad," *Proc. 7th Int. Conf. Conflu. 2017 Cloud Comput. Data Sci. Eng.*, pp. 25–29, 2017, doi: 10.1109/CONFLUENCE.2017.7943118.

[2] Y. Muliono, H. Ham, and D. Darmawan, "Keystroke Dynamic Classification using Machine Learning for Password Authorization," *Procedia Comput. Sci.*, vol. 135, pp. 564–569, 2018, doi: 10.1016/j.procs.2018.08.209.

[3] M. M. Hoobi, "Keystroke Dynamics Authentication based on Naïve Bayes Classifier ادانتسا ىلع ةيكيمانيد ةطغض حاتفملا ةقداصم يبوه دمحم سيم زيادبا فيند فنصم مادختسابا مدختسملا ةقداصم," vol. 56, no. 2, pp. 1176–1184, 2015.

[4] D. R. Chandranegara, H. Wibowo, and A. E. Minarno, "Combined scaled manhattan distance and mean of horner's rules for keystroke dynamic authentication," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, pp. 770–775, 2020, doi: 10.12928/TELKOMNIKA.v18i2.14815.

[5] Z. M. Fadhil, "Hybrid of K-means clustering and naive Bayes classifier for predicting performance of an employee," *Period. Eng. Nat. Sci.*, vol. 9, no. 2, pp. 799–807, 2021, doi: 10.21533/pen.v9i2.1898.

[6] V. Vanitha and D. Akila, "Image Segmentation and classification Hepatitis viral infection in human blood smear with a hybrid algorithm combining Naive Bayes Classifier Input Image Image preprocessing Random Forest Naive Bayes Classifier Feature extraction Clustering Image Result," vol. 12, no. 11, pp. 5873–5881, 2021.

[7] M. Wahyudi and A. Andriani, "Application of C4.5 and Naïve Bayes Algorithm for Detection of Potential Increased Case Fatality Rate Diarrhea," *J. Phys. Conf. Ser.*, vol. 1830, no. 1, pp. 0–12, 2021, doi: 10.1088/1742-6596/1830/1/012016.
[8] K. S. Killourhy and R. A. Maxion, "Comparing anomaly-detection algorithms for keystroke dynamics," *Proc. Int. Conf. Dependable Syst. Networks*, pp. 125–134, 2009, doi: 10.1109/DSN.2009.5270346.
[9] Y. Zhong, Y. Deng, and A. K. Jain, "Keystroke dynamics for user authentication," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, no. June, pp. 117–123, 2012, doi: 10.1109/CVPRW.2012.6239225.
[10] Z. E. Rasjid and R. Setiawan, "Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques," *Procedia Comput. Sci.*, vol. 116, pp. 107–112, 2017, doi: 10.1016/j.procs.2017.10.017.
[11] T. Anusas-Amornkul, "Strengthening password authentication using keystroke dynamics and smartphone sensors," *ACM Int. Conf. Proceeding Ser.*, pp. 70–74, 2019, doi: 10.1145/3357419.3357425.
[12] M. S. Mubarok, A. Adiwijaya, and M. D. Aldhi, "Aspect-based sentiment analysis to review products using Naïve Bayes," *AIP Conf. Proc.*, vol. 1867, 2017, doi: 10.1063/1.4994463.
[13] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier," *Int. J. Inf. Eng. Electron. Bus.*, vol. 8, no. 4, pp. 54–62, 2016, doi: 10.5815/ijieeb.2016.04.07.
[14] O. A.-M. A. H. Mohammad, T. Alwada'n, "Arabic Text Categorization Using Support vector machine, Naïve Bayes and Neural Network Adel," *GSTF J. Comput.*, vol. Volume 5, no. 1, pp. 40–44, 2016, doi: 10.5176/2251-3043.
[15] N. F. Rusland, N. Wahid, S. Kasim, and H. Hafit, "Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 226, no. 1, 2017, doi: 10.1088/1757-899X/226/1/012091.
[16] R. Joyce and G. Gupta, "Identity Authentication Based on Keystroke Latencies," *Commun. ACM*, vol. 33, no. 2, pp. 168–176, 1990, doi: 10.1145/75577.75582.
[17] W. YANG and F. FANG, "Application of a Dynamic Identity Authentication Model Based on an Improved Keystroke Rhythm Algorithm," *Int. J. Commun. Netw. Syst. Sci.*, vol. 02, no. 08, pp. 714–719, 2009, doi: 10.4236/ijcns.2009.28082.