**TAMIKA**: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akuntansi
**Vol. 4 No. 2 (Desember 2024)** – ***Edisi Khusus SEMNASTIK 2024***
https://doi.org/10.46880/tamika.Vol4No2(SEMNASTIK).pp168-174 e-ISSN: 2828-1276

# COMPARISON OF SVM, KNN, AND NAÏVE BAYES ALGORITHMS IN MONKEYPOX DISEASE CLASSIFICATION

**Kelvin Leonardi Kohsasih**✉
Informatics Engineering Department, STMIK TIME, Medan, Indonesia
Email: enjun@upi.edu

## *ABSTRACT*

*Advances in medical technology have enabled the application of machine learning for disease classification, including monkeypox. Monkeypox is a zoonotic disease caused by the monkeypox virus and can be detected through patient data. This study aims to compare the performance of Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Naïve Bayes algorithms in building a monkeypox classification model. The dataset used consists of 25,000 patient records. The results show that the SVM model with a linear kernel achieved the best accuracy compared to KNN and Naïve Bayes. These findings demonstrate that the SVM model with a linear kernel is highly effective in classifying monkeypox, offering great potential for further medical applications.*

***Keyword: Monkeypox Disease, Machine Learning, Support Vector Machine, K-Nearest Neighbors, Naïve Bayes.***

## INTRODUCTION

The unexpected appearance of the monkeypox virus (MPXV-2022) in several non-endemic countries has caused global concern in recent times. Monkeypox virus (MPXV) belongs to the Orthopoxvirus genus within the Poxviridae family. It is a DNA virus closely related to the smallpox virus (Altindis et al., 2022). Epidemiological data reveal that MPXV was initially discovered in Central and West Africa. The first case of MPXV infection reported outside Africa was in the United States (Wang et al., 2023). In 2017, Nigeria experienced an outbreak of MPXV, which was later followed by a case in the United Kingdom associated with recent travel to Nigeria (Alakunle et al., 2020).

According to the WHO publication report for the period from January 1, 2022, to June 30, 2024, there have been a total of 99,176 confirmed monkeypox cases, including 208 deaths, across 116 locations (World Health Organization, 2024). The most common symptoms of monkeypox are: fever, headache, muscle aches, back discomfort, low energy, and swollen lymph nodes. These symptoms are often followed by a rash appearing on the face, palms, soles, groin, genital area, and/or anal regions (Haque et al., 2022; Huang et al., 2022). Efforts are ongoing to curb the spread of monkeypox, with a focus on improving early detection.

Advances in medical technology have made it possible to apply machine learning to classify diseases such as breast cancer, health diseases, HIV/AIDS, and even the COVID-19 virus (Adapala et al., 2023; Belete & Huchaiah, 2021; Nagavelli et al., 2022; Shaban et al., 2021). Machine learning algorithms are essential for accurate prediction and accurate analysis (Kohsasih & Situmorang, 2022). There are many studies that have applied algorithms to classify medical diseases. In 2024, Anugrah W et al. conducted a study on the classification of monkeypox using an SVM algorithm with RBF kernel. The study found that the developed model achieved an accuracy rate of 65% when the SVM parameters were set as C=10 and gamma (γ)=1(Anugrah et al., 2024). The following study presents a Bayesian Optimization-Support Vector Machine (BO-SVM) model to classify individuals with Parkinson's disease, achieving optimal results through hyperparameter tuning for six machine learning models. Among these, the SVM model demonstrated the highest accuracy of 92.3% after optimization (Elshewey et al., 2023).

In 2023, Pattimura B et al, conducted research to identify monkeypox using the Naive Bayes algorithm. The study revealed that the classification of monkeypox image feature extraction with the Naive Bayes method achieved an accuracy of 75% (Bagas Pattimura et al., 2023). The following study presents a K-Nearest Neighbors (KNN) model for classifying cardiovascular diseases. The results indicate that the KNN model achieved an accuracy of 90%, precision of 89%, recall of 90%, and an F1-score of 90% (Artanti, 2024).

Based on previous research, many studies have utilized machine learning algorithms such as SVM, KNN, and Naive Bayes for disease prediction and classification. In this study, the researcher aims to classify monkeypox to develop a machine learning model capable of detecting the disease. Additionally, the researcher will analyze and compare performance metrics, such as accuracy, precision, and recall, of the

SVM, KNN, and Naive Bayes algorithms in predicting monkeypox.

## RESEARCH METHODS

Several stages were carried out in this research, starting with conducting a literature study on machine learning algorithms and monkeypox virus (MPXV-2022) and collecting datasets. The data is then subjected to preprocessing steps such as data selection, data cleaning, and data transformation. These processes are intended to ensure that the data used for analysis or model training is clean, relevant, and formatted optimally to achieve accurate and reliable results. The data will then be divided into three parts used for the training, testing and validation set by applying the machine learning algorithm such as SVM, KNN and Naive bayes. Finally, each architecture will have its performance evaluated. Figure 1 illustrates the procedure carried out in the study.

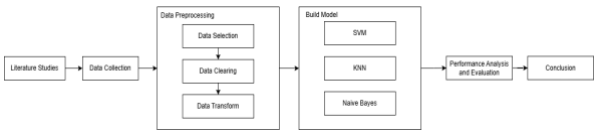



**Figure 1.** Summary of Research Methods

### Dataset

The dataset used in this study is the Monkey-Pox PATIENTS Dataset, a public dataset available on Kaggle. It includes records from 25,000 patients with relevant features and a target variable indicating whether the patient has monkeypox. The dataset comprises 11 attributes [16]. The attributes used in this research are presented in Table 1

**Table 1.** Dataset Attributes

| No | Attribute | Description |
|---|---|---|
| 1 | Patient_ID | Patients' Unique ID |
| 2 | Systemic Illness | Type of illness |
| 3 | Rectal Pain | Do they have Rectal Pain |
| 4 | Sore Throat | Do they have Sore Throat |
| 5 | Penile Oedema | Do they have Penile Oedema |
| 6 | Oral Lesions | Do they have Oral Lesions |
| 7 | Solitary Lesion | Do they have Solitary Lesion |
| 8 | Swollen Tonsils | Do they have Swollen Tonsils |
| 9 | HIV Infection | Do they have HIV Infection |
| 10 | Sexually Transmitted Infection | Do they have any sexually transmitted infection |
| 12 | MonkeyPox | Do they have MonkeyPox (Positive) or not (Negative) |

### Support Vector Machine

Support Vector Machine (SVM) algorithm is a widely-used machine learning technique for text classification and performs well across various domains (Espejel & Cantu-Ortiz, 2021). SVM adjusts the model to enable linear separation of the domain. SVM can be categorized into linear and nonlinear models. To address nonlinear problems, the kernel concept is applied in high-dimensional spaces to define the hyperplane, maximizing the margin between the data classes (Syafika & Karisma, 2023). A kernel function returns a value equivalent to the dot product of feature vectors mapped into a higher-dimensional space, without explicitly performing the mapping. It allows efficient computation and facilitates linear separation of data in higher dimensions by avoiding explicit processing of the vectors (Cichosz, 2015). Common kernel functions used in SVM include:

#### *Linear*

The linear kernel calculates the dot product of two input vectors in the original space without transforming them into a higher-dimensional feature space(Safitri et al., 2019). The mathematical representation of the linear kernel is given in equation 1.

$$K(x, y) = x^T\ y$$

Where:
$K(x, y)$ is the kernel function that measures the similarity between vectors x and y.
$x^T\ y$ is The dot product of vectors x and y, where $x^T$ is the transpose of x.

#### *Polinomial*

The polynomial kernel evaluates the polynomial relationship between two input vectors in the original space (Rabbani et al., 2023). The mathematical representation of the polinomial kernel is given in equation 2.

$$K(x, y) = (x \cdot y + c)^d$$

Where:
$K(x, y)$ is the kernel function that measures the similarity between vectors x and y.
$x \cdot y$ is the dot product of vectors x and y
c is A constant term that allows the kernel function to be shifted
d is the degree of the polynomial, which determines the complexity of the decision boundary.

***Radial Basis Function (RBF)***

The RBF kernel is the most frequently used method of kernelization in nonlinear scenarios because of its resemblance to the Gaussian distribution (Leni et al., 2023). The mathematical representation of the RBF kernel is given in equation 3.

$$K(X_1, X_2) = exp(-\gamma \,||\, X_1 - X_2||^2)$$

Where:
$K(X_1, X_2)$ is the kernel function measuring similarity between data points $X_1$ and $X_2$.
$X_1, X_2$ is input vectors (data points).
$\gamma$ is hyperparameter controlling the kernel's sensitivity to differences between $X_1$ and $X_2$.
$||\, X_1 - X_2||^2$is squared Euclidean distance, representing how far apart $X_1$ and $X_2$ are in the feature space.

***Sigmoid***

The sigmoid function, often used in neural networks, can resemble the RBF kernel and model complex nonlinear interactions. However, it may not always be suitable as it is not always positive definite. (Safitri et al., 2019). The mathematical representation of the RBF kernel is given in equation 4.

$$K(x, y) = \boldsymbol{tanh}(\alpha x^T y + c)$$

Where:
$K(x, y)$ is the kernel function that measures the similarity between vectors x and y.
$\alpha x^T y$ is the dot product of vectors x and y
α is scaling factor that controls the influence of the dot product in the kernel function.
c is constant term that shifts the result of the dot product before applying the hyperbolic tangent function.

**K-Nearest Neighbor**

K-Nearest Neighbor (KNN) is a supervised machine learning algorithm used for classification, where the class of a new test sample is determined by the majority category among its K-nearest neighbors in the dataset. Two critical factors that influence the performance of KNN are the distance function used and the chosen value of K. In practice, the Euclidean distance function is commonly used to measure the proximity between training data points and test data (Lonang et al., 2023). The mathematical representation of Euclidean distance is provided in the equation 5.

$$Euclidean\ d_{(a,b)} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

Where:
$d_{(a,b)}$ : distance
$a$ : Training data
$b$ : Testing data
$i$ : number of attributes
$n$ : dimension data

**Naïve Bayes**

Naïve Bayes is a classical machine learning algorithm based on Bayesian networks, commonly applied to classification tasks and known for its strong performance (Rochim et al., 2021). When using Naïve Bayes for classification, from a probabilistic perspective, it calculates the probability of an item belonging to each target category and then selects the highest probability to assign the item to the corresponding category as the classification result (Guo et al., 2023). The mathematical representation of Euclidean distance is provided in the equation 6.

$$P(Q \mid X) = \frac{P(X|Q)\ P(Q)}{P(X)}$$

Where:
$X$ : Data with an unknown class
$Q$ : Hypothesis $X$ for a specific class
$P(Q \mid X)$ : Probability of hypothesis Q given $X$
$P(Q)$ : Probability of hypothesis Q
$P(X \mid Q)$ : Probability of $X$ given hypothesis Q
$P(X)$ : Probability of $X$

**Evaluate Metrics**

The classification results will be evaluated using a confusion matrix. The confusion matrix summarizes classification performance by including predictions from the testing phase. Also known as a mistake matrix, it displays the results as TP (true positive), TN (true negative), FP (false positive), and FN (false negative) (Kohsasih et al., 2022). The confusion matrix is illustrated in Figure 2.

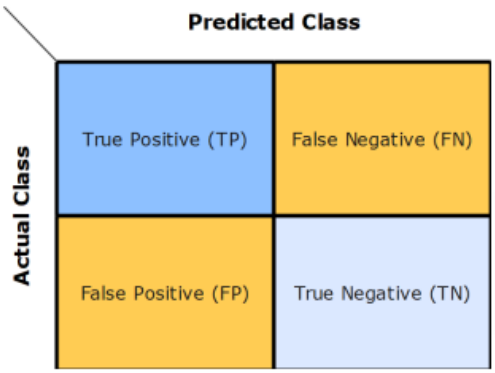



**Figure 2.** Illustration Confusion Matrix for Binary Classification

Based on the data from the confusion matrix, performance metrics such as accuracy, precision, recall, and F1-score can be calculated. The formulas for calculating these metrics are provided in equations (7), (8), (9), and (10), respectively.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

$$Precision = \frac{(TP)}{(TP + FP)}$$

$$Recall = \frac{(TP)}{(TP + FN)}$$

$$F1 - Score = 2 \times \frac{Precision\,.Recall}{Precision + Recall}$$

Where TP represents the number of correctly classified positive data points, TN refers to the number of correctly classified negative data points, FN is the number of incorrectly classified negative data points, and FP represents the number of false positive data points.

## RESULT AND DISCUSSION

In this study, the data used is a public dataset containing 25,000 monkeypox patient records. The data is divided into two classes, with 15,909 positive cases and 9,091 negative cases. Several preprocessing steps were applied, including data selection, where relevant attributes were chosen; data cleaning, which involved removing damaged or incomplete records; and data transformation, where the data was converted into numerical form. After preprocessing, the dataset size was reduced to 18,784 records, consisting of 12,585 positive cases and 6,199 negative cases. The preprocessed data is illustrated in Figure 3.

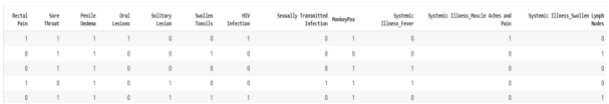


**Figure 3.** Illustration of Preprocessed Data

The processed data was then split into three subsets: a training set, a validation set, and a testing set. The training set was used to train the model, the validation set to optimize model performance during training, and the testing set to evaluate the overall performance of the model.

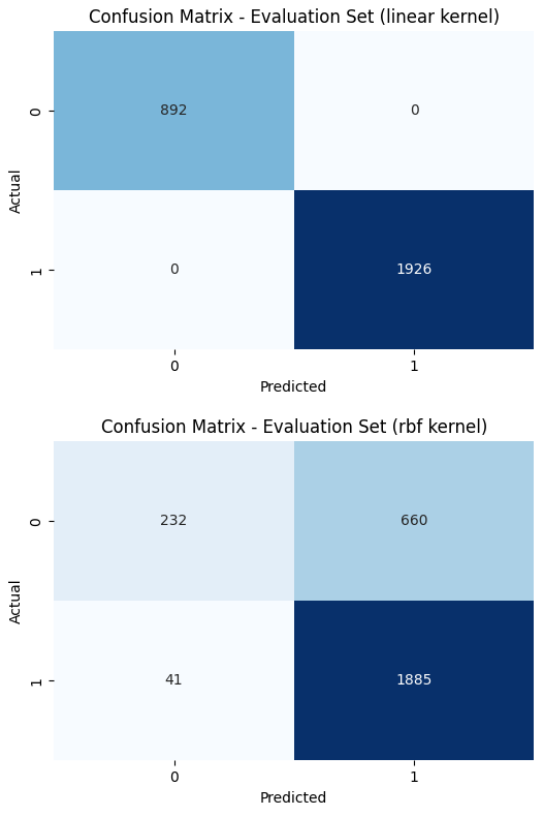
The next step after data processing is to build models using machine learning algorithms. In this study, the algorithms employed include Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes. To build and train the models, we used Google Colab along with the Keras and TensorFlow deep learning frameworks. Utilizing Keras and TensorFlow facilitates efficient model implementation and takes advantage of Google Colab's computational power for faster model training.

The accuracy results obtained from the classification on the test set are then saved as the classification model. The classification was conducted using the Support Vector Machine (SVM) algorithm with various kernels. The classification results using SVM with different kernels can be seen in Table 2.

**Table 2.** Comparison Result of SVM Kernel Evaluation

| Kernel | AUC | Precision | Recall | F1 |
|---|---|---|---|---|
| Linear | 100 | 100 | 100 | 100 |
| Polinomial | 72.96 | 78.62 | 72.96 | 65.76 |
| RBF | 74.84 | 77.97 | 74.84 | 69.69 |
| Sigmoid | 88.11 | 88.80 | 88.11 | 87.50 |

Based on Table II, it is evident that the SVM model built using the linear kernel achieved the best performance, with accuracy, precision, recall, and F1-score all reaching 100%. The linear kernel outperforms other SVM kernels. The confusion matrices for each kernel are shown in Figure 4.



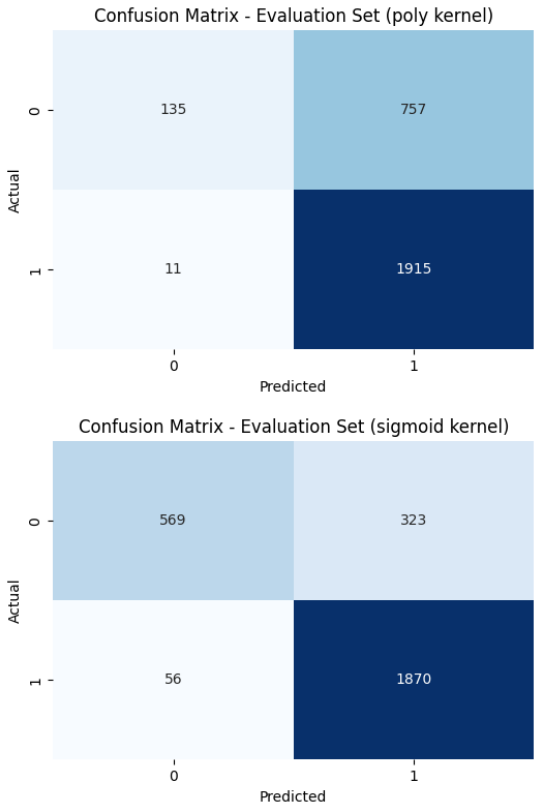



**Figure 4.** Confusion Matrix Evaluation of SVM Model

Subsequently, we developed a model using the K-Nearest Neighbors (KNN) algorithm, with the chosen value of k set to 5. The performance results for the KNN model showed an accuracy of 96.27%, precision of 96.34%, recall of 96.27%, and an F1-score of 96.23%. These results indicate that the KNN model also provides very good performance in classifying data. For a detailed view of how the KNN model performs in predicting the various classes, the confusion matrix for the KNN model is shown in Figure 5.

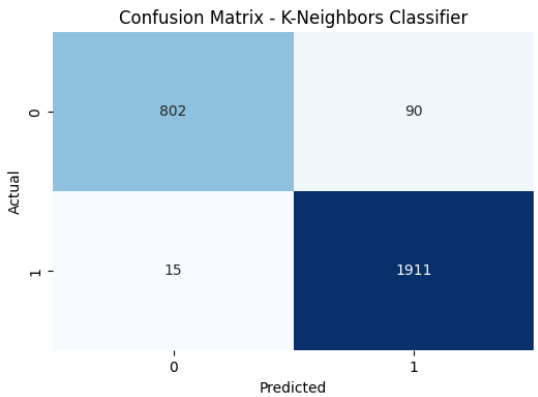



**Figure 5.** Confusion Matrix Evaluation of KNN Model

The final model developed in this study employs the Naïve Bayes algorithm, which is based on probabilistic principles, allowing the model to make predictions based on the probability distribution of the classes. The performance results of the Naïve Bayes model show an accuracy of 74.59%, precision of 74.26%, recall of 74.59%, and an F1-score of 74.40%. These figures indicate that while the Naïve Bayes model performs well, its performance is not as high as that of the SVM and KNN models tested. For a more detailed analysis of the predictions and performance of the Naïve Bayes model, the confusion matrix is provided in Figure 6.

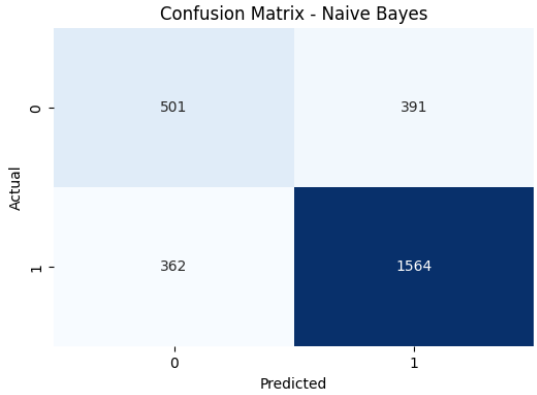



**Figure 6.** Confusion Matrix Evaluation of Naive Bayes Model

A comprehensive comparison of the performance results for each model can be seen in Table 3, which summarizes the outcomes of the three tested algorithms.

**Table 3.** Comparison Result of Model Evaluation

| Model | AUC | Precision | Recall | F1 |
|---|---|---|---|---|
| SVM | 100 | 100 | 100 | 100 |
| KNN | 96.27 | 96.34 | 96.27 | 96.23 |
| Naïve Bayes | 74.59 | 74.26 | 74.59 | 74.40 |

The data presented in Table III indicates that the model using the Support Vector Machine (SVM) with a linear kernel achieved the best performance, with accuracy, precision, recall, and F1-score all at 100%. In contrast, the model using the Naïve Bayes algorithm demonstrated lower performance, with an accuracy of 74.59%, precision of 74.26%, recall of 74.59%, and an F1-score of 74.40%. These results suggest that the SVM with a linear kernel provides more optimal results compared to Naïve Bayes for monkeypox disease classification. This study also suggests avenues for further research, such as comparing performance with other algorithms or using alternative datasets, such as skin images from patients, for monkeypox classification.

## CONCLUSION

This study aims to classify monkeypox disease by comparing several machine learning algorithms, including Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Naïve Bayes, and to enhance the performance from previous research, which achieved only 65% accuracy. The results indicate that the SVM model with a linear kernel achieved the highest accuracy of 100%, significantly outperforming the KNN and Naïve Bayes models, which achieved accuracies of 96.27% and 74.59%, respectively. These findings highlight that the SVM with a linear kernel is highly effective for monkeypox classification, surpassing the results of previous studies that utilized SVM with an RBF kernel.

## DISEMINATION

This article has been disseminated at the National Seminar on Information and Communication Technology (SEMNASTIK) APTIKOM Year 2024 held by Universitas Methodist Indonesia on October 24-26, 2024.

## REFERENCES

Adapala, J. S. S., Gontla, K. V. S., Koka, V., Modugula, S. L., Mothukuri, R., & Bulla, S. (2023). Breast Cancer Classification using SVM and KNN. *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, 1617–1621. https://doi.org/10.1109/ICEARS56392.2023.10085546

Alakunle, E., Moens, U., Nchinda, G., & Okeke, M. I. (2020). Monkeypox virus in nigeria: Infection biology, epidemiology, and evolution. In *Viruses* (Vol. 12, Issue 11). MDPI AG. https://doi.org/10.3390/v12111257

Altindis, M., Puca, E., & Shapo, L. (2022). Diagnosis of monkeypox virus – An overview. In *Travel Medicine and Infectious Disease* (Vol. 50). Elsevier Inc. https://doi.org/10.1016/j.tmaid.2022.102459

Anugrah, W., Haerani, E., & Oktavia, L. (2024). Klasifikasi Penyakit Cacar Monyet Menggunakan Metode Support Vector Machine. *Journal of Computer System and Informatics*, *5*(3), 558–566. https://doi.org/10.47065/josyc.v5i3.5149

Artanti, V. (2024). Classification of Cardiovascular Diseases Using the K-Nearest Neighbors (KNN) Algorithm. *IEESE International Journal of Science and Technology (IJSTE)*, *13*(2), 12–27.

Bagas Pattimura, Y., Paitin Kanoena, M., & Dwi Hartanto, A. (2023). Implementasi Algoritma Naïve Bayes dalam Identifikasi Citra Jenis Penyakit Cacar Dengan Image Processing. *Information Technology Journal*, *5*(1).

Belete, D. M., & Huchaiah, M. D. (2021). Performance Evaluation of Classification Models for HIV/AIDS Dataset. In *Lecture Notes on Data Engineering and Communications Technologies* (Vol. 70, pp. 109–125). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-981-16-2934-1_7

Cichosz, P. (2015). Kernel methods. In *Data Mining Algorithms* (pp. 454–497). Wiley. https://doi.org/10.1002/9781118950951.ch16

Elshewey, A. M., Shams, M. Y., El-Rashidy, N., Elhady, A. M., Shohieb, S. M., & Tarek, Z. (2023). Bayesian Optimization with Support Vector Machine Model for Parkinson Disease Classification. *Sensors*, *23*(4). https://doi.org/10.3390/s23042085

Espejel, A. H., & Cantu-Ortiz, F. J. (2021). Data Mining Techniques to Build A Recommender System. *2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*, 217–221. https://doi.org/10.1109/ISCSIC54682.2021.00047

Guo, W., Wang, G., Wang, C., & Wang, Y. (2023). Distribution network topology identification based on gradient boosting decision tree and attribute weighted naive Bayes. *Energy Reports*, *9*, 727–736. https://doi.org/10.1016/j.egyr.2023.04.256

Haque, Md. E., Ahmed, Md. R., Nila, R. S., & Islam, S. (2022). *Classification of Human Monkeypox Disease Using Deep Learning Models and Attention Mechanisms*. https://doi.org/https://doi.org/10.48550/arXiv.2211.15459

Huang, Y., Mu, L., & Wang, W. (2022). Monkeypox: epidemiology, pathogenesis, treatment and prevention. In *Signal Transduction and Targeted Therapy* (Vol. 7, Issue 1). Springer Nature. https://doi.org/10.1038/s41392-022-01215-4

Kohsasih, K. L., Hayadi, B. H., Robet, Juliandy, C., Pribadi, O., & Andi. (2022). Sentiment Analysis for Financial News Using RNN-LSTM Network. *2022 4th International Conference on Cybernetics and Intelligent System, ICORIS 2022*. https://doi.org/10.1109/ICORIS56080.2022.10031595

Kohsasih, K. L., & Situmorang, Z. (2022). Analisis Perbandingan Algoritma C4.5 dan Naïve Bayes Dalam Memprediksi Penyakit Cerebrovascular. *Jurnal Informatika*, *9*(1), 13–17. https://doi.org/10.31294/inf.v9i1.11931

Leni, D., Chamim, M., Sumiati, R., & Rosa, Y. (2023). Modeling Mechanical Component Classification Using Support Vector Machine with A Radial Basis Function Kernel. *JURNAL Teknik Mesin*, *16*(2), 165–174. http://ejournal2.pnp.ac.id/index.php/jtm

Lonang, S., Yudhana, A., & Biddinika, M. K. (2023). Performance Analysis for Classification of Malnourished Toddlers Using K-Nearest Neighbor. *Scientific Journal of Informatics*, *10*(3), 313. https://doi.org/10.15294/sji.v10i3.45196

Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine Learning Technology-Based Heart Disease Detection Models. *Journal of Healthcare Engineering*, *2022*. https://doi.org/10.1155/2022/7351061

Rabbani, S., Safitri, D., Rahmadhani, N., Sani, A. A. F., & Anam, M. K. (2023). Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(2), 153–160. https://doi.org/10.57152/malcom.v3i2.897

Rochim, A. F., Kusumastuti, R., & Windasari, I. P. (2021). Comparison of Feature Selection for Naive Bayes Classification Method in A Case Study of The Corona virus Lockdown. *2021 International Conference on Data Science and Its Applications (ICoDSA)*, 215–220. https://doi.org/10.1109/ICoDSA53588.2021.9617471

Safitri, L. R., Chamidah, N., Saifudin, T., & Alpandi, G. T. (2019). Comparison of Kernel Support Vector Machine In Stroke Risk Classification (Case Study: IFLS data). *The 1st International Conference on Global Development*. https://doi.org/http://dx.doi.org/10.12962/j23546026.y2019i6.6394

Shaban, W. M., Rabie, A. H., Saleh, A. I., & Abo-Elsoud, M. A. (2021). Accurate detection of COVID-19 patients based on distance biased Naïve Bayes (DBNB) classification strategy. *Pattern Recognition*, *119*, 108110. https://doi.org/10.1016/j.patcog.2021.108110

Syafika, V. A. N., & Karisma, R. D. L. N. (2023). On Implementation of Support Vector Machine (SVM) in Determining the Classification of Stunting-Specific Intervention Index in Indonesia. *Seminar Nasional Official Statistics*, *2023*(1), 267–276. https://doi.org/10.34123/semnasoffstat.v2023i1.1595

Wang, L., Shang, J., Weng, S., Aliyari, S. R., Ji, C., Cheng, G., & Wu, A. (2023). Genomic annotation and molecular evolution of monkeypox virus outbreak in 2022. *Journal of Medical Virology*, *95*(1). https://doi.org/10.1002/jmv.28036

World Health Organization. (2024). *Multi-country outbreak of mpox*. https://www.who.int/publications/m/item/multi-country-outbreak-of-mpox--external-situation-report-35--12-august-2024