The Role of Dimensionality Reduction and Feature Extraction in Improving Malaria Detection Models

Adithya Kusuma Whardana *adithya@tau.ac.id*
Faculty of Engineering and Technology/Informatics Engineering
Tanri Abeng University
Jakarta, 12250, Indonesia

Elfira Yolanda Reza *elfira.yolanda@student.tau.ac.id*
Faculty of Engineering and Technology/Informatics Engineering
Tanri Abeng University
Jakarta, 12250, Indonesia

Hari Suharto *hari.suharto@student.tau.ac.id*
Faculty of Engineering and Technology/Informatics Engineering
Tanri Abeng University
Jakarta, 12250, Indonesia

Azriel Putra Pradiva *azriel.putra@student.tau.ac.id*
Faculty of Engineering and Technology/Informatics Engineering
Tanri Abeng University
Jakarta, 12250, Indonesia

Acep Rifa Al Aziz *aceprifa@student.tau.ac.id*
Faculty of Engineering and Technology/Informatics Engineering
Tanri Abeng University
Jakarta, 12250, Indonesia

Abstract

Malaria is an infectious disease caused by Plasmodium parasites, with the P. falciparum species being the primary cause of mortality globally. Conventional microscopy-based approaches for malaria diagnosis possess considerable drawbacks, such as prolonged analysis durations and a requisite high degree of expertise. This study proposes an artificial intelligence-based methodology that integrates feature extraction via Histogram of Oriented Gradients (HOG) and dimensionality reduction through Principal Component Analysis (PCA) to enhance malaria detection efficacy utilizing a Support Vector Machine (SVM) model. This research additionally contrasts the efficacy of this approach with that of the Convolutional Neural Network (CNN) method. This study utilizes two dataset sizes, comprising 200 and 2000 photographs, with 80% allocated for training and 20% for testing. The experimental findings indicate that the fundamental SVM model attains 67.5% accuracy on short datasets, which increases to 90% with HOG, but decreases to 80% with PCA. In extensive datasets, the fundamental SVM model attained an accuracy of 70.5%, which increased to 87% using HOG, but subsequently decreased to 81.25% following PCA. Compared to the CNN method, which achieved 97% accuracy, it exhibited superior generalization capability on the test data. This work illustrates that the integration of HOG and PCA enhances malaria detection efficacy, albeit with certain trade-offs. This work examines whether hybrid classical methods, such as HOG and PCA in conjunction with SVM, may provide efficient and accurate malaria detection, particularly in resource-limited settings, emphasizing its practical applicability in real-world scenarios.

Keywords: Malaria, Histogram of Oriented Gradients, Principal Component Analysis, Support Vector Machine.

1. INTRODUCTION

Malaria is a lethal infectious disease caused by the Plasmodium parasite, which is transmitted when the female Anopheles mosquito bites a host. One of the most deadly diseases in the world, malaria is caused by the most deadly species of Plasmodium falciparum and accounts for hundreds of thousands of deaths every year (Saba et al., 2022). Over 249 million instances were registered in 85 countries by the World Health Organization in 2022, highlighting the ongoing threat of the disease (World Health Organization, 2023).

Effective treatment of malaria requires a prompt and precise diagnosis. Microscopical examination of Giemsa-stained blood smears is considered the gold standard; nonetheless, it is a labor-intensive process that requires experts (Memon et al., 2019). There are new possibilities for automated malaria detection thanks to the rise of artificial intelligence, especially in picture categorization. In this field, deep learning techniques, especially CNNs, have shown impressive accuracy(Alfayat, M. P et al., 2024). Yet, in underdeveloped areas or remote clinics, the computer resources and huge datasets needed to run such models are frequently inaccessible.

Support Vector Machines (SVMs) and other traditional ML techniques, along with feature extraction tools like Histogram of Oriented Gradients (HOGs) and dimensionality reduction methods like Principal Component Analysis (PCA), provide an alternative that is both lightweight and easy to understand. While some research has demonstrated the efficacy of these methods on their own, very little is known about how well they work when tailored to meet the unique clinical needs of individual patients.

This work is unique since it directly compares its performance to CNN-based methods with limited data and uses HOG and PCA parameters to identify malaria in confined situations. Scalability in environments with restricted access to computational infrastructure is addressed by this work, which shows that a standard machine learning pipeline may attain competitive accuracy with fewer resource requirements.

Feature-based approaches such as Histogram of Oriented Gradients (HOG) have shown competitive performance in facial expression recognition, particularly when paired with SVM classifiers. Their ability to detect subtle facial geometry changes makes them suitable for real-time emotion-aware systems, as demonstrated by their success on datasets like JAFFE and CK+. Additionally, HOG remains relevant in human detection tasks due to its strong performance in structured environments, with reliable accuracy in scenarios involving crowd occlusion, making it particularly suitable for surveillance applications when combined with machine learning classifiers like SVM (Rajaa et al., 2021)(Hossain et al., 2019).To further improve classification effectiveness and save computational time, Principal Component Analysis (PCA) is employed to decrease the dimensionality of the retrieved image features. This process preserves the relevant information while increasing classification efficiency (Hasan & Abdulazeez, 2021). Particularly helpful in medical image processing, which frequently encounters huge datasets, principal component analysis (PCA) is renowned for its capacity to reduce high-dimensional data while preserving a substantial amount of information (Sarowar et al., 2019).

This paper proposes a hybrid CNN-SVM approach for semantic concept detection in videos, achieving a Mean Average Precision (MAP) of 0.58 on the TRECVID 2007 dataset, surpassing methods like NTT-MD-DUT and Cascading CNN. It integrates global features from keyframes with deep features from the fc7 layer of AlexNet, utilizing linear score fusion. While effective for multimedia indexing and video retrieval, the approach lacks comparison with modern deep learning methods, details on SVM parameter optimization, and statistical analysis. The small dataset scale limits generalization, and minor grammatical errors reduce professionalism, requiring further validation for robustness (Patil & Sawarkar, 2019). Similarly, CNNs have proven to be incredibly accurate in medical image categorization, particularly in detecting malaria, demonstrating their versatility across different domains (Vijayalakshmi A & Rajesh Kanna B, 2020). Support vector machines (SVM) are useful for high-dimensional data because they maximize the margin between classes by locating the hyperplane with the best fit. In an effort to

enhance the precision of malaria parasite identification, this work employs SVM as the principal classifier for malaria detection.

Support Vector Machine (SVM) is consistently recognized for its high classification accuracy and robustness, particularly when combined with effective feature extraction and dimensionality reduction techniques. Principal Component Analysis (PCA) enhances computational efficiency by reducing feature dimensionality while preserving essential data variance. Additionally, Histogram of Oriented Gradients (HOG) efficiently captures edge and shape information in images, making it highly effective for object recognition and defect detection. The integration of HOG, PCA, and SVM offers a reliable, fast, and accurate solution for various image classification tasks, outperforming conventional multi-class strategies in specific applications (Uddin et al., 2019; Mao et al., 2019; Sharma & Singh, 2022).

An improved methodology for detecting malaria by the optimal integration of HOG, PCA, and SVM is the principal contribution of this research. A practically oriented framework tailored to malaria diagnostics, integrating well-established techniques (HOG, PCA, and SVM) into an optimized pipeline adapted to limited-resource environments applicable to real-world clinical situations is proposed in this paper, which integrates feature extraction, dimensionality reduction, and machine learning classification. The ultimate goal of this study is to provide the groundwork for malaria detection technologies in the future, which should improve diagnostic precision and cut down on medical staff mistakes.

2. LITERATURE REVIEW

With millions of cases and high fatality rates, malaria continues to be a major worldwide health concern, especially in sub-Saharan Africa. Especially in settings with low resources, traditional diagnostic methods like microscopic analysis of blood samples have a number of drawbacks, such as the need for trained staff, lengthy processing periods, and the possibility of human error (Saba et al., 2022). So, there's been a lot of buzz about how AI and ML may improve malaria diagnosis in terms of accuracy, efficiency, and accessibility.

The study highlights the advantages of Support Vector Machine (SVM), particularly its ability to achieve high classification accuracy and efficiency when combined with effective feature selection and dimensionality reduction techniques. By integrating PCA, KPCA, and an improved Reliefalgorithm, the proposed SVM-based approach demonstrated superior performance on complex datasets, confirming SVM's robustness in handling high-dimensional data and its effectiveness in extracting discriminative features for accurate classification (Ding, 2019).Feature extraction using SVM in conjunction with Histogram of Oriented Gradients (HOG) has been the subject of multiple studies. HOG excels in capturing object features like shape and texture, making it a great tool for categorizing blood cells infected with malaria. To illustrate the efficacy of HOG characteristics in reaching high accuracy in classification, showed that they could detect malaria parasites in thin blood smear pictures (Rajaraman et al., 2019). To improve classification effectiveness, a hybrid of HOG and SVM has been employed to extract pertinent picture characteristics that characterize the look and feel of RBCs (BEKTAŞ, 2019).

To further reduce data dimensionality while keeping important information, Principal Component Analysis (PCA) is commonly employed in medical image processing. Principal Component Analysis (PCA) is recognized as an effective technique in medical image classification for enhancing model performance and maintaining accuracy, particularly through its ability to reduce computational load by minimizing data dimensionality (Velliangiri et al., 2019). Combining principal component analysis (PCA), hidden Markov model (HOG), and support vector machine (SVM) has improved malaria diagnosis accuracy and simplified data analysis (Sarowar et al., 2019).

Because of their capacity to automatically learn and extract hierarchical characteristics from images, Convolutional Neural Networks (CNN) have becoming more and more used in the area of malaria detection alongside traditional machine learning methods. Convolutional Neural

Networks (CNN) offer superior performance in image recognition tasks due to their ability to automatically extract hierarchical features, handle spatial dependencies, and achieve higher accuracy compared to conventional methods (Fan, 2021). The application of Convolutional Neural Networks (CNN) combined with transfer learning has been shown to achieve high classification accuracy even when using relatively small datasets, particularly in tasks such as plant disease detection (Alfayat, M. P et al., 2024).

Convolutional Neural Networks (CNN) demonstrate superior performance in classification tasks however, their practicality is often limited by high computational demands, extensive training time, and the need for significant resources. To address these limitations, combining classical methods such as Support Vector Machines (SVM) with feature extraction techniques like Histogram of Oriented Gradients (HOG) and Principal Component Analysis (PCA) offers a more computationally efficient alternative. While CNN generally achieves higher overall accuracy, studies have shown that SVM, when supported by well-selected features, can produce competitive results especially when working with smaller datasets (BEKTAŞ, 2019).Previous studies have shown that the combination of Principal Component Analysis (PCA) and Histogram of Oriented Gradients (HOG) is effective for object classification in images. PCA efficiently reduces dimensionality, while HOG provides robust feature representations, making this approach well-suited for large-scale image data. In the context of malaria diagnosis, this combination enables high classification performance with reduced computational overhead(Sarowar et al., 2019).

While convolutional neural networks (CNN) and other deep learning models have been effective in malaria diagnosis, there are still obstacles to overcome, such as dealing with massive datasets and the associated high computing demands (Asim, 2023). For smaller datasets or systems with restricted resources, the computational efficiency of combining SVM, HOG, and PCA is significantly higher.

This study is distinguished from previous work by explicitly focusing on this hybrid classical machine learning approach (HOG + PCA + SVM), which has rarely been directly compared to CNN within the same experimental framework. While prior research has demonstrated the success of deep learning models, few studies have assessed whether traditional techniques can still achieve competitive accuracy in constrained environments. This work fills that gap by providing a comparative evaluation and highlighting how traditional techniques can be optimized to match deep learning performance using limited data and resources. The findings contribute a practical solution suitable for real-world deployment in under-resourced settings, such as rural clinics or mobile health units. Unlike prior studies, this work applies the hybrid approach within a unified experimental setup using the same dataset and evaluation metrics, thereby offering a direct and practical performance comparison. Additionally, the study fine-tunes PCA component selection to preserve 99% data variance and utilizes HOG feature extraction with fixed parameter settings, showing that the classical pipeline can still perform effectively without relying on complex parameter optimization to enhance feature representation, which has not been concurrently emphasized in previous research.

3. METHODOLOGY

Based on the research that has been carried out, there are several stages of the process carried out in the detection of malaria cells, namely: collection, pre-processing of datasets, training & testing of data.

This study uses a deductive research strategy, starting with existing theory and previous findings in medical image analysis and machine learning. The goal is to compare the efficiency of combining Histogram of Oriented Gradients (HOG), Principal Component Analysis (PCA), and Support Vector Machine (SVM) as a lightweight classification pipeline with deep learning-based methods. Data was collected using publicly available malaria datasets from Kaggle. To model the scenario in a low data environment, 200 images of small blood smears, including infected and uninfected cells. Data analysis began with normal preprocessing methods, followed by feature

extraction, dimensionality reduction, and classification. To validate the comparative performance of the models, measures such as accuracy, precision, recall, and F1-score. Given the limited sample size used in this study (200 images), this study used an 80-20 trial split to simulate real-world scenarios. Research using this approach may result in optimistic performance estimates. Cross-validation was not applied in this experiment, but it is considered an important direction for future research to improve statistical reliability.

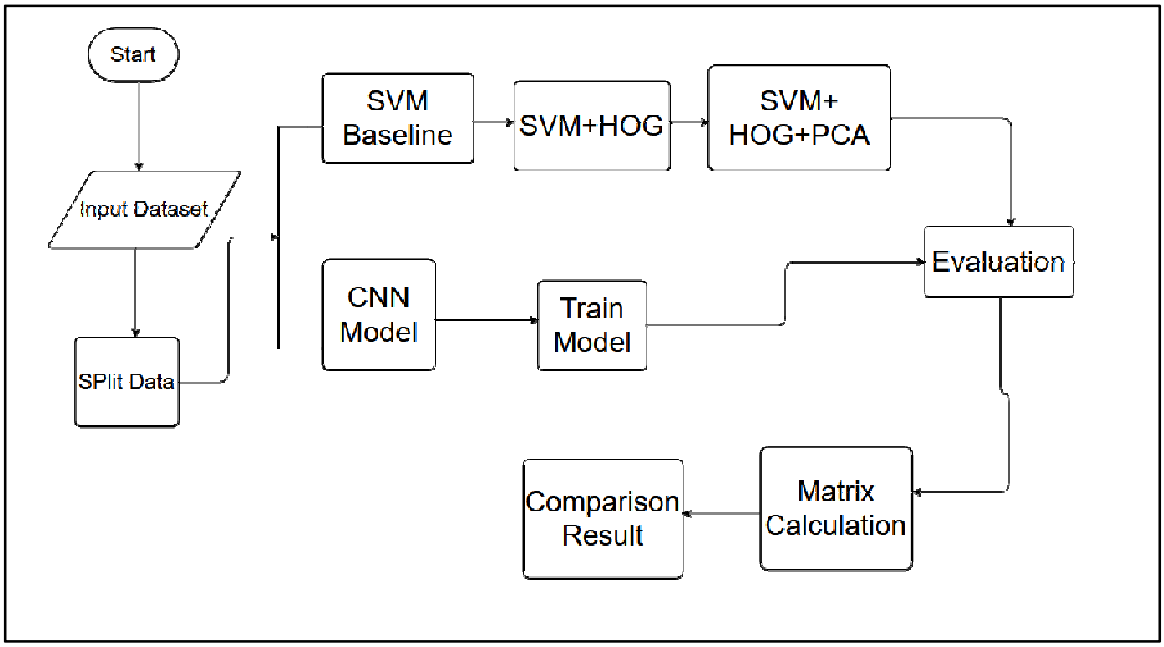

FIGURE 1: Research Flow Diagram.

Support Vector Machine (SVM), Histogram of Oriented Gradients (HOG), Principal Component Analysis (PCA), and Convolutional Neural Networks (CNN) were some of the machine learning models utilized in this study for malaria diagnosis. The research flow diagram gives an overview of the technique as a whole. Data collection, pre-processing, feature extraction, model training and testing, and method comparison are the main processes in the systematic approach shown in the flowchart.

The gathering of a dataset including pictures of parasitized and non-infected red blood cells is the initial stage of the study procedure. This study's dataset consisted of 27,558 data points culled from a thorough set of blood cell pictures made available on the Kaggle platform. An experimental subset of 200 photographs was the primary emphasis of this research. Eighty percent of the dataset was utilized for training the model, while the remaining twenty percent was utilized for testing the developed model. This separation guarantees a balanced strategy, avoids overfitting, and gives an unbiased assessment of the model's efficacy (Geevaretnam et al., 2022).

After data collection, the next crucial step is pre-processing the data to prepare it for feature extraction and model training. This includes: To guarantee that the model receives homogeneous input dimensions, resize the photos to 128x128 pixels. Converting to grayscale to remove unnecessary color information and concentrate on intensity gradients, which are more useful for recognizing malaria-infected cells (Sepahvand, 2021). Pixel values are normalized to a range of 0 to 1, ensuring that the data is uniform for efficient model training. This pre-processing improves the model's convergence during training by guaranteeing that all features are of equal scale. These pre-processing stages are critical for boosting the model's ability to learn key patterns while reducing noise and irrelevant features in the input, resulting in improved classification performance.

Feature extraction is an important stage in the malaria detection process because it converts raw images into more informative representations. This study includes two basic feature extraction techniques: HOG and PCA. HOG (Histogram of Oriented Gradients) is utilized to record the shape and textural characteristics of malaria-infected red blood cells. Histogram of Oriented Gradients (HOG) is effective for capturing local edge and shape information, making it well-suited for image classification tasks such as fashion item recognition. By encoding gradient orientation patterns, HOG provides a robust representation of object structure, contributing to reliable classification performance when combined with machine learning classifiers like SVM(Greeshma et al., 2020). Principal Component Analysis (PCA) is widely valued for its ability to reduce high-dimensional data into a lower-dimensional space while preserving most of the original data variance. This improves computational efficiency, mitigates overfitting risks, and enhances classification performance, particularly in complex datasets such as microarray gene expression and high-dimensional image features. By transforming correlated features into uncorrelated principal components, PCA effectively simplifies data structure, facilitating faster and more accurate processing in machine learning models(Ma'Ruf et al., 2019).PCA is very useful when working with high-dimensional data since it allows the model to focus on the most relevant components of the data, improving categorization by eliminating minor deviations.

Following feature extraction, the models are trained and tested using the extracted features. Baseline SVM, a simple Support Vector Machine, is first applied to the raw dataset with no feature extraction or reduction procedures. The accuracy metrics True Positives (TP), False Negatives (FN), and True Negatives (TN) are used to assess the model's capacity to distinguish between infected and non-infected red blood cells. The combined use of Support Vector Machine (SVM), Histogram of Oriented Gradients (HOG), and Principal Component Analysis (PCA) offers significant advantages in image-based classification tasks. HOG effectively captures contour and shape features, while PCA reduces feature dimensionality and eliminates redundancy, improving computational efficiency. When integrated with SVM, this combination enhances classification accuracy and speed, demonstrating robust performance in complex environments such as coal and gangue identification, where precise feature extraction and efficient dimensionality reduction are critical (Cheng et al., 2023).Convolutional Neural Networks (CNN), as a deep learning technique, are employed to automatically learn hierarchical features directly from image data, thereby eliminating the need for manual feature extraction (Alfayat, M. P et al., 2024). CNN are extremely useful in picture classification problems because they can automatically extract hierarchical features from input data. In this study, the CNN model is compared to SVM-based models to assess its ability to detect malaria in microscopic images of red blood cells (Vijayalakshmi A & Rajesh Kanna B, 2020).

After training and testing the models, the next step is to compare the performance of the various approaches (Baseline SVM, SVM + HOG, SVM + HOG + PCA, and CNN). The comparison is based on numerous performance parameters. Accuracy is the percentage of correctly classified instances out of all instances. Precision is the percentage of true positive predictions (properly discovered malaria cells) among all positive predictions. Recall that the model can accurately identify all malaria-infected cells. F1-Score: A composite metric that considers both precision and recall.

The performance comparison aids in determining which method offers the optimum balance of accuracy, computing efficiency, and model complexity. CNN achieved 97% accuracy, but SVM-based models indicated a trade-off between accuracy and sensitivity.

3.1 Data Collection

The data set of this study was taken from the Kaggle platform and consisted of 27,558 data in image format. The data were classified into two main categories, namely parasites, which included images of red blood cells infected with malaria parasites, and uninfection, which included images of normal red blood cells. In this study, the dataset used was 200 data. Next, this dataset is processed by dividing it into two main parts: training data and test data. The segmentation process is carried out with a proportion of 80% of the data allocated for training,

which aims to train the Support Vector Machine (SVM) model, while the remaining 20% is used for testing to evaluate the performance of the trained model. This division is designed to ensure a balanced distribution of data between training and testing, so that the model can be properly optimized and provide representative evaluation results(Geevaretnam et al., 2022).

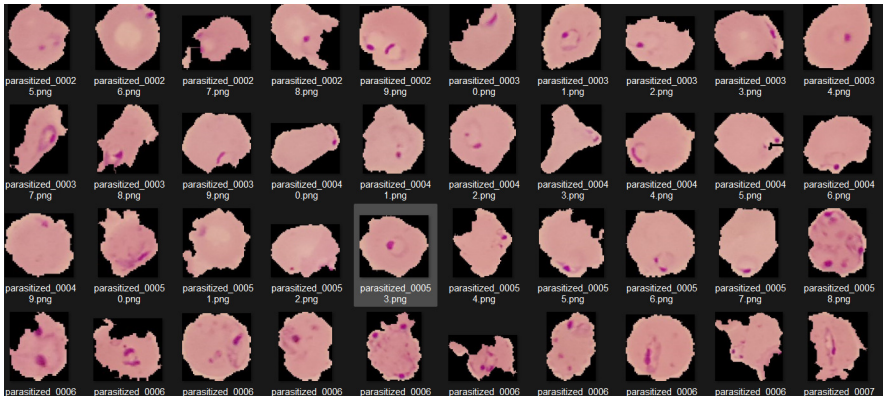

FIGURE 2: Precitized Red Blood Cells.

Figure 2 shows red blood cells that have been parasitized, possibly by parasites such as Plasmodium that cause malaria. Each box in this image shows different shapes and conditions of infected cells. The pink color of the cells and the dark details inside are characteristic of the presence of parasitic infections.

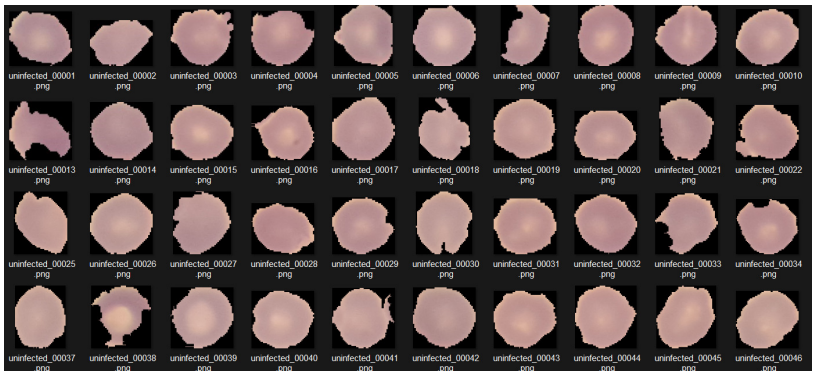

FIGURE 3: Uninfected Red Blood Cells.

Figure 3 shows uninfected red blood cells with a normal shape and a uniform pink color with no signs of parasite infection. This image provides a visual comparison of the infected cells in Figure 2.

3.2 Pre-processing

The preprocessing stage in this study includes several main steps designed to prepare data optimally according to the needs of the malaria detection model.

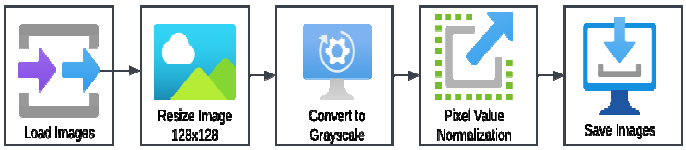

FIGURE 4: Image Preprocessing.

Figure 4 illustrates the stages of image pre-processing, starting with resizing the image to 128×128 pixels to ensure the dimensional uniformity of the model input. The image is then converted to grayscale to simplify the data by eliminating irrelevant color information. The final step is to normalize the pixel values to the range of 0–1, with 0 indicating an uninfected image and 1 indicating an infected image. After normalization, there is a data division where the value is 80% for training and 20% for testing with the aim that the model can be tested objectively, the division of training and data testing is carried out randomly.

3.3 Feature Extraction

Feature extraction is an important step in the malaria detection process, which aims to identify specific characteristics of red blood cell images infected by malaria parasites. The Histogram of Oriented Gradients (HOG) method was used to extract the features in this study. HOG is an excellent technique for gathering information about the shape and texture of objects in images, which is particularly relevant for the identification of infected red blood cells (Sepahvand, 2021). Histogram of Oriented Gradients (HOG) is highly effective for capturing local shape and edge information in images, making it suitable for object detection and classification tasks. Principal Component Analysis (PCA) complements this by reducing the high dimensionality of HOG feature vectors, preserving essential variance while eliminating redundancy, thereby improving computational efficiency and enhancing model performance in high-dimensional image analysis (Cheng et al., 2023).

3.4 Method

In this study, the study used 3 methods to conduct research, where to get the maximum accuracy value.

3.4.1 Baseline SVM

Baseline SVM (Support Vector Machine) is an initial approach without applying feature extraction or reduction techniques. The model uses the original data without modification, so its performance depends directly on the quality and dimensions of the input data.

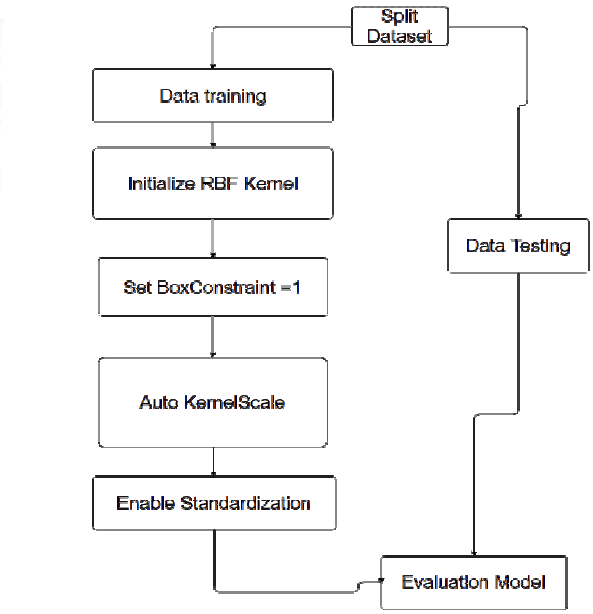

FIGURE 5: SVM Baseline Process.

Figure 5 depicts the training and assessment process of the Support Vector Machine (SVM) model using a baseline method. The method starts by separating the dataset into two parts: training data and testing data. Training data is used to train the model using a specific configuration, and testing data is used to evaluate the model's performance. In SVM training,

Kernel RBF is used to handle non-linear data by mapping data to higher dimensions, BoxConstraint=1 balances between overfitting and underfitting, KernelScale automatically helps find optimal parameters for classification, while data standardization ensures all features are at a comparable scale to improve model performance.

3.4.2 SVM with HOG Feature Extraction

In this method, features from the image are extracted using the Histogram of Oriented Gradients (HOG), which specifically captures the patterns of shape, texture, and contours of objects in the image. The HOG feature is then used as input to the SVM for classification.

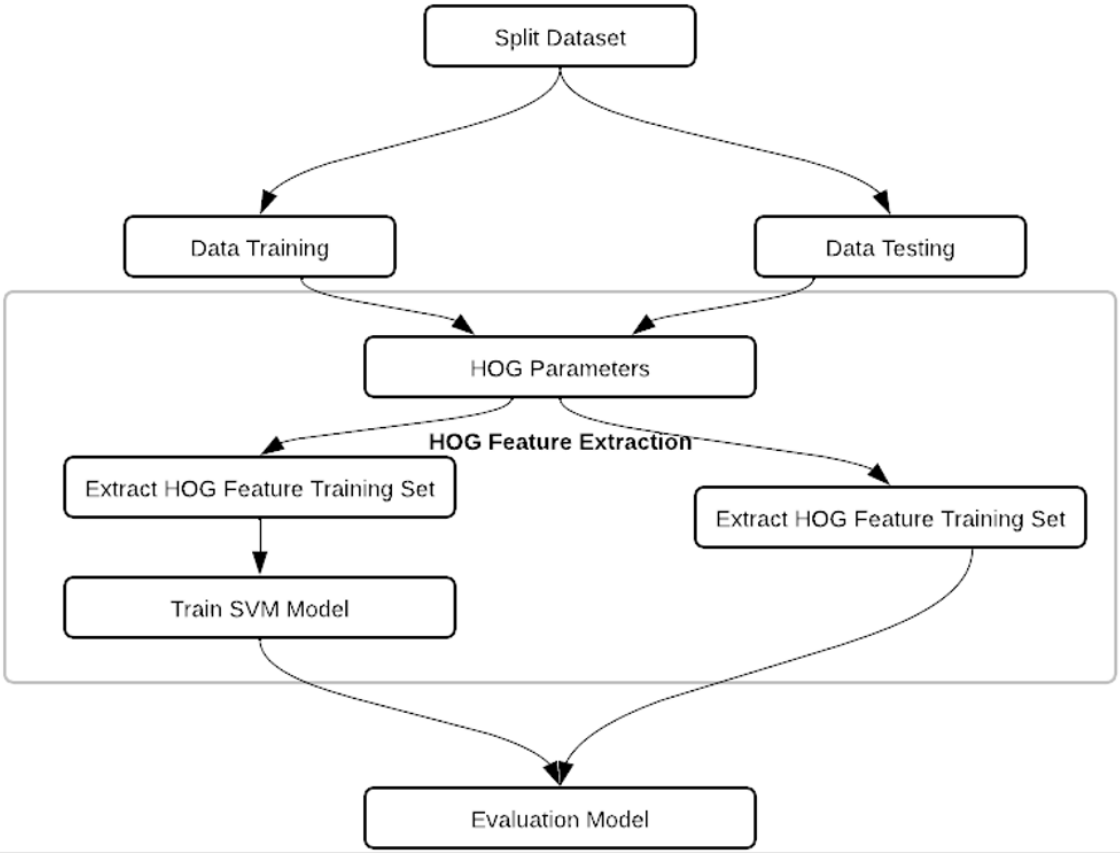

FIGURE 6: SVM Process + HOG Feature.

Figure 6 for training and evaluation process of the model based on the Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM) methods. The process begins with dividing the dataset into two parts, namely training data to train the model and testing data for model evaluation. In the HOG Feature Extraction stage, feature extraction is carried out using the same HOG parameters for both parts of the data. Training data goes through the HOG feature extraction process to produce a feature set which is then used to train the SVM model. In parallel, testing data also undergoes the HOG feature extraction process using the same parameters. After the SVM model is drilled using features from the training data, the model is evaluated using features extracted from the testing data. The evaluation results are used to assess model performance with certain metrics, such as accuracy or others. This diagram illustrates a structured workflow in integrating the HOG and SVM methods to build and test models.

The HOG parameters used consist of CellSize which specifies the base area of the histogram calculation from the gradient orientation, BlockSize which sets the normalization of the histogram to reduce the influence of lighting variations by grouping 2x2 cells, and NumBins [9] which divides

the gradient orientation into 9 angle ranges, while BoxConstraint=10 on the SVM is used to penalize more classification errors because the data is already more structured after going through the process HOG feature extraction. The value of BoxConstraint was chosen based on empirical observation during initial trials, aiming to balance the trade-off between model complexity and classification accuracy. A higher value was selected to better define the decision boundary after structured feature extraction using HOG.

3.4.3 SVM with HOG + PCA Feature Extraction

Combination of HOG features and dimension reduction with Principal Component Analysis (PCA). PCA reduces the number of dimensions of HOG data while retaining the most important information. This combination improves the efficiency and accuracy of the model.

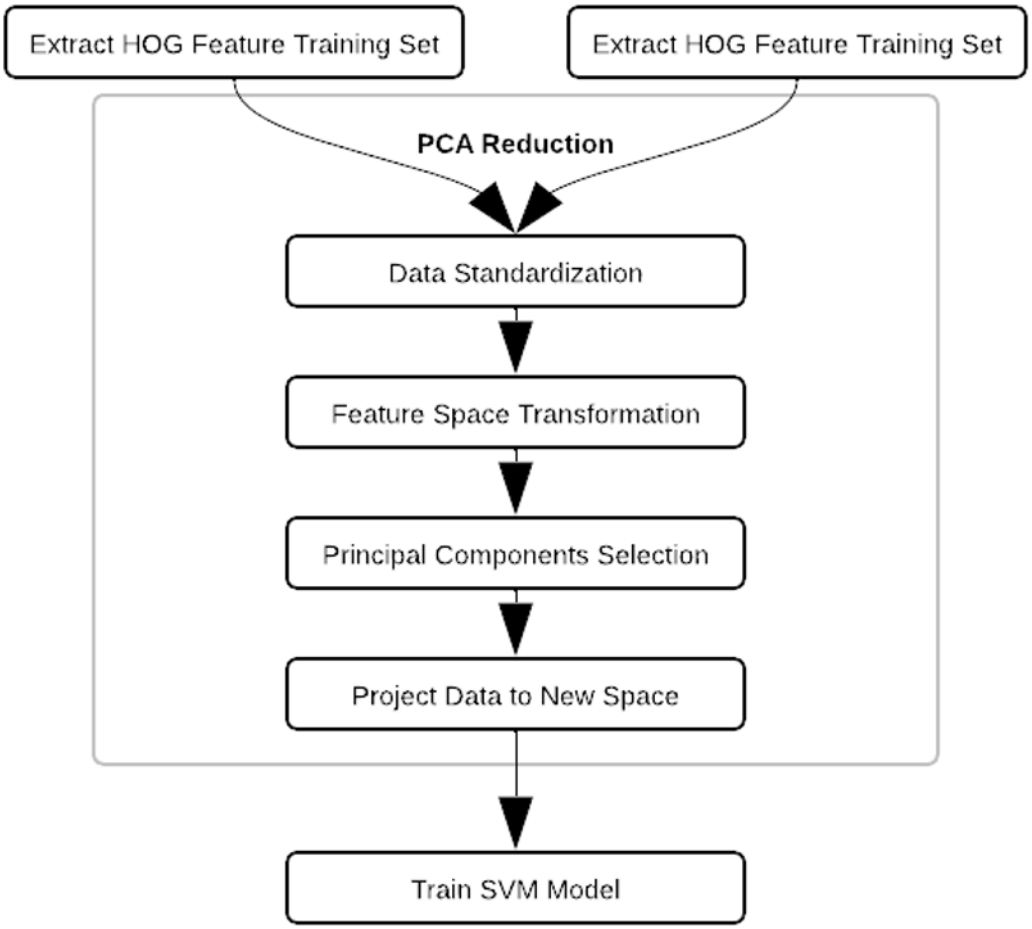

FIGURE 7: SVM+HOG+PCA Process.

Figure 7 describes the dimension reduction process using Major Component Analysis (PCA) for Histogram of Oriented Gradients (HOG) feature extraction data, before training the Support Vector Machine (SVM) model. The process begins with the extraction of HOG features from the dataset, resulting in a feature set for training and testing the data. The PCA Reduction phase begins with standardizing the data to ensure that all features are at the same scale. Next, the data is transformed into a new feature space, where key components are selected based on their contribution to the variance of the data. After that, the data is projected onto a new, lower-dimensional space, represented by the main component. These projections result in a simpler but still informative representation of the data. The reduced data is then used to train the SVM model, which aims to build a classification model with optimal performance. This diagram illustrates how PCA is implemented to reduce data complexity before the SVM model training process.

The process of dimensionality reduction with PCA includes standardizing the data using z-scores, transforming to a new space through decomposition of key components, selecting components that maintain a 99% variance, and projecting the data into a lower dimensional space. The result is a more efficient representation of the data, which is then used as input for SVM model training.

3.4.4 Convolutional Neural Network (CNN)

CNN models are more specialized in deep learning which is designed to process data that has a grid structure, especially images.The CNN model has two convolutional layers, followed by two fully linked layers. To avoid overfitting, a dropout rate of 0.3 was used, and the batch size was set to 32. To optimize the model, we employed the Adam optimizer at a learning rate of 0.0001.

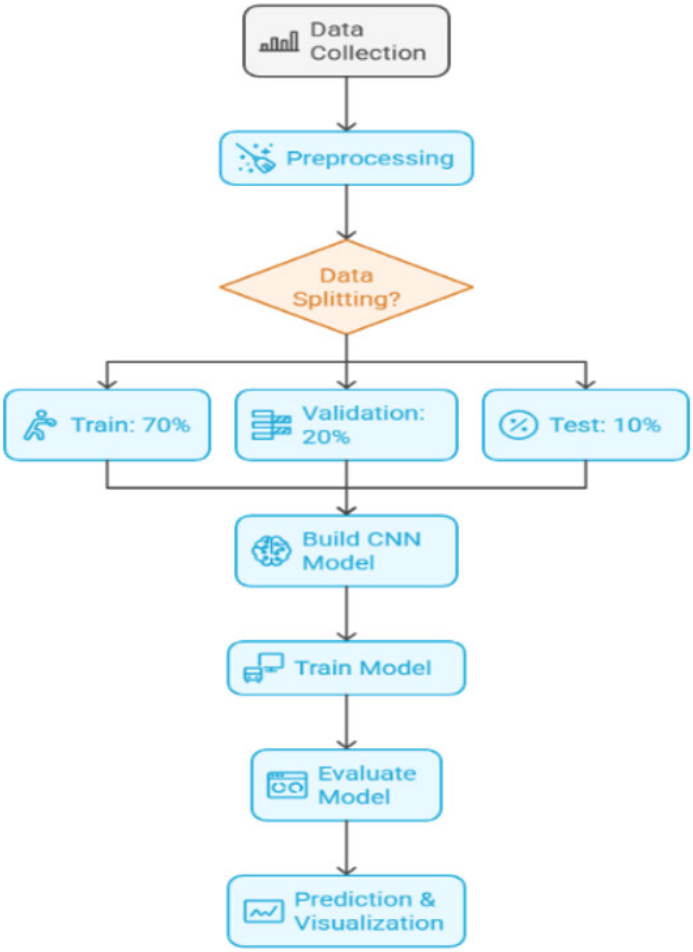

FIGURE 8: CNN Process.

3.5 Detection and Validation Accuracy

To determine the level of validity of each algorithm tested, a validation process was carried out by comparing the results of malaria parasite cell detection with the results of red blood cells. The results of this comparison were then analyzed using a confusion matrix. The detection results of each algorithm used will be compared to produce four evaluation values: True Positive (TP), Value False Negative (FN), Value True Negative (TN). The True Positive (TP) value indicates the number of malaria parasite cells that are actually present and successfully detected by the model False Negative Values (FN) indicate cases in which malaria cells are detected by the model even though they are not actually present. The True Negative (TN) value indicates the number of samples that do not have malaria parasite cells and are successfully identified as uninfected (Setiawan et al., 2020).

Model Performance Evaluation:

Accuracy: Measures the proportion of correct predictions to total data, calculated by the formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Accuracy Percentage: The accuracy percentage is calculated by multiplying the accuracy value by 100.

$$\text{Accuracy percentage} = \text{Accuracy x } 100 \quad (2)$$

Precision: Measures the accuracy of a positive prediction, which is the ratio between a correct positive prediction (TP) and all positive predictions (TP + FP).

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Recall: Measures the model's ability to detect all true positive samples, i.e. the ratio between the correct positive prediction (TP) and the total positive cases (TP + FN).

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

F1-Score: The harmonic average of Precision and Alert, used to balance the two in a single metric.

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

4. RESULTS

This study evaluated the performance of the Support Vector Machine (SVM) algorithm using a dataset of 200 images by experimenting with various feature extraction methods, including Histogram of Oriented Gradients (HOG) and Principal Component Analysis (PCA). The dataset was split into 80% training and 20% testing, ensuring balanced data distribution and objective evaluation.

The baseline SVM model, which used raw image data without any feature extraction, achieved an accuracy of 67.5%. When HOG was introduced for feature extraction, the accuracy significantly increased to 90%, highlighting its effectiveness in capturing essential shape and texture features necessary for distinguishing malaria-infected cells (Sarowar et al., 2019). However, incorporating PCA to reduce dimensionality led to a drop in accuracy to 80%. While PCA effectively reduces noise and computational cost, it may also eliminate important discriminative features, explaining the decline in performance (Hasan & Abdulazeez, 2021; Velliangiri et al., 2019).

This performance trade-off is particularly relevant for real-time diagnostic applications, where computational efficiency is essential. Despite the slight reduction in accuracy, the HOG + PCA + SVM combination remains favorable in low-resource settings. Compared to previously reported methods, the proposed approach demonstrates competitive performance even with smaller datasets, highlighting its practical applicability (Rajaraman et al., 2019; BEKTAŞ, 2019).

In conclusion, while CNN models offer higher accuracy, the hybrid classical approach combining HOG, PCA, and SVM presents a viable and computationally efficient alternative for malaria detection, particularly in resource-constrained environments.

4.1 SVM Model Performance Evaluation

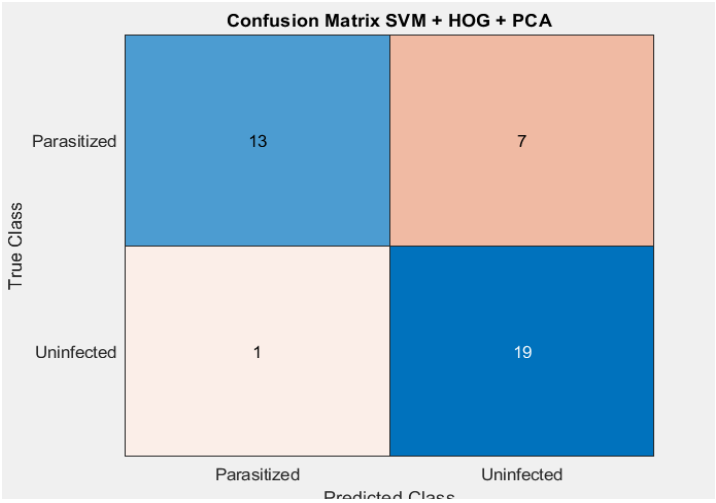

FIGURE 9: Convolution Matrix SVM + HOG PCA.

The combination of the HOG feature with Principal Component Analysis (PCA) (Figure 7) provides attractive results, with varying detection performance. For the Parasitized class, the model recorded TP 13 and FN of only 1, showing a very accurate detection although there was a slight decrease compared to the HOG feature alone. Meanwhile, for the Uninfected class, TN increased to 19, but FP also increased to 7, indicating a compromise on the ability to detect uninfected samples.

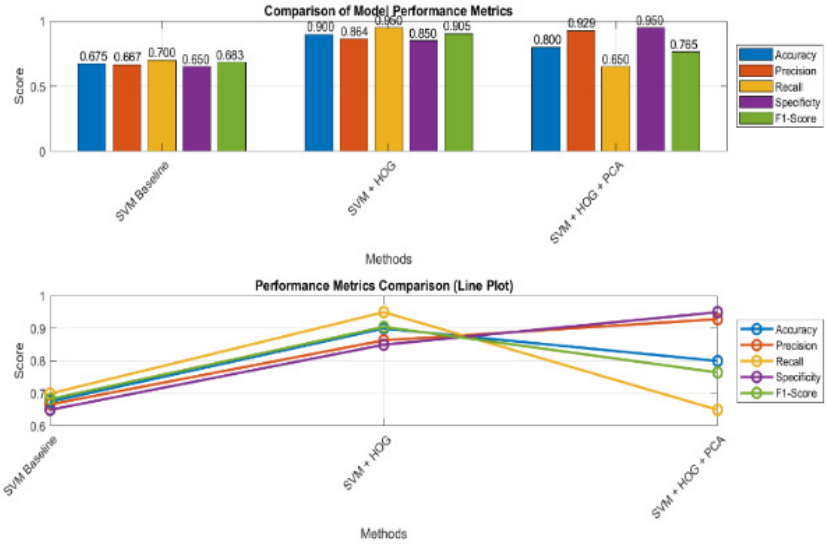

FIGURE 10: Performance Matrics comparison.

This graph (Figure 10) shows the visual change in the performance metric values of each method (SVM Baseline, SVM + HOG, and SVM + HOG + PCA) on the 200 dataset. From this graph, it can be seen that the SVM + HOG method shows a significant improvement in all metrics compared to the baseline. Meanwhile, the SVM + HOG + PCA method tends to lower some metrics, especially on the Recall and F1-Score aspects, although it provides an improvement in Precision and Specificity.

4.2 CNN Model Performance Evaluation

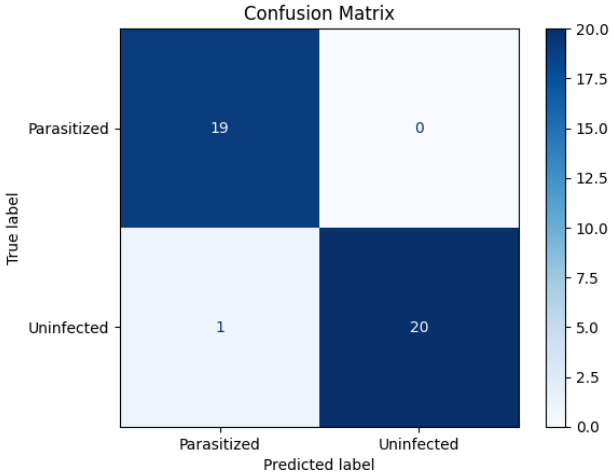

FIGURE 11: CNN Convolution Matrix.

Convolutional Neural Network (CNN) are a class of deep learning models specifically designed to process visual data. In this study, CNN is utilized as the primary method for classifying red blood cell images into infected or uninfected categories. Each image is first resized to 128×128 pixels during the preprocessing phase to ensure consistency in input dimensions and reduce computational complexity.

The CNN architecture consists of several convolutional layers with 3×3 filters, followed by ReLU activation functions to introduce non-linearity into the model. MaxPooling layers with a 2×2 kernel are inserted after certain convolutional layers to reduce the spatial dimensions and preserve the most prominent features, helping to prevent overfitting and accelerate training. Batch normalization is applied to stabilize and accelerate the training process, while dropout layers (with rates between 0.3 and 0.5) are introduced to prevent overfitting by randomly deactivating certain neurons during training.

The extracted features are then passed to fully connected (dense) layers, which process the high-level representations and perform final classification. A sigmoid activation function is used in the output layer to handle the binary classification task (parasitized or uninfected). The model is trained using the Adam optimizer with a learning rate of 0.0001, and binary cross-entropy is employed as the loss function. Training is conducted over 25 epochs with a batch size of 32, which provides optimal convergence for small-to-moderate datasets.

In addition to the custom CNN architecture, the study also implements a transfer learning approach using EfficientNetV2M as the base model. The pretrained model, originally trained on ImageNet, is imported with the include_top=False configuration to remove the original classification head. This allows the model to be fine-tuned for the malaria detection task by adding custom classification layers. Transfer learning significantly reduces training time and enhances accuracy, especially on smaller datasets. In this experiment, the model achieved an accuracy of 97%, demonstrating high generalization and minimal prediction errors.

The success of this CNN-based approach highlights its capability to automatically learn complex hierarchical features from microscopic images of red blood cells, outperforming traditional methods that rely heavily on manual feature engineering. This makes CNN a highly promising candidate for automated diagnostic systems in resource-constrained clinical environments.

Based on the confusion matrix, the model generates predictions with minimal error rates. From the total test data, there are 19 images of the "Parasitized" class that are correctly classified, with

no classification errors in this class (True Positives). Meanwhile, for the "Uninfected" class, as many as 20 images were correctly classified, with only one classification error (False Negative). This indicates that the model has a high sensitivity to both classes, although there is a slight decrease in detecting images of the "Uninfected" class.

4.3 Method Comparison

To evaluate the efficacy of various approaches for malaria detection, this study compared the performance of multiple methods, with a particular emphasis on Support Vector Machine (SVM) integrated with Histogram of Oriented Gradients (HOG) and Principal Component Analysis (PCA). Additionally, a Convolutional Neural Network (CNN) model using the EfficientNetV2M architecture and transfer learning was included to benchmark deep learning performance. The evaluation was conducted using several performance metrics, including F1 Score, Accuracy, Precision, and Recall. The results of this comparative analysis are presented in the table below.

Method	F1 Score	Precision	Recall	Accuracy (%)
Baseline SVM	0.65	0.67	0.68	67.5
SVM + HOG	0.88	0.91	0.9	90
SVM + HOG + PCA	0.75	0.78	0.76	80
CNN	0.95	0.96	0.97	97

TABLE 1: Result of Comparison Method.

When developing a Support Vector Machine (SVM)-based classification model, both computing efficiency and accuracy must be considered. Various preprocessing and dimensionality reduction procedures, such as Histogram of Oriented Gradients (HOG) and Principal Component Analysis (PCA), are used to improve the model's performance. The basic SVM model based on native data has the advantage of high computing efficiency, but it is constrained by low accuracy, especially with large datasets. For instance, a baseline SVM classifier achieved an accuracy of 67.5%, while the incorporation of Histogram of Oriented Gradients (HOG) features significantly enhanced classification performance by effectively capturing local shape and texture patterns, increasing accuracy to 90%, particularly in the recognition of malaria-infected cells (Sarowar et al., 2019).

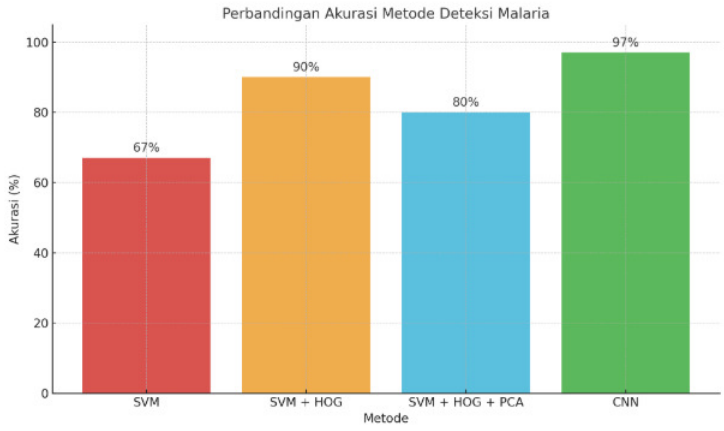

FIGURE 12: Method comparison bar.

The bar chart presents a comparative analysis of four malaria detection methods based on their classification accuracy using a dataset of 200 blood cell images.

The baseline SVM model, which does not incorporate feature extraction or dimensionality reduction, achieved an accuracy of 67%, indicating limited capability in distinguishing parasitized from healthy cells using raw image data. Incorporating Histogram of Oriented Gradients (HOG) significantly enhances performance, with the SVM + HOG model reaching 90% accuracy. This improvement is attributed to HOG's effectiveness in capturing shape and texture features relevant to malaria detection.

Adding Principal Component Analysis (PCA) to reduce feature dimensionality results in a slight decline in performance. The SVM + HOG + PCA model achieves 80% accuracy, suggesting that while PCA improves computational efficiency, it may also lead to the loss of important discriminative features. The CNN model, employing EfficientNetV2M with transfer learning, outperforms all other methods with 97% accuracy, demonstrating strong generalization capabilities in classifying malaria-infected cells from image data.

In summary, while traditional machine learning methods like SVM benefit from feature engineering, deep learning models such as CNN provide superior performance and automation, making them especially suitable for large-scale or real-time diagnostic applications.

The combination of HOG and PCA offers an optimal balance between accuracy and computational efficiency, achieving 80% accuracy while reducing feature dimensionality. PCA enhances efficiency by retaining only the most relevant features, thereby reducing the complexity of the data. However, determining the optimal number of principal components is essential to maintain performance, as dimensionality reduction techniques can vary in their impact on model accuracy and efficiency depending on how they are configured (Velliangiri et al., 2019). These findings highlight the trade-off between accuracy, computational complexity, and efficiency. While SVM combined with HOG provides high accuracy, the addition of PCA offers improved computational efficiency by reducing feature dimensionality, making the HOG + PCA + SVM approach suitable for large datasets or environments with limited computational resources. However, the CNN model, particularly when implemented with transfer learning using EfficientNetV2M, achieves the highest overall accuracy while automating feature extraction. Although CNN demands greater computational resources, it remains the most effective solution in terms of accuracy and generalization capability. Table 2 summarizes the accuracy performance of each method.

Method	**Accuracy**
SVM	67%
SVM+HOG	90%
SVM+HOG+PCA	80%
CNN	97%

TABLE 2: Result Accuracy.

The experimental results show that conventional techniques, specifically HOG + PCA + SVM, reach 90% accuracy, which is much greater than baseline SVM models. While CNN-based approaches achieve even higher accuracy (97%), they demand more data and processing resources. The proposed approach's strength is its simplicity, quickness, and applicability for low-resource clinical settings. These findings support the use of optimized classical techniques in situations where deep learning deployment is problematic due to technology restrictions or limited training data. The experimental results show that conventional techniques, specifically HOG + PCA + SVM, reach 90% accuracy, which is much greater than baseline SVM models. While CNN-based approaches achieve even higher accuracy (97%), they demand more data and processing resources. The proposed approach's strength is its simplicity, quickness, and applicability for low-resource clinical settings. These findings support the use of optimized classical techniques in situations where deep learning deployment is problematic due to technology restrictions or limited training data.

Study	Method	Accuracy	Dataset Size
Sarowar et al. (2019)	HOG + SVM	90%	1,000 images
Rajaraman et al. (2019)	Deep Neural Ensembles	94%	2,000 images
This Study	HOG + PCA + SVM	90%	200 images
This Study (CNN)	CNN	97%	200 images

TABLE 3: Comparative Results with Previous research.

This study compares the HOG + PCA + SVM algorithm for malaria diagnosis to numerous other methodologies from prior studies. According to Table 3, the Sarowar et al. (2019) technique using HOG + SVM obtained 90% accuracy with 1,000 photos, demonstrating high performance in feature extraction and classification even with bigger datasets. Rajaraman et al. (2019) used the Deep Neural Ensembles approach, which surpassed SVM in terms of accuracy (94%), had a larger dataset (2,000 images), and required more processing resources, making it less efficient than older methods. On the other hand, this study indicates that the combination of HOG + PCA + SVM with 200 images achieves 90% accuracy, providing superior computational efficiency with a smaller dataset and making it a very valuable option for resource-constrained circumstances. The EfficientNetV2 + Transfer Learning strategy obtained 97% accuracy in CNN with the same dataset, although it required more CPU resources. Overall, while CNN demonstrated the best accuracy, HOG + PCA + SVM offer considerable benefits in terms of computational efficiency and practicality in data- and hardware-constrained scenarios. Table 3 demonstrates the trade-off between accuracy and computational economy, with the CNN-based model providing the highest accuracy while needing more resources, and the HOG + PCA + SVM method providing a more efficient solution ideal for resource-constrained applications.

5. CONCLUSION

This study has successfully demonstrated that the integration of HOG and PCA with SVM significantly improves classification accuracy compared to baseline models, while maintaining computational efficiency.

This paper proposes a practical and effective malaria detection method by combining Histogram of Oriented Gradients (HOG) for feature extraction and Principal Component Analysis (PCA) for dimensionality reduction, integrated with Support Vector Machines (SVM). The study addressed the central research question: Can classical machine learning approaches, specifically the combination of HOG and PCA with SVM, provide accurate and efficient malaria detection suitable for low-resource environments?

The findings indicate that while Convolutional Neural Network (CNN) demonstrated the highest accuracy (97%), the hybrid approach of SVM with HOG achieved a strong 90% accuracy. Although the addition of PCA slightly reduced performance due to feature loss, it contributed significantly to computational efficiency, making the method viable for implementation in settings with limited resources.

The practical implications of this research are especially relevant for health facilities in rural or underdeveloped areas, where access to high-performance computing is limited. Target beneficiaries include healthcare workers in remote clinics, developers of lightweight diagnostic tools, and public health organizations aiming to deploy cost-effective AI-based malaria detection systems.

Future research may concentrate on adjusting PCA parameters to enhance the equilibrium between dimensionality reduction and feature preservation. Furthermore, evaluating the model with larger and more heterogeneous datasets would facilitate the assessment of its generalizability in practical clinical applications. Examining alternative feature extraction methods, like Gabor filters and wavelet transforms, as well as investigating lightweight transfer learning models, may further improve the system's performance and flexibility. Nonetheless, the small dataset size used in this study is recognized as a limitation. Future work will incorporate cross-

validation techniques and larger datasets to improve the generalizability and statistical robustness of the results.

6. REFERENCES

Asim, M. (2023). Classification of Microscopic Malaria Parasitized Images Using Deep Learning Feature Fusion. *Lahore Garrison University Research Journal of Computer Science and Information Technology*, *7*(02), 7. https://doi.org/10.54692/lgurjcsit.2023.0702473.

Alfayat, M. P., & Whardana, A. K. (2024). DETEKSI DINI ALZHEIMER PADA OTAK DENGAN KOMBINASI METODE. Scan: Jurnal Teknologi Informasi dan Komunikasi, 19(1), 32-41.https://doi.org/10.33005/scan.v19i1.47351

BEKTAŞ, J. (2019). Comparison of CNNs and SVM for Detection of Activation in Malaria Cell Images. *Natural and Applied Sciences Journal*, *2*(2), 38–50. https://doi.org/10.38061/idunas.632709

Cheng, G., Chen, J., Wei, Y., Chen, S., & Pan, Z. (2023). A Coal Gangue Identification Method Based on HOG Combined with LBP Features and Improved Support Vector Machine. *Symmetry*, *15*(1). https://doi.org/10.3390/sym15010202

Ding, W. (2019). SVM-Based feature selection for differential space fusion and its application to diabetic fundus image classification. *IEEE Access*, *7*, 149493–149502. https://doi.org/10.1109/ACCESS.2019.2944899

Fan, T. (2021). Image Recognition and Simulation Based on Distributed Artificial Intelligence. *Complexity*, *2021*. https://doi.org/10.1155/2021/5575883

Geevaretnam, J. L., Megat Mohd. Zainuddin, N., Kamaruddin, N., Rusli, H., Maarop, N., & Wan Hassan, W. A. (2022). Predicting the Carbon Dioxide Emissions Using Machine Learning. *International Journal of Innovative Computing*, *12*(2), 17–23. https://doi.org/10.11113/ijic.v12n2.369

Greeshma, K., Gripsy, J. V., & others. (2020). Image Classification using HOG and LBP Feature Descriptors with SVM and CNN. *Int J Eng Res Technol*, *8*(4), 1–4. www.ijert.org

Hasan, B. M. S., & Abdulazeez, A. M. (2021). A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. *Journal of Soft Computing and Data Mining*, *2*(1), 20–30. https://doi.org/10.30880/jscdm.2021.02.01.003

Hossain, B. M., Karungaru, S., & Tereda, K. (2019). *Robust Motion Detection and Tracking of Moving Objects using HOG Feature and Particle Filter*. *13*, 9–16.

Ma'Ruf, F. A., Adiwijaya, & Wisesty, U. N. (2019). Analysis of the influence of Minimum Redundancy Maximum Relevance as dimensionality reduction method on cancer classification based on microarray data using Support Vector Machine classifier. *Journal of Physics: Conference Series*, *1192*(1). https://doi.org/10.1088/1742-6596/1192/1/012011

Memon, M. H., Khanzada, T. J. S., Memon, S., & Hassan, S. R. (2019). Blood image analysis to detect malaria using filtering image edges and classification. *Telkomnika (Telecommunication Computing Electronics and Control)*, *17*(1), 194–201. https://doi.org/10.12928/TELKOMNIKA.v17i1.11586

Monteiro-Guerra, F., Rivera-Romero, O., Fernandez-Luque, L., & Caulfield, B. (2020). Personalization in Real-Time Physical Activity Coaching Using Mobile Applications: A Scoping Review. *IEEE Journal of Biomedical and Health Informatics*, *24*(6), 1738–1751. https://doi.org/10.1109/JBHI.2019.2947243

Patil, N. S., & Sawarkar, S. D. (2019). Semantic Concept Detection in Video Using Hybrid Model of CNN and SVM Classifiers. *International Journal of Image Processing*, *13*(2), 13–28.

Rajaa, S., Harrabi, R., & Ben Chaabane, S. (2021). Facial expression recognition system based on SVM and HOG techniques. *International Journal of Image Processing (IJIP)*, *15*, 14. https://www.cscjournals.org/library/manuscriptinfo.php?mc=IJIP-1215

Rajaraman, S., Jaeger, S., & Antani, S. K. (2019a). Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images. *PeerJ*, *7*. https://doi.org/10.7717/PEERJ.6977

Rajaraman, S., Jaeger, S., & Antani, S. K. (2019b). Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images. *PeerJ*, *7*(May). https://doi.org/10.7717/PEERJ.6977

Saba, N., Balwan, W. K., & Mushtaq, F. (2022). Burden of Malaria - A Journey Revisited. *Scholars Journal of Applied Medical Sciences*, *10*(6), 934–939. https://doi.org/10.36347/sjams.2022.v10i06.013

Sarowar, M. G., Razzak, M. A., & Fuad, M. A. Al. (2019). HOG feature descriptor based PCA with SVM for efficient accurate classification of objects in image. *Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing, IACC 2019*, 171–175. https://doi.org/10.1109/IACC48062.2019.8971585

Sepahvand, K. K. (2021). Structural damage detection using supervised nonlinear support vector machine. *Journal of Composites Science*, *5*(11). https://doi.org/10.3390/jcs5110303

Setiawan, A., Diyasa, I. G. S. M., Hatta, M., & Puspaningrum, E. Y. (2020). Mixture gaussian v2 based microscopic movement detection of human spermatozoa. *International Journal of Advances in Intelligent Informatics*, *6*(2), 210–222. https://doi.org/10.26555/ijain.v6i2.507

Uddin, M. P., Mamun, M. Al, & Hossain, M. A. (2019). Effective feature extraction through segmentation-based folded-PCA for hyperspectral image classification. *International Journal of Remote Sensing*, *40*(18), 7190–7220. https://doi.org/10.1080/01431161.2019.1601284

Velliangiri, S., Alagumuthu krishnan, S., & Thankumar Joseph, S. I. (2019). A Review of Dimensionality Reduction Techniques for Efficient Computation. *Procedia Computer Science*, *165*, 104–111. https://doi.org/10.1016/j.procs.2020.01.079

Vijayalakshmi A, & Rajesh Kanna B. (2020). Deep learning approach to detect malaria from microscopic images. *Multimedia Tools and Applications*, *79*(21–22), 15297–15317. https://doi.org/10.1007/s11042-019-7162-y

World Health Organization. (2023). World malaria report 2023. https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2023.