J-Icon : Jurnal Informatika dan Komputer
Vol. 11 No. 2, Oktober 2023, pp. 237~245
DOI: 10.35508/jicon.v11i2.12658

CLASSIFICATION OF DISTRACTED DRIVER USING SUPPORT VECTOR MACHINE BASED ON PRINCIPAL COMPONENT ANALYSIS FEATURE REDUCTION AND CONVOLUTIONAL NEURAL NETWORK

Achmad Yuneda Alfajr[1*], Kartini[2], Anggraini Puspita Sari[3]

[1,2,3]Program Studi Informatika, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Jl. Rungkut Madya No.1, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur
[1]Email*: 19081010128@student.upnjatim.ac.id
[2]Email: kartini.if@upnjatim.ac.id
[3]Email: anggraini.puspita.if@upnjatim.ac.id

ABSTRAK

Penggunaan alat transportasi darat di Indonesia terutama di kota-kota besar seperti Surabaya, mengalami pertumbuhan yang pesat. Namun, peningkatan penggunaan ini juga berdampak pada peningkatan kecelakaan lalu lintas. Salah satu faktor penyebab utamanya adalah distraksi yang dialami oleh pengemudi. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan metode untuk mendeteksi pengemudi yang terganggu menggunakan teknologi klasifikasi gambar dengan model *Convolutional Neural Networks* (CNN) dan *Support Vector Machine* (SVM). Dalam penelitian ini, data diperoleh dari dataset "*State Farm Distracted Driver Detection*" yang berisi gambar-gambar pengemudi yang sedang berkendara dalam kondisi terganggu atau tidak fokus. Proses awal melibatkan pra-pemrosesan data, seperti mengubah ukuran gambar menjadi 50 x 50 piksel dan membagi dataset menjadi data latih dan data uji. Kemudian, ekstraksi fitur dilakukan menggunakan model CNN dengan tiga lapisan konvolusi, tiga lapisan *Maxpooling*, dan satu lapisan *Flatten*. Setelah ekstraksi fitur, metode *Principal Component Analysis* (PCA) digunakan untuk mengurangi dimensi data. Selanjutnya, model SVM dilatih menggunakan data hasil reduksi fitur oleh PCA dengan pembagian data yang digunakan adalah 60:40. Penelitian ini melakukan perbandingan antara menggunakan PCA dan tanpa PCA. Berdasarkan hasil pengujian, penggunaan PCA tidak hanya meningkatkan akurasi klasifikasi hingga 96.28% dibandingkan dengan 92.46% tanpa PCA, tetapi juga mempercepat waktu pelatihan menjadi 10.64 detik dari 19.67 detik tanpa PCA.
Kata kunci: Pengemudi Terganggu, CNN, PCA, SVM

ABSTRACT

The use of ground transportation in Indonesia especially in major cities like Surabaya has experienced rapid growth. However, this increased usage has also led to a rise in traffic accidents. One of the main contributing factors is driver distraction. Therefore, this study aims to develop a method for detecting distracted drivers using image classification technology with Convolutional Neural Networks (CNN) and Support Vector Machine (SVM) models. In this research, data was obtained from the "State Farm Distracted Driver Detection" dataset, which contains images of drivers who are either distracted or not focused while driving. The initial process involves data preprocessing, such as resizing images to 50 x 50 pixels and dividing the dataset into training and testing data. Next, feature extraction is performed using a CNN model with three convolutional layers, three Maxpooling layers, and one flattened layer. After feature extraction, the Principal Component Analysis (PCA) method is used to reduce the dimensionality of the data. Furthermore, an SVM model was trained using data reduced by PCA with a 60:40 data split. This research conducted a comparison between the use of PCA and not using PCA. Based on the test results, the use of PCA not only improved the classification accuracy to 96.28% compared to 92.46% without PCA but also accelerated the training time to 10.64 seconds from 19.67 seconds without PCA.
Keywords: Distracted Drivers, CNN, PCA, SVM

1. INTRODUCTION

The utilization of ground transportation in Indonesia especially in major cities like Surabaya is exceptionally high. This is driven by the presence of well-developed road infrastructure designed to facilitate traffic flow [1]. However, the flip side of this coin is the high incidence of traffic accidents. In developing countries, statistics indicate that 1.25 million people die each year as a result of these accidents.

***) Correspondent Author**
Submitted: 25 September 2023
Accepted: 24 October 2023
Published: 31 Oktober 2023

ISSN: 2337-7631 (Printed)
ISSN: 2654-4091 (Online)

One of the primary causes is driver impairment or distraction, such as the use of electronic devices, eating, and drinking while driving [2][3].

Data from Surabaya in 2017 recorded 1.338 accidents, with 461 of them involving private cars [4]. These accidents resulted in 71 fatalities and 47 people suffering severe injuries. Driver negligence, particularly due to distractions like texting or eating, is a primary factor that elevates the risk of accidents. To address this issue, the primary focus is on road safety and driver behavior. One proposed approach is to leverage cutting-edge technology in image classification. This technology enables the monitoring of driver actions and categorizing them as either distracting or not. Previous studies have extensively delved into this field. One research has found that CNN models outperform Vision Transformers in detecting distracted drivers, achieving an accuracy of 90% using a pre-trained DenseNet model [5]. In December 2019, Qingzhi Bu and colleagues conducted a study titled "Research on a Driver's Distracted Behavior Detection Method Based on Multiclass Classification and SVM." This research utilized the Histogram of Oriented Gradient (HOG) method for feature extraction and Support Vector Machine (SVM) as the classifier. In this study, the obtained accuracy result was 90.84% [6].

Taking into consideration all the existing findings and methods, researchers are embarking on further research into detecting distracted drivers. The method to be employed involves feature extraction using CNN, followed by feature reduction through PCA before the training process using SVM models. The primary objective of this research is to determine the accuracy in distinguishing between distracted and non-distracted drivers, thus potentially making a significant contribution to road safety improvement efforts.

2. MATERIAL AND METHODS

Distracted Driving

Distraction or impairment while driving is a leading cause of road accidents, accounting for 25% of the total accidents that occur. This distraction is defined as a shift of focus from the primary task of driving to secondary tasks, such as looking at billboards, using a smartphone, or even conversing with passengers. According to the National Highway Traffic Safety Administration, some indications of distracted drivers include talking on the phone, even in "hands-free" mode, texting, adjusting the radio or AC, as well as eating and drinking [7].

This level of distraction can be gauged through various indicators. For instance, drivers who are completely focused on driving will be recorded as "Undistracted," while those engaged in phone conversations or operating the radio will be categorized as distracted. This situation can also be exacerbated by the presence of other passengers in the vehicle, further diverting the driver's attention. Figure 1 serves as an example of the driver being in a distracted state.

Figure 1. Example of distracted driver

Convolutional Neural Networks (CNN)

CNN are a subset of neural networks specifically crafted for processing data with grid-like or lattice structures, such as images. Utilizing mathematical operations known as convolution, CNNs effectively extract essential features from input data [8]. The typical structure of a CNN typically involves three main types of layers: convolutional layers, activation layers, and pooling layers. In the convolutional layer, a convolution operation occurs between the input image matrix and filters, and it is described as [9].

$$n_{out} = \left(\frac{n_{in} - k + 2p}{s}\right) + 1 \quad (1)$$

ISSN: 2337-7631 (Printed)
ISSN: 2654-4091 (Online)

Activation functions like ReLU (Rectified Linear Unit) are employed to introduce non-linearity elements into the network, enabling CNN to grasp more intricate patterns. The ReLU function is defined by Equation 2 [10].

$$f(x) = max(0, x) \tag{2}$$

Pooling layers, such as max pooling or average pooling, serve to reduce the dimensionality of the feature map. This not only accelerates computations but also extracts dominant features from the data [8]. The final stage in the CNN architecture typically involves the "Flatten" process, where the two-dimensional feature map is transformed into a one-dimensional vector. This facilitates the integration between CNN and other classification algorithms such as SVM, which require one-dimensional vector format inputs.

Principal Component Analysis (PCA)

PCA is a statistical technique utilized for dimensionality reduction of data without significant information loss. By rotating the original coordinate system, PCA establishes a new, simpler coordinate system that facilitates data interpretation and minimizes errors in the classification process [11][12]. This technique is effective in dealing with highly correlated data with numerous variables, consolidating them into a smaller number of principal components. PCA operates through specific algorithmic steps, including calculations of data mean and covariance matrices, as well as the extraction of eigenvectors and eigenvalues. Mathematical formulas like Equation 3 and 4 are utilized in this process. Subsequently, the resulting principal components are employed as a new representation of the original data, often with lower dimensions but retaining high information retention.

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{3}$$

$$C_x = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right) \left(X_i - \overline{X}\right)^T \tag{4}$$

Support Vector Machine (SVM)

SVM is a pivotal method in machine learning introduced by Vapnik in 1992. This method emphasizes Structural Risk Minimization (SRM) to discover the optimal "hyperplane" that separates data between classes [13]. SVM is employed for various tasks, including classification and regression. In classification, for instance, SVM seeks the hyperplane that maximizes the distance between classes [14]. Furthermore, SVM has also been extended to handle multi-class classification. This approach involves techniques like *'one-vs-rest'* and *'one-vs-one'*. In this model, each class is represented by a vector w_m, and classification is performed by maximizing the objective function in Equation 5, where $w_m^T x$ represents the score of class m against vector x [15].

$$\hat{y} = \underset{m \in [k]}{\operatorname{argmax}}\, w_m^T x \tag{5}$$

Model parameters, such as weight vectors w_m, are derived from the optimization process. In the context of multi-class problems, this optimization typically involves minimizing the objective function related to margin and classification loss. The problem formulation can take either a "primal" or "dual" form, and the proper parameter selection is crucial for model performance [16].

Confusion Matrix

Confusion Matrix is a table that compares the number of correct and incorrect predictions made by a classification model. The rows in this matrix represent the actual data classes, while the columns represent the predicted classes. From the Confusion Matrix, one can gauge the model's performance and take steps to enhance its performance, such as increasing training data or adjusting model parameters. Furthermore, there are three crucial evaluation metrics in multi-class classification model performance analysis: accuracy, recall, precision, and F1-score. Accuracy measures how precisely the model classifies data overall by dividing the number of correct predictions by the total data, using Equation 6. On the other hand, Recall measures the extent to which the model can identify a specific class by dividing the True Positives (correct predictions) of that class by the total actual data of that class, using Equation 7. Precision measures how accurately the model predicts a specific class by dividing the True Positives for that class by the total predictions for that class, using Equation 8 [17]. The F1-Score, which is a metric used to assess the performance of a classification model, represents a harmonious balance between precision (the ability of the model to correctly identify relevant instances) and recall (the ability of the model to capture all relevant instances), using Equation 9 [18].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \tag{6}$$

$$Recall = \frac{TP}{TP+FP}\ x\ 100\% \tag{7}$$

ISSN: 2337-7631 (Printed)
ISSN: 2654-4091 (Online)

$$Precision = \frac{TP}{TP+FN} x\ 100\% \quad (8)$$

$$F1 - score = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall} \quad (9)$$

Research Flow

In this research, the method to be employed involves data preprocessing, and feature extraction using CNN, followed by feature reduction with PCA before the training process using an SVM model for the classification of distracted drivers. Figure 2 illustrates the research flow, which begins with data collection, data preprocessing, data splitting, CNN modeling, PCA modeling, SVM modeling, and testing scenarios.

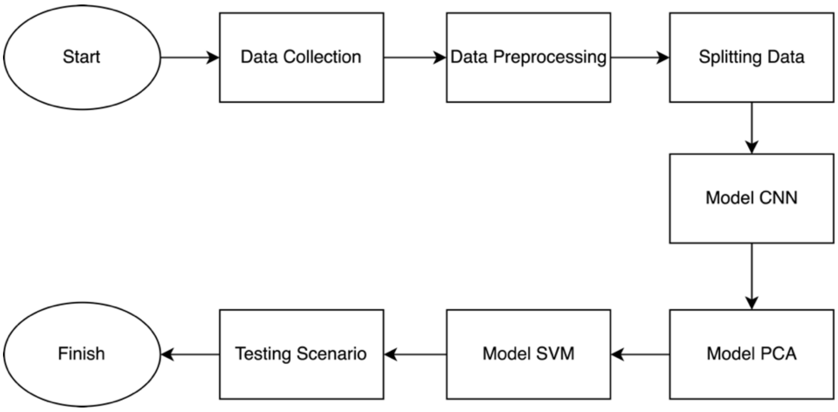

Figure 2. Research Flow

Data Collection

In the data collection process, the data source utilized is obtained from Kaggle, an open-source platform, under the title "State Farm Distracted Driver Detection". This data is supplied by the insurance company State Farm and comprises a series of images of drivers captured by a camera located in front of the driver. Each image is labeled to indicate whether the driver is focused or distracted [19].

Table 1. Data Distribution

No	Class	Number of Data
1	safe driving	2489
2	texting – right	2267
3	talking on the phone – right	2317
4	texting – left	2346
5	talking on the phone – left	2326
6	operating the radio	2312
7	Drinking	2325
8	reaching behind	2002
9	hair and makeup	1911
10	talking to passenger	2129
	Total Data	22424

The issue of distracted driving has been a primary focus in efforts to enhance road safety. This is because distracted driving can lead to accidents, both minor and major. Therefore, State Farm has provided this dataset to assist in the development of artificial intelligence models that can detect driver conditions, aiming to reduce accidents and improve road safety. This dataset comprises approximately 22,424 labeled images used for model training. Each image in this dataset is labeled with one of ten different labels, each depicting a specific activity being carried out by the driver. The distribution of the dataset will be presented in Table 1.

Data Preprocessing

Before performing feature extraction on the "Distracted Drivers" dataset, preprocessing steps need to be carried out, as illustrated in Figure 3. The preprocessing step is to resize the image to 50 x 50 pixels from its original size of 480 x 640 pixels. This resizing is done to expedite the model training process. By reducing the image size to 50 x 50 pixels, the total number of pixels to be processed becomes significantly lower. This can reduce the computational load required during training, thus speeding up the time needed

ISSN: 2337-7631 (Printed)
ISSN: 2654-4091 (Online)

to generate a suitable model. Additionally, this uniform resizing enhances visual consistency across the dataset, enabling the model to identify consistent patterns irrespective of variations in image size or aspect ratio. Furthermore, this preprocessing step helps in reducing visual noise commonly found in high-resolution images, which may be irrelevant to the task at hand. By resizing the images, this visual noise is effectively diminished, allowing the model to focus more on essential features and patterns. Figure 3 outlines the Data Preprocessing flow.

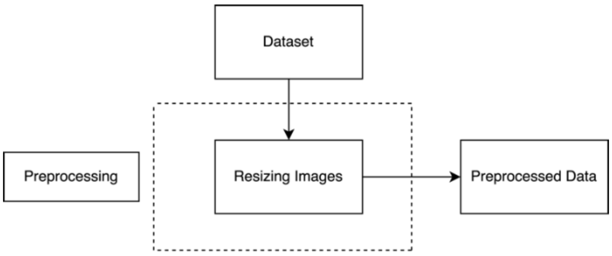

Figure 3. Data Preprocessing Flow

Splitting Data

The data splitting in this research serves several crucial purposes. It enables the evaluation of the model's effectiveness in predicting unknown data, providing insights into the model's ability to generalize. Secondly, this process aids in preventing overfitting, which is a condition where the model becomes overly focused on the training data and fails to accurately predict new data. If this occurs, corrective actions such as regularization or model simplification can be taken. Thirdly, data splitting supports parameter optimization, allowing for comparisons between various models or parameter sets to find the most effective one [20]. The data splitting scenario to be used in the testing phase will be 60:40.

Model CNN (Feature Extraction)

Feature extraction in this research is handled by CNN, designed with three convolutional layers, three MaxPooling layers, and one Flatten layer. This first convolutional layer operates on input images that are dimensioned at 50x50 pixels and possess three color channels (comprising the RGB spectrum) as seen in Figure 4. Through this process, the convolutional layer generates a set of 32 feature maps, each intended to capture specific patterns and information within the image. The MaxPooling layers reduce the size of these feature maps, speeding up computations and reducing model complexity.

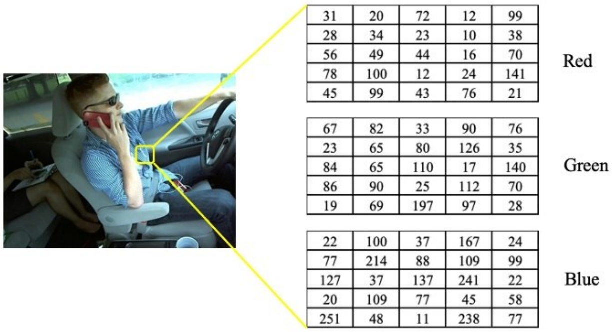

Figure 4. Illustration of RGB channel on dataset

The convolutional layers play a pivotal role in identifying image features, such as edges or curves, through a process called convolution. In this process, filters or kernels are slid across the entire image, and the multiplication results between filter elements and image elements are summed to create a matrix that indicates the filter's response to image features. Convolutional layers are typically applied repeatedly, producing a series of feature maps used as input for the next layers in the CNN architecture. Subsequently, the flattened layer transforms the feature maps into a one-dimensional vector. The entire architecture, as a whole, functions to extract relevant features from the input images.

As seen in Figure 5, the CNN architecture comprises key layers, starting with a Conv 2D layer equipped with 32 filters, each sized (3,3), and applying the ReLU activation function using Equation 2. This layer conducts convolution on the input images, resulting in 32 feature maps. The input shape for this layer is (50, 50, 3), indicating the input images are 50x50 pixels with three color channels (as seen in standard RGB images). Subsequently, a MaxPool 2D layer, featuring a (2,2) filter, effectively reduces the

ISSN: 2337-7631 (Printed)
ISSN: 2654-4091 (Online)

dimensions of the feature maps by half simplifying model complexity and enhancing computational efficiency. Following this, a second Conv 2D layer incorporates 64 filters, each with a (3,3) filter size, and employs the ReLU activation function using Equation 2. This layer conducts convolution on the feature maps obtained from the initial MaxPool 2D layer, resulting in 64 feature maps. A subsequent MaxPool 2D layer further reduces these feature map dimensions by half. For the third Conv 2D layer, 128 filters with a (3,3) filter size are employed, using the ReLU activation function specified in Equation 2. This layer conducts convolution on the feature maps stemming from the second MaxPool 2D layer, leading to the production of 128 feature maps. Once again, a MaxPool 2D layer follows to reduce dimensions by half. The final layer, the Flatten Layer, transforms these feature maps from the third MaxPool 2D layer into a one-dimensional vector.

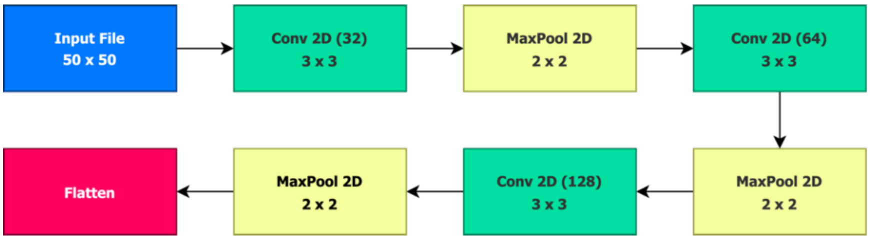

Figure 5. CNN Architecture

Model PCA (Feature Reduction)

After the CNN model extracts features from the data, PCA is employed to reduce its dimensionality. The purpose of using PCA is to expedite training time by reducing the number of features without losing crucial information from the images. Dimensionality reduction is particularly important when the features extracted by CNN have a very high dimension, making the training process more efficient with PCA. The Process Diagram of the PCA model is presented in Figure 6.

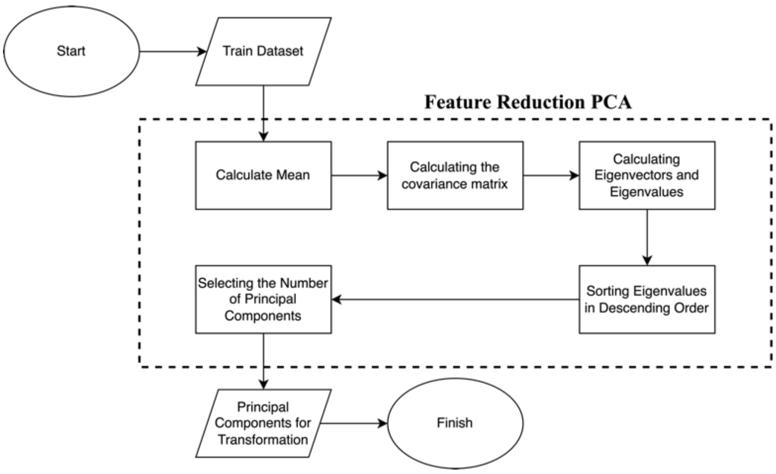

Figure 6. Feature Reduction Flow using PCA

Model SVM (Classifier)

The training of the model involves the utilization of SVM as a classifier as seen in Figure 7. Before the classification stage can commence, a crucial preprocessing step is undertaken with feature reduction using PCA. The result from PCA is employed with the dual objective of expediting the SVM training process while ensuring that essential information is retained. The output of PCA yields pairs of training data and corresponding labels, encompassing a total of 10 unique labels. These labeled training data are then fed into the SVM algorithm, which leverages this information to learn the underlying patterns and relationships within the data.

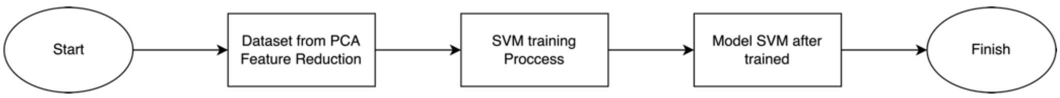

Figure 7. Training Proccess using SVM

ISSN: 2337-7631 (Printed)
ISSN: 2654-4091 (Online)

Testing Scenario

The system testing is conducted with the primary objective of attaining the highest possible classification accuracy. In pursuit of this goal, the testing process involves the careful selection of the optimal number of feature reductions. The testing phase is specifically focused on determining the most suitable number of PCA feature reductions for the SVM. This comparative testing is conducted based on the results of PCA feature reduction, which reduces the initially extracted features from the CNN. Additionally, the study involves a comparison between the features obtained through PCA reduction and the direct use of the initial features from the CNN, thereby assessing the influence of PCA on improving the SVM's performance.

3. RESULT AND DISCUSSION

Training and Testing with PCA Feature Reduction

The phase of training and testing is aimed at comprehensively understanding the influence of feature reduction through PCA on the performance of the SVM model when it comes to classifying distracted drivers. Within this testing framework, the emphasis lies in comparing the effects of PCA-based feature reduction to the original features extracted from the CNN, yielding valuable insights into their influence on classification performance, as illustrated in Table 2.

Table 2. Comparison between with PCA and without PCA

Metric	With PCA (%)	Without PCA (%)
Percentage of Variance	98.9556	100
Precision	96.28	93.28
Recall	96.13	92.45
F1-Score	96.17	92.47
Accuracy	96.28	92.46
Train Time (seconds)	10.6472	19.6743

Based on the results presented in Table 2, it is evident that the feature reduction through PCA had a significant impact on the classification performance of the SVM model in the context of classifying distracted drivers. The experimental findings reveal intriguing insights into the effects of different feature reduction levels. When the number of features is reduced using PCA, the resulting accuracy significantly improves to 96.28% compared to the scenario without PCA, which achieves an accuracy of 92.46%. This enhancement in accuracy underscores the utility of PCA in improving the model's ability to correctly classify distracted drivers. Additionally, the training time is substantially reduced to 10.6472 seconds compared to the 19.6743 seconds required without PCA. This signifies that PCA not only enhances accuracy but also accelerates the training process, making it a valuable tool for optimizing the SVM model's efficiency. Furthermore, it's important to mention that the percentage of variance explained by PCA is approximately 98.9556%. This indicates that PCA effectively captures a substantial portion of the data's variability while reducing the dimensionality.

Figure 8. The result of the SVM classification

This dimensionality reduction not only contributes to improved efficiency but also aids in preserving the essential patterns within the data. The experimental results highlight the advantageous impact of PCA in

ISSN: 2337-7631 (Printed)
ISSN: 2654-4091 (Online)

terms of accuracy enhancement and training time reduction. The percentage of variance explained by PCA underscores its effectiveness in preserving important data characteristics while reducing feature dimensions, ultimately contributing to the overall improvement in model performance. The result of the SVM classification with an example of the classification outcome is shown in Figure 8.

4. CONCLUSION AND SUGGESTION

The research findings underscore the effectiveness of employing a combination of Convolutional Neural Networks (CNN) for feature extraction, Principal Component Analysis (PCA) for feature reduction, and Support Vector Machine (SVM) for the classification of distracted drivers. The experimentation involving PCA-based feature reduction emerges as a crucial setup. The outcomes reveal a substantial improvement in classification accuracy, with an impressive increase to 96.28% compared to the 92.46% achieved without PCA. This reduction in feature dimensions contributes to a remarkable reduction in training time, from 19.67 seconds without PCA to just 10.6472 seconds with PCA.

For future research, several suggestions can be proposed based on these findings. First, alternative methods beyond CNN, PCA, and SVM could be explored to compare their accuracy and training duration. Second, the machine learning models developed could be integrated into web or mobile applications to facilitate real-time classification of distracted driving behavior.

REFERENCES

[1] H. Asyari, F. Maulana, K. Muhammad, dan R. Aulia Imran, “Pengaruh Driving Distraction Penggunaan Smartphone Terhadap Pengemudi Sebagai Penyebab Kecelakaan Lalu Lintas Dengan Multilevel Factorial,” *Din. Rekayasa*, vol. 18, no. 244159, hal. 99–108, 2022.

[2] L. Lady dan A. Umyati, “Human Error dalam Berkendara Berdasarkan Kebiasaan Pelanggaran oleh Pengemudi,” *J. Manaj. Transp. Logistik*, vol. 8, no. 1, hal. 21, 2021, doi: 10.54324/j.mtl.v8i1.510.

[3] N. C. for S. and Analysis, “Distracted Driving 2018,” *Dot Hs 812 132*, no. April 2020, hal. 1–7, 2020, [Daring]. Tersedia pada: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812381

[4] I. W. Agustin, C. Meidiana, dan S. Muljaningsih, “Studi Simulasi Model Kecelakaan Pengendara Mobil untuk Meningkatkan Keselamatan Lalu Lintas di Daerah Perkotaan,” *War. Penelit. Perhub.*, vol. 32, no. 2, hal. 93–102, 2020, doi: 10.25104/warlit.v32i2.1513.

[5] H. V. Koay, J. H. Chuah, dan C. O. Chow, “Convolutional Neural Network or Vision Transformer? Benchmarking Various Machine Learning Models for Distracted Driver Detection,” *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2021-Decem, hal. 417–422, 2021, doi: 10.1109/TENCON54134.2021.9707341.

[6] Q. Bu, J. Qiu, H. Wu, dan C. Hu, “Research on driver’s distracted behavior detection method based on multiclass classification and SVM,” *IEEE Int. Conf. Robot. Biomimetics, ROBIO 2019*, no. December, hal. 444–448, 2019, doi: 10.1109/ROBIO49542.2019.8961551.

[7] NHTSA, “Research Note Distracted Driving 2021,” *Dot Hs 812 132*, vol. 2019, no. April 2015, hal. 1–8, 2023, [Daring]. Tersedia pada: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812381

[8] I. Goodfellow, Y. Bengio, dan A. Courville, *Deep learning*. MIT Press, 2016. [Daring]. Tersedia pada: https://www.deeplearningbook.org/

[9] A. Hibatullah dan I. Maliki, “Penerapan Metode Convolutional Neural Network Pada Pengenalan Pola Citra Sandi Rumput,” hal. 1–8, 2019.

[10] A. P. Sari, H. Suzuki, T. Kitajima, T. Yasuno, D. A. Prasetya, dan N. Nachrowie, “Prediction model of wind speed and direction using convolutional neural network - Long short term memory,” *PECon 2020 - 2020 IEEE Int. Conf. Power Energy*, hal. 356–361, 2020, doi: 10.1109/PECon48942.2020.9314474.

[11] M. Z. Nasution, “Penerapan Principal Component Analysis (PCA) Dalam Penentuan Faktor Dominan Yang Mempengaruhi Prestasi Belajar Siswa (Studi Kasus : SMK Raksana 2 Medan),” *J. Teknol. Inf.*, vol. 3, no. 1, hal. 41, 2019, doi: 10.36294/jurti.v3i1.686.

[12] Adiwijaya, U. N. Wisesty, E. Lisnawati, A. Aditsania, dan D. S. Kusumo, “Dimensionality reduction using Principal Component Analysis for cancer detection based on microarray data classification,” *J. Comput. Sci.*, vol. 14, no. 11, hal. 1521–1530, 2018, doi: 10.3844/jcssp.2018.1521.1530.

[13] I. P. Monika dan M. T. Furqon, “Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 10, hal. 3165–3166, 2018, [Daring]. Tersedia pada: http://j-ptiik.ub.ac.id

[14] N. Maulidah, R. Supriyadi, D. Y. Utami, F. N. Hasan, A. Fauzi, dan A. Christian, “Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes,” *Indones. J.*

Softw. Eng., vol. 7, no. 1, hal. 63–68, 2021, doi: 10.31294/ijse.v7i1.10279.

[15] M. Blondel, A. Fujino, dan N. Ueda, "Large-scale multiclass support vector machine training via euclidean projection onto the simplex," *Proc. - Int. Conf. Pattern Recognit.*, no. 4, hal. 1289–1294, 2014, doi: 10.1109/ICPR.2014.231.

[16] K. Crammer dan Y. Singer, "On The Algorithmic Implementation of Multiclass Kernel-based Vector Machines," *J. Mach. Learn. Res.*, vol. 2, hal. 265–292, 2001.

[17] B. P. Pratiwi, A. S. Handayani, dan S. Sarjana, "Pengukuran Kinerja Sistem Kualitas Udara Dengan Teknologi Wsn Menggunakan Confusion Matrix," *J. Inform. Upgris*, vol. 6, no. 2, hal. 66–75, 2021, doi: 10.26877/jiu.v6i2.6552.

[18] M. Anand, A. Velu, dan P. Whig, "Prediction of Loan Behaviour with Machine Learning Models for Secure Banking," *J. Comput. Sci. Eng.*, vol. 3, no. 1, hal. 1–13, 2022, doi: 10.36596/jcse.v3i1.237.

[19] W. K. Anna Montoya, Dan Holman, SF_data_science, Taylor Smith, "State Farm Distracted Driver Detection," *Kaggle*, 2016. https://www.kaggle.com/competitions/state-farm-distracted-driver-detection

[20] I. O. Muraina, "Ideal Dataset Splitting Ratios in Machine Learning Algorithms: General Concerns for Data Scientists and Data Analysts," *7th Int. Mardin Artuklu Sci. Res. Conf.*, no. February, hal. 496–504, 2022.

ISSN: 2337-7631 (Printed)
ISSN: 2654-4091 (Online)