INDONESIAN SIGN LANGUAGE IMAGE DETECTION USING CONVOLUTIONAL NEURAL NETWORK (CNN) METHOD

Andreas Nugroho Sihananto*[1], Erista Maya Safitri[2], Yoga Maulana[3], Fikri Fakhruddin[4], Mochammad Ervinda Yudistira [5]
Department of Informatics, Universitas Pembangunan Nasional Veteran Jawa Timur[1,3,4,5]
Department of Information System, Universitas Pembangunan Nasional Veteran Jawa Timur[2]
E-mail address: andreas.nugroho.jarkom@upnjatim.ac.id[1], maya.si@upnjatim.ac.id [2], 18081010075@student.upnjatim.ac.id[3], 18081010085@student.upnjatim.ac.id[4], 18081010149@student.upnjatim.ac.id[5]

Received: 05, February, 2023 ***Revised: 12, February, 2023*** ***Accepted: 17, May, 2023***

ABSTRACT

In Indonesia, there are two sign languages utilized by the deaf community, SIBI and BISINDO. Unfortunately, the majority of non-deaf individuals and deaf companions are not proficient in sign language. To address this communication gap, information systems can play a pivotal role in recognizing sign language speech. Recently, researchers conducted a study using the Convolutional Neural Network (CNN) algorithm to predict sign language for both SIBI and BISINDO datasets. The aim was to develop a model that could accurately translate sign language into written or spoken language, thus bridging the gap between deaf and non-deaf individuals. The research found that the CNN algorithm performed optimally on epoch 50 for SIBI with a testing accuracy of 93.29%, while for BISINDO, it achieved the best result on epoch 40 with a testing accuracy of 82.32%. These results suggest that the CNN algorithm has the potential to accurately recognize and translate sign language, thus improving communication between deaf and non-deaf individuals in Indonesia.

Keywords: SIBI, BISINDO, CNN, Neural Network, Accuracy,

1. INTRODUCTION

Language is a medium of communication. Likewise with people who have certain limitations, one of which is speech impaired. A speech impediment is someone who is unable to speak. A person who is speech impaired communicates using sign language. Sign language is a language used by using body movements without the need for sound. As we know that the speech impaired not only communicate with other speech impaired people but also communicate with normal people who do not necessarily understand the gestures they give(Rahma et al., 2020). Without a mediator, the deaf may experience isolation and frustration. This situation becomes more problematic if those around them are not skilled in sign language. Furthermore, they may be unaware of urgent surrounding sounds, such as fire alarms, bells, and car horns, which causes them to make delayed decisions that may put their lives at risk. Additionally, the need for a mediator between a deaf person and other people may cause misunderstandings and miscommunication depending on the skillfulness of the mediator. Moreover, such dialogs sometimes include confidential information that requires a trusted mediator, who may not be readily available. Moreover, certain critical situations may require a DHH person to

communicate with a hearing person, such as in an emergency or visiting a doctor, where there is neither an available interpreter nor a layperson who knows sign language(Alkhalifa & Al-Razgan, 2018).

Moreover, there are various kinds of sign language used by the deaf. In Indonesia alone, there are two variants of sign language used by the deaf, namely SIBI (Indonesian Sign Language) and BISINDO (Indonesian Sign Language). The SIBI is based on American Sign Language and usually used in formal occasion but BISINDO is more widely used in informal situation(Mursita, 2015; Wedayanti, 2019). Because there is 2 variants of the sign language, sometimes it is very hard for common people to distinguish the sign language. Effort to make a system to detect the language is very important because WHO stated there are an estimated 360 million Deaf and Hard of Hearing (DHH) individuals globally (328 million adults and 32 million children).

As technology progresses, computer and machine vision systems have increasingly been utilized in the public health sector over the past few years. Deep learning, a recent breakthrough in the field of Artificial Intelligence, has been widely employed by researchers to classify images, and has emerged as a rapidly growing area of computer vision(Gill & Khehra, 2021). With the sign language image detection system using the conventional neural network method, it is hoped that it will make it easier for someone who wants to understand sign language. This makes the accuracy of the image classification process in this SIBI and BISINDO dataset important. Therefore, the author uses the Convolutional Neural Network (CNN) algorithm which is part of the Artificial Neural Network (ANN) algorithm which has good performance for image classification.

The public health sector has incorporated computer and machine vision systems with recent advances in the field of Artificial Intelligence, particularly in deep learning. Image classification using deep learning applications has become an emerging area in computer vision research (Gill & Khehra, 2021). The conventional neural network method is used for sign language image detection, which can facilitate understanding of sign language. The accuracy of image classification in the SIBI and BISINDO dataset is crucial. Therefore, the Convolutional Neural Network (CNN) algorithm is employed, which is a part of the Artificial Neural Network (ANN) algorithm known for its superior performance in image classification.

2. THEORY

Sistem Isyarat Bahasa Indonesia (SIBI)

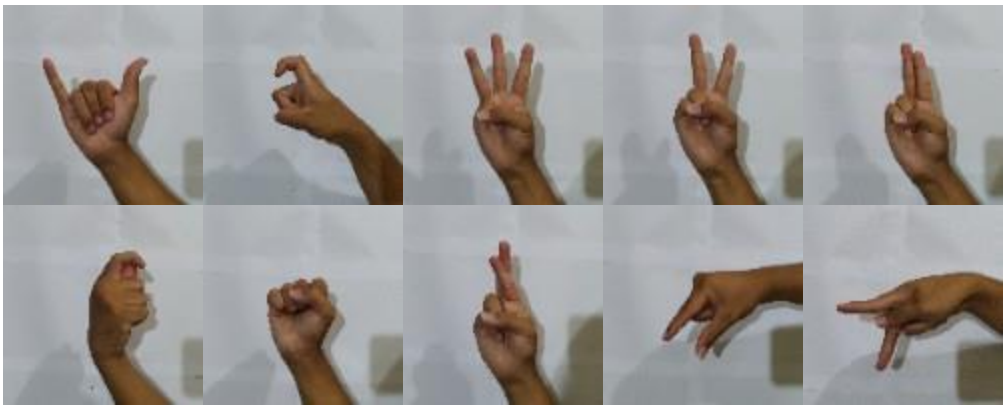

Figure 1. Example of SIBI Signs

SIBI is standardized sign language of Indonesia and formally used in school or important event. There is another variation of Sign Language that is called BISINDO but government of Indonesia formally suggest using SIBI for education and state business. It was modeled after American Sign Language with some modification.

Bahasa Isyarat Indonesia (BISINDO)

BISINDO is a sign language that appears naturally in from daily interaction of Indonesian's deaf and mute. It is widely used in everyday life as daily conversation. BISINDO is one of 100 sign languages that develop naturally among the world's deaf people thus has several variations in each region. This sign language has a grammar that is different from the spoken language used by hearing people in general, covering all elements, starting from phonology, morphology, syntax, pragmatics, and other elements.

Figure 2. Example of BISINDO Signs

While more commonly used than SIBI, BISINDO have many variants and somehow according to many parties accompanying deaf people is a more practical and effective communication system for deaf people in Indonesia than SIBI. Because BISINDO was developed by deaf Indonesians to be used as communication between hearing people. BISINDO itself originates from the initial language / mother tongue of the deaf, where the use of BISINDO itself adapts to the understanding of the deaf language from various deaf backgrounds without providing an Indonesian language affix structure(Mursita, 2015). But with many variants of BISINDO it will be hard for accompanying parties to learn all variants and it can be happened that one

people may not communicate with other people if they only know BISINDO if they use different variant of BISINDO.

Convolutional Neural Network (CNN)

Convolutional Neural Network, also known as convnet or CNN, is a machine learning method that belongs to the family of artificial neural networks (ANNs). It has a wide range of applications and can be utilized with different types of data sets. CNN is specifically designed for deep learning algorithms that require image recognition and processing of pixel data. While there are other types of neural networks for deep learning, CNNs are particularly well-suited for object identification and recognition tasks. As such, they are the preferred network design for computer vision (CV) tasks, and are particularly useful for applications where precise object recognition is critical, such as facial recognition or self-driving vehicle systems.

A convolutional layer, a pooling layer, and a fully connected (FC) layer make up a deep learning CNN. The first layer is the convolutional layer, while the final layer is the FC layer. The complexity of the CNN grows from the convolutional layer to the FC layer. The CNN is able to detect ever bigger and more intricate aspects of a picture until it successfully recognizes the complete thing as a result of the rising complexity.

a neural network specifically designed to process image data, in this case chest X-ray images of COVID-19 patients. CNN works by performing a convolution operation on the input image, in which a filter or kernel will be shifted at each input image pixel to extract important and relevant features. CNN then uses different layers to combine these features and produce predictions or classifications on the input image. In this article, 18 different CNN models with transfer learning techniques are used to diagnose COVID-19 on chest X-ray images, and quantitative and qualitative analyzes are performed to evaluate the performance of each model (Chow et al., 2023).

Convolutional layer. The convolutional layer, the central component of a CNN, is where most calculations take place. The first convolutional layer may be followed by a subsequent convolutional layer. A kernel or filter inside this layer moves over the image's receptive fields during the convolution process to determine if a feature is present (Wang et al., 2023).

The kernel traverses the entire picture over a number of iterations. A dot product between the input pixels and the filter is calculated at the end of each cycle. A feature map or convolved feature is the result of the dots being connected in a certain pattern. In this layer, the picture is ultimately transformed into numerical values that the CNN can understand and extract pertinent patterns from.

Pooling layer. The pooling layer functions in a similar manner to the convolutional layer in that it scans a kernel or filter across the input image. However, it differs from the convolutional layer in that it has fewer input parameters and results in some loss of information. Despite this, the pooling layer plays a beneficial role in streamlining the CNN and enhancing its efficiency (Sunkara & Luo, 2023).

Fully connected layer. The categorization of images in a CNN occurs in the fully connected (FC) layer, which utilizes the features extracted in the preceding layers. "Fully connected"

refers to the fact that each activation unit or node in the subsequent layer is connected to every input or node from the previous layer. However, not all layers in the CNN are fully connected, as this would create a dense network that is computationally expensive, increases losses, and ultimately reduces the quality of the output (Yin & Zhao, 2023).

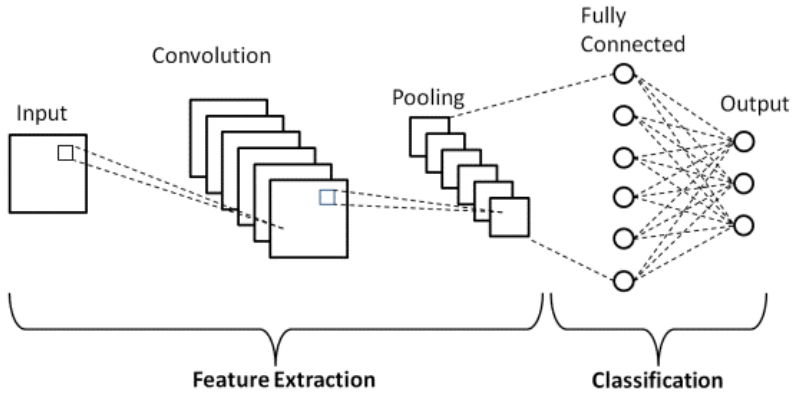

Figure 3. CNN Architecture

3. METHOD

For dataset we will use sample of alphabetic sign language from A to Z both in SIBI form and BISINDO form. Every alphabet training data containing 12 samples for BISINDO 18 samples for SIBI. The parameters for CNN can be seen on Table 1 meanwhile numbers of data can be seen on Table 2.

Table 1. CNN Models and Parameter

Layer	Size	Output
Input	(28, 28, 1)	-
Convolution + Relu	16 (3 x 3) filters	(26, 26, 16)
Batch Normalization	-	(26, 26, 16)
Max Pooling + Dropout	(2 x 2) filters	(13, 13, 16)
Convolution + Relu	32 (3 x 3) filters	(11, 11, 32)
Batch Normalization	-	(11, 11, 32)
Max Pooling + Dropout	(2 x 2) filters	(5, 5, 32)
Global Average Pooling	-	32
Flatten	-	32
Dense + Dropout	-	128
Softmax	24	24

Table 2. Numbers of data

Dataset	Testing	Training
SIBI	468 files	26 files
BISINDO	312 files	26 files

The image pre-processing phase involved splitting the dataset into train and test data, with 780 train data and 52 test data. The images were then converted to grayscale and resized to 28x28 pixels. The pre-processed data was stored in an array and labeled with 26 classes of alphabets. To build the CNN model, the parameters in Table 1 were used. The CNN architecture comprised of 2 convolutional layers, 2 pooling layers, a fully-connected layer, and a softmax

output layer. ReLu activation, batch normalization, and dropout regularization were applied in each layer. The fully-connected layer used global average pooling, flatten, and dense with Dropout regularization. The CNN was developed using the TensorFlow library.

4. RESULTS AND DISCUSSION

The result on testing on SIBI dataset can be seen on Table 3. It will be presented in graphic form on Figure 4 for Model Loss function and Figure 5 for Model Accuracy result.

Table 3. SIBI Dataset Cross validation result

Epoch	Train Loss	Test Loss	Training Accuracy	Testing Accuracy
10	0.0855	0.104	0.8704	0.7704
20	0.1159	0.121	0.9616	0.8616
30	0.092	0.109	0.9717	0.7717
40	0.1016	0.117	0.9711	0.9111
50	0.0736	0.091	0.9719	0.9329
60	0.0936	0.1	0.9501	0.9203

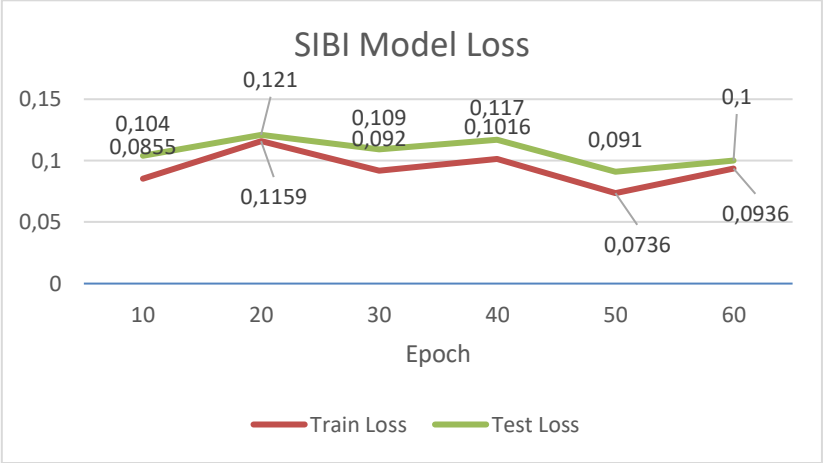

Figure 4. SIBI Model Loss

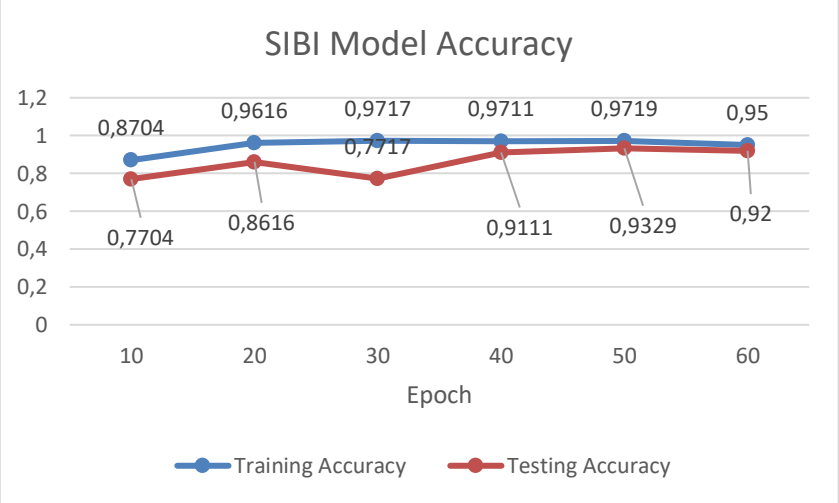

Figure 5. SIBI Model Accuracy

Based on result from Table 3, Figure 4 and Figure 5, we found that Model Loss Function for SIBI model reached minimum value at epoch value 50 with training test loss function valued at 0.0736, testing loss function at 0.091 with accuracy 97,19 % for training data. It resulted accuracy of testing data at 93,29%.

For BISINDO dataset, the cross-validation test result can be seen on Table 4. It will be presented in graphic form on Figure 6 for Model Loss function and Figure 7 for Model Accuracy result.

Table 4. BISINDO Dataset Cross validation result

Epoch	Train Loss	Test Loss	Training Accuracy	Testing Accuracy
10	0.1553	0.10885	0.92625	0.7907
20	0.15455	0.15485	0.93155	0.8224
30	0.16125	0.1977	0.92715	0.8055
40	0.1744	0.1569	0.9343	0.8232
50	0.158	0.1628	0.93055	0.8161

Meanwhile for BISINDO we found that Model Loss Function for BISINDO model reached minimum value at epoch value 40 with training test loss function valued at 0.1774, testing loss function at 0.1569 with accuracy 93,43% for training data. It resulted accuracy of testing data at 82,32%.

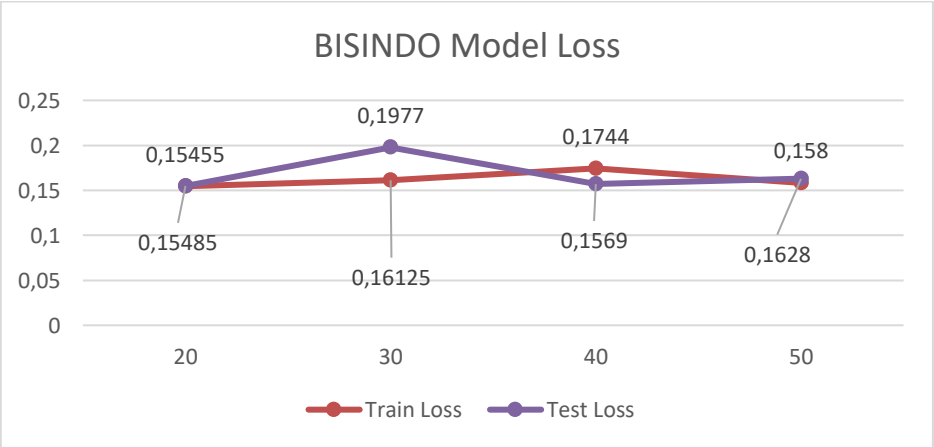

Figure 6. BISINDO Model Loss

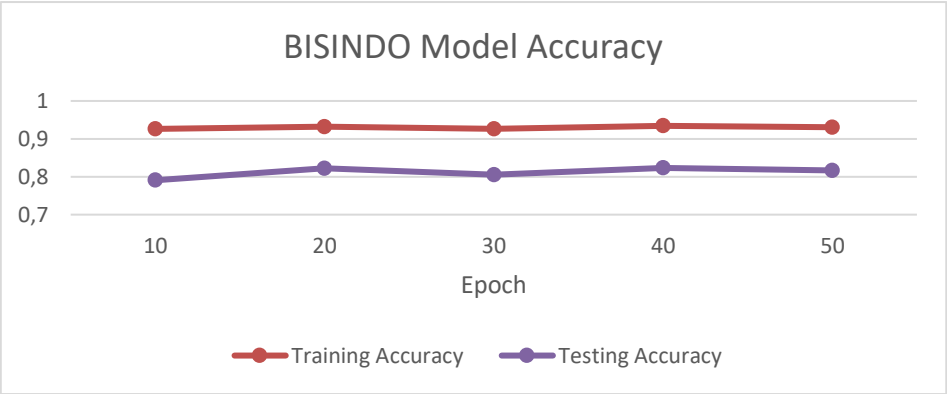

Figure 7. Model Accuracy of BISINDO

The results suggest that the model performs more effectively on the SIBI dataset. This could be attributed to the fact that the SIBI dataset is clearer and less noisy compared to the BISINDO dataset. The lower accuracy of the BISINDO dataset may be due to the use of standard preprocessing without any modification. However, it is worth noting that the BISINDO dataset's alphabetic characters require fewer epochs than SIBI. This may be due to the smaller dataset size of BISINDO in comparison to SIBI.

5. CONCLUSIONS AND SUGGESTIONS

The CNN model was utilized to detect the sign language alphabets for both SIBI and BISINDO datasets, and it yielded optimum results. The SIBI dataset had an optimum parameter on epoch 50, with a training accuracy of 97.19% and a testing accuracy of 93.29%. On the other hand, the BISINDO dataset had its optimum parameter at epoch 40, with a training accuracy of 93.43% and a testing accuracy of 82.32%. Since the BISINDO dataset contains more noise and lower resolution than the SIBI dataset, the noise reduction preprocessing method should be applied to the dataset in the future to improve its accuracy. Algorithms such as threshold and low rank minimization may be utilized to reduce the noise. Additionally, building a modified CNN model such as CNN-RNN hybrid, CNN-LSTM hybrid, CNN-SVM, or YOLO may also be useful for this dataset to determine whether accuracy can be improved.

6. ACKNOWLEDGEMENTS

We want to express our thanks to Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) Universitas Pembangunan Nasional “Veteran” Jawa Timur who provide grant via “Uber Publikasi Jurnal Nasional Terakreditasi” Schema that allowed us to write and complete this research.

REFERENCES

Alkhalifa, S., & Al-Razgan, M. (2018). Enssat: wearable technology application for the deaf and hard of hearing. *Multimedia Tools and Applications*, *77*(17), 22007–22031. https://doi.org/10.1007/s11042-018-5860-5

Chow, L. S., Tang, G. S., Solihin, M. I., Gowdh, N. M., Ramli, N., & Rahmat, K. (2023). Quantitative and Qualitative Analysis of 18 Deep Convolutional Neural Network (CNN) Models with Transfer Learning to Diagnose COVID-19 on Chest X-Ray (CXR) Images. *SN Computer Science*, *4*(2), 1–17. https://doi.org/10.1007/s42979-022-01545-8

Gill, H. S., & Khehra, B. S. (2021). An integrated approach using CNN-RNN-LSTM for classification of fruit images. *Materials Today: Proceedings*, *51*(xxxx), 591–595. https://doi.org/10.1016/j.matpr.2021.06.016

Mursita, R. A. (2015). Respon Tunarungu Terhadap Penggunaan Sistem Bahasa Isyarat Indonesa (Sibi) Dan Bahasa Isyarat Indonesia (Bisindo) Dalam Komunikasi. *Inklusi*, *2*(2), 221. https://doi.org/10.14421/ijds.2202

Rahma, U., Perwiradara, Y., Ikawikanti, A., Mayasari, B. M., Rinanda, T. D., Brawijaya, U., & Malang, K. (2020). School Wellbeing Analysis Among Visual Impairments, Deaf and Physical Disability Students in College Inclusion. *Jurnal Fakultas Psikologi Universitas Wisnuwardhana Malang*, *24 No : 1*(1), 16–32.

Sunkara, R., & Luo, T. (2023). No more strided convolutions or pooling: a new CNN building block for low-resolution images and small objects. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part III*, 443–459. https://doi.org/10.1007/978-3-031-26409-2_27

Wang, Z., Wang, Z., Zeng, C., Yu, Y., & Wan, X. (2023). High-quality image compressed sensing and reconstruction with multi-scale dilated convolutional neural network. *Circuits, Systems, and Signal Processing*, *42*(3), 1593–1616. https://doi.org/10.1007/s00034-022-02181-6

Wedayanti, Ni Putu Luhur. TEMAN TULI DIANTARA SIBI DAN BISINDO. Proceedings, [S.l.], p. 137-146, oct. 2019. Available at: <http://ojs.pnb.ac.id/index.php/Proceedings/article/view/1513>.

Yin, L., & Zhao, M. (2023). Inception-embedded attention memory fully-connected network for short-term wind power prediction. *Applied Soft Computing*, *141*, 110279. https://doi.org/10.1016/j.asoc.2023.110279