A Transfer Learning Strategy for Owl Sound Classification by Using Image Classification Model with Audio Spectrogram

Kevin William Gunawan[1], Alam Ahmad Hidayat[3], Tjeng Wawan Cenggoro[2,3] and Bens Pardamean[1,3]

[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science Program, Bina Nusantara University, Jakarta, Indonesia
[2]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
[3]Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta, Indonesia

Abstract: This paper presents an improved approach to train models that can be used to accurately predicting animal presence based on its sound with a limited dataset. Currently, deep-learned models dominate the state-of-the-art methods for audio classification tasks for their predictive capabilities. However, an immense amount of data is needed to build an accurate deep learned classifier. Such immensity of data is usually hard to be satisfied on endemic or endangered animals. For example, collecting an Indonesian scops owl audio dataset for our experiment in an adequate amount is insatiable, thus may reduce the predictive capability of a deep-learned model. To overcome such an issue, we propose a transfer learning strategy that alleviates overfitting in a deep model and a way to maximize the use of datasets by extracting two acoustic features: Mel's spectrogram and Mel Frequency Cepstral Coefficient (MFCC) from each data point. In this study, we employ a dual-input scalable Convolutional Neural Network (CNN) derived from EfficientNet [1] which utilizes and learn from both acoustic features. Our experimental pre-trained dual-input network achieves 99.27% mAP on our testing data accuracy whereas a trained-from-scratch Resnet-50 model used as the baseline model achieves 99% mAP on the same testing set.

1. Introduction

Manual processing of an avian sound annotation and detection is inconvenient and can hamper bird conservation efforts. Therefore, the need for accurate, scalable, and automated bird presence recognition is vital for wildlife monitoring and can be beneficial for avian biodiversity conservation. In this regard, deep learning methods of bird detection are a suitable choice to combat the rapid loss of avian diversity. The applications of deep learning in sound identification have been widely recognized and have dominated several annual bird sound identification competitions such as Bird Cross-Language Evaluation Forum (BirdCLEF) [2] and Detection and Classification of Acoustic Scenes and Events (DCASE) [3].

While deep learning models can achieve excellent predictive performance, such a model still needs an enormous amount of unique data-point to be able to reach said performance. This proves to be hard to satisfy on endangered or endemic bird species, as an inadequate number of data tends to overfit a deep learning model. In this case, transferring trained networks on similar large-scale image recognition datasets such as ImageNet [4] to be repurposed on bird recognition tasks is feasible. Furthermore, utilizing learned low-level and mid-level feature definitions in a transfer-learned model could speed up convergence thus accelerate the model preparation process. In addition, the model needs to be compact both in size and computational resources needed during inference to cover the larger ground with small compute devices. In this regard, we employ a dual-input model that feeds from two popular image representations of acoustic features, namely Mel's spectrogram and Mel Frequency Cepstral Coefficient (MFCC). We used EfficientNet [1] as our model backbone which uses compound scaling that enables efficient CNN architecture scaling, as animal sound datasets may be widely varying in size. Therefore, our strategy can be replicated efficiently on datasets of any size. Using this approach, we achieve an efficient and scalable model that is small both in model size and computational resource which is suitable to be used in conservation scenarios. Our dual-input network achieves 99.47% classification accuracy on our Indonesian scops owl validation set, exceeding classification

Received: February 18th, 2021. Accepted: September 24th, 2021
DOI: 10.15676/ijeei.2021.13.3.3

accuracy of trained-from-scratch Resnet-50 [5] validated with exact set and substantially reduce the model size and computational cost of simple dual-input CNN [6]. While our experiment uses an Indonesian scops owl audio dataset as representative, implementation of our methodology includes but is not limited to owl scops and can be replicated on other birds or animals that have a distinct pitch.

2. Literature Review

So far, there have been numerous studies that show promising results on the subject with various approaches. M. A. Acevedo, C. J. Corrada-Bravo, H. Corrada-Bravo, L. J. Villanueva-Rivera and T. M. Aide [7] suggest machine learning solutions such as linear discriminant analysis, decision tree, and support vector machine to classify numbers of bird and frog species. N. Turpault, R. Serizel, J. Salamon and A. P. Shah [3] suggest using a randomized decision tree that utilizes features derived from audio recordings statistics and ranks those feature importance with a decision tree. Pioneering deep learning solutions detect bird presence by using wild scene imagery [8][9]. While using wild scene imagery is simple and may be beneficial in some scenarios, such imagery may have shortfalls such as occlusion and pose variance which is harder to properly collect on endemic or endangered animals. In this regard, audio data is more popular to be used by the researcher in this scenario. Beforehand, features from audio recordings were manually extracted and analyzed to learn distinct features and information to classify species [10],[11]. However, recent trends show that CNN models that utilized Mel's spectrogram or MFCC derived from audio data dominate BirdCLEF winning solutions for the past year. E. Sprengel, M. Jaggi, Y. Kilcher, and T. Hofmann's solution used 5 convolutional layers network wins BirdCLEF 2016 [12]. C. Koh et al. [13] suggest using popular CNN architecture such as ResNet and Inception, along with Mel's spectrogram as a visual representation of acoustic features. J. Martinsson [14] also suggests that ResNet trained on Mel's spectrogram and MFCC is advisable. M. Lasseck [15] won the BirdCLEF 2018 challenge by improving his previous winning solution on NIPS4B 2013 Competition [16] that utilizes spectrogram derived from audio recordings with InceptionV3 architecture.

3. Methods

Our proposed methodology consists of three stages: audio data preprocessing, model training, and measure trained model predictive accuracy with a testing dataset.

A. Dataset

In this experiment, we used 7 genera of Indonesian scops owl described in [17], which are Rinjani scops owl (Otus jolandae), Sunda scops owl (Otus angelinae), Otus lempiji, Wallace's scops owl (Otus magicus), Sulawesi scops owl (Otus manadensis), Otus reufescens, and Otus silvicola that were obtained from the xeno-canto database. We use 7 genera of endemic Indonesian owl sound recordings as described above to conduct our experiment. The summary of original audio signals and augmented signal counts used from each genus is shown in table 4.

Table 1. Dataset properties used in this experiment

Genus	Signal Count	Augmented Signal Count
Otus Sangelinae	18	3020
Otus Jolandae	89	3020
Otus Lempiji	127	3020
Otus Magicus	13	3020
Otus Manadensis	128	3020
Otus Rufescens	104	3020
Otus Silvicola	54	3020

A total of 21.140 Mel's spectrograms and MFCC pairs are divided into training, validation, and testing set with an 8:1:1 ratio. Therefore, a total of 16.812, 2.114, and 2.114 sets of data are used in the training, validation, and testing set respectively.

B. Data Pre-processing

In our strategy, we treat the audio classification task as an image classification task. Therefore, audio recordings data should be represented in the visual domain to be used in CNN models. In this regard, we use Mel's spectrogram [18] and MFCC [19] are widely used in sound recognition tasks [6], [13], [18], [20]–[23], and can accurately map auditorial features in a visual domain. Temporal information from audio recordings such as timbre and pitch can also be represented in a spectrogram. In this case, each owl species have signature temporal properties that can distinguish one owl species from another. For this reason, mel's spectrogram and MFCC can be used to accurately identify an owl species based on its sound recordings.

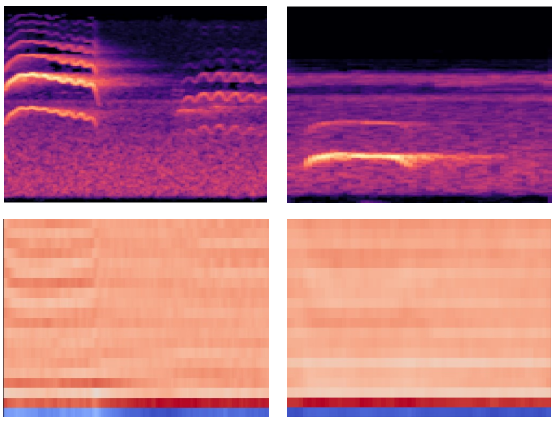

Figure 1. Mel's spectrogram and MFCC sample derived from *otusangelinae* and *otusjolandae*

Mel's spectrogram is computed by applying a Short-Time Fourier Transform with a fourier transform window of 2048 size with 75% overlap between each frame on an audio recording and then converted into the log amplitude scale which considers human auditory sensitivity, producing a log spectrogram with 75 x 128 pixels resolution. An example of generated Mel's spectrogram computed using librosa.feature.melspectrogram function from librosa python package for audio analysis is shown below.

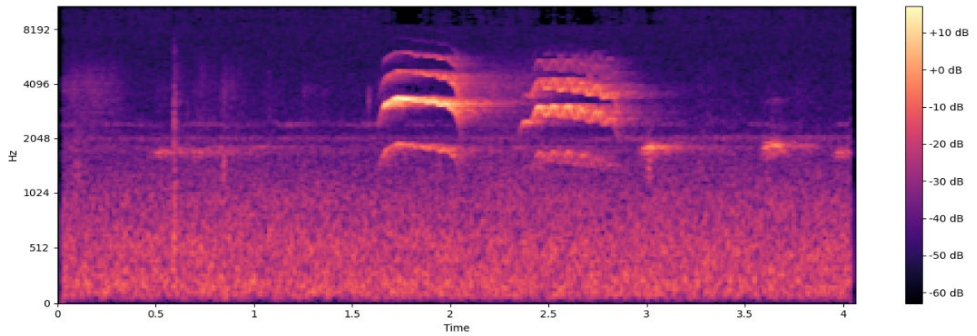

Figure 2. Mel's spectrogram generated from an audio recording

On the other hand, MFCC was also derived from a spectrogram. MFCC can be computed by taking Discrete Cosine Transform (DCT) on Mel's spectrogram using 20 MFCC coefficients, resulting in a compressible visual representation of an audio recording with 75 x 128 resolution. An example of MFCC computed using librosa.feature.mfcc from librosa is shown below.

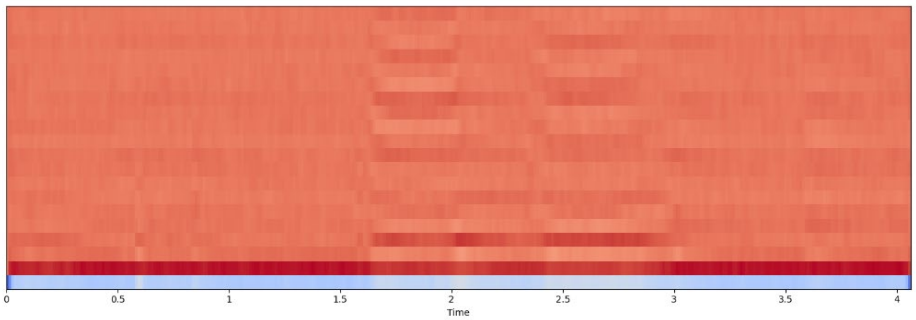

Figure 3. MFCC generated from an audio recording

The lengths of audio recordings obtained from the xeno-canto database vary from few seconds to minutes. We cut and unify each audio signal duration to a five-second clip, which represents a single bird chirp sound. The remaining audio clips that are not used are saved for augmentation purposes. Once unified, the resulting audio chunk is then converted from a compressed mp3 format into a lossless wav format. Each wav file is then re-sampled at 44.1 kHz 16-bit format. To further raise the number of unique data and reduce data inequality among classes, we employ a unique data augmentation strategy on the audio domain, where each audio signal is combined with random background noise derived from remainders of the trimmed audio signal. We prefer this strategy as opposed to other data augmentations that are more commonly used in this scenario {Formatting Citation} as scops owls tend to share natural similar habitat. Therefore, it is more beneficial to mimic the original natural habitat settings than to employ a data augmentation strategy such as high pitch shifting that may accidentally deviate some characteristics unique to a species. In the end, we split preprocessed dataset into train, validation, and test split with an 8:1:1 ratio.

C. Proposed Network Architecture and Training Methodology

We use a pre-trained image classification model that has previously been trained on the ImageNet dataset. Our model takes a pair of Mel's spectrograms and MFCC of a given data point as input data. Therefore, we divide the network into 2 blocks of layers, where the first block (block A) is responsible to learn lower-level features from both Mel's spectrogram and MFCC, while the later block (block B) may extract higher-level features derived from the previous block. The detailed structure of both block A and block B networks are described in the tables below.

Table 2. EfficientNet Block A – Initial feature extraction block

Operator	Output Resolution (C x H x W)	Channels	Layers
Conv 3x3	32 x 64 x 37	32	1
MBConv Block 1 3x3	16 x 64 x 37	16	1
MBConv Block 2 3x3	24 x 32 x 18	24	2
MBConv Block 3 5x5	40 x 16 x 9	40	2
MBConv Block 4 3x3	80 x 8 x 4	80	3

Table 3. EfficientNet Block B – Intermediate feature extraction block

Operator	Output Resolution (C x H x W)	Channels	Layers
MBConv Block 5 5x5	112 x 8 x 4	112	3
MBConv Block 6 5x5	192 x 4 x 2	192	4
MBConv Block 7 3x3	320 x 4 x 2	320	1

Table 4. The complete architecture of our network

Operator	Output Resolution (C x H x W)	Channels
Image input layer	3 x 128 x 75	3
Conv 3x3	32 x 64 x 37	32
EfficientNet Block A	80 x 8 x 4	80
EfficientNet Block B	320 x 4 x 2	320
Dropout	1280 x 1	-
Linear	7 x 1	-

We employ a set of block A to accommodate both Mel's spectrogram and MFCC as our model input, enabling independent feature extraction for both Mel's spectrogram and MFCC.

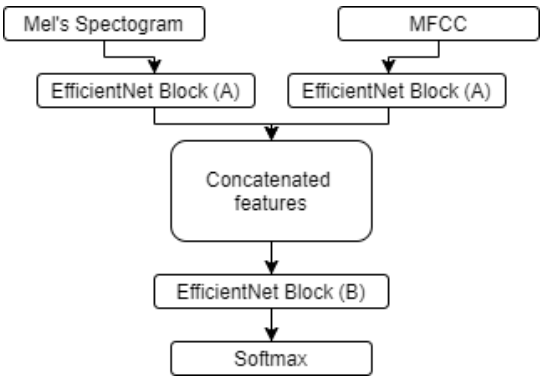

Figure 4. Proposed model architecture

We use efficientNet as our model reference, as the compound scaling used in the efficientNet family allows our model to be scalable and adaptable to an extensive level of data complexity. Specifically, we adapt the most concise efficientNet B0 variant from the efficientNet model family. While having the smallest model size and Floating Operations Per Second (FLOPS) amongst other efficientNet variants, it performed excellently on our dataset. In our case, using heavier and larger efficientNet variants such as B4 or B5 offers no consistent improvement in classification accuracy while having prodigally increased model size and computational burden.

4. Result and Discussion

In our experiment, we trained the baseline and proposed model by feeding the augmented pairs of Mel's spectrograms and MFCC as stated before. As for the single input network, we only feed Mel's spectrogram counterpart. At each end of an epoch, we evaluated its current accuracy on the validation set. Finally, we test the model generalization on the testing dataset. Apart from the model's predictive performance, we also measure epochs needed to achieve 95%

training accuracy to measure its convergence speed. We choose a Resnet-50 and dual input CNN as stated in [6] as our experiment baselines. We trained and identically evaluated the baseline models as our experimental model training method.

To confirm the advantage transfer-learned network as opposed to a network that is learned from scratch, we trained two variants of a modified EfficientNet B0 network that are transfer learned and trained from scratch. The final evaluation of the model is presented in Table 5.

Table 5. Model performance comparisons

Model	mAP	Epochs to reach 95% training accuracy	Weight Size
Resnet-50	99%	3	94,4 MB
Dual-Input CNN	98.49%	1	937,3 MB
Pre-trained dual-input EfficientNet B0	99.27%	3	17,6 MB
Trained-from-scratch dual-input EfficientNet B0	99.06%	5	

From the table, we can infer that while the baseline models have already had a good performance in our research dataset, both models use a substantial amount of disk space to store their model weight, thus cannot be implemented effectively in an environment or device with limited storage capacity and computing power such as micro-controllers. With EfficientNet, our model can be efficiently scaled depending on data complexity and target deployment device storage capacity. In our experiment, model size can be reduced up to 98.2% with EfficientNet B0 while still achieving comparable predictive performance with baseline models.

The experiment result shows that our proposed architecture converges faster due to the transfer-learned weights learned from ImageNet. Our proposed model also uses a substantially smaller weight size compared to baseline models while also surpassing the predictive accuracy of baseline models.

5. Conclusion

Deep-learning-based sound classifications have succeeded to improve the development of automatic identification of endemic and endangered animal species, which may greatly contribute to successful animal conservation efforts. We provided an example by developing a novel methodology to be used for identifying Indonesia scops owl species based on their existing sound recordings. By using a cost-effective and scalable approach, the proposed model can be deployed into low-cost devices, therefore additional devices can be used to cover more grounds. On the other hand, an immense number of quality data of endemic or endangered animals needed to build a performant model is not easy to collect. Therefore, audio recording data is preferable, as audio recordings are free of occlusion and can be augmented more easily. In this experiment, it is shown that a transfer-learned network that has been previously trained on ImageNet shows a better predictive capability and accelerates convergence compared with the same network architecture that is trained from scratch. For that reason, a pre-trained model is beneficial to be used in a situation where the number of quality datasets is limited.

6. References

[1]. M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 10691–10700, 2019.

[2]. S. Kahl *et al.*, “Overview of BIRDCLEF 2019: Large-scale bird recognition in

soundscapes," *CEUR Workshop Proc.*, vol. 2380, 2019.
[3]. N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis," pp. 253–257, 2019, doi: 10.33682/006b-jx26.
[4]. J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
[5]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.
[6]. A. A. Hidayat, T. W. Cenggoro, and B. Pardamean, "Convolutional Neural Networks for Scops Owl Sound Classification," *To Appear Int. Conf. Comput. Sci. Comput. Intell. 2020*, 2020.
[7]. M. A. Acevedo, C. J. Corrada-Bravo, H. Corrada-Bravo, L. J. Villanueva-Rivera, and T. M. Aide, "Automated classification of bird and amphibian calls using machine learning: A comparison of methods," *Ecol. Inform.*, vol. 4, no. 4, pp. 206–214, 2009, doi: 10.1016/j.ecoinf.2009.06.005.
[8]. R. Yoshihashi, R. Kawakami, M. Iida, and T. Naemura, "Bird detection and species classification with time-lapse images around a wind farm: Dataset construction and evaluation," *Wind Energy*, vol. 20, no. 12. pp. 1983–1995, 2017, doi: 10.1002/we.2135.
[9]. S. J. Hong, Y. Han, S. Y. Kim, A. Y. Lee, and G. Kim, "Application of deep-learning methods to bird detection using unmanned aerial vehicle imagery," *Sensors (Switzerland)*, vol. 19, no. 7, pp. 1–16, 2019, doi: 10.3390/s19071651.
[10]. M. F. Kacamarga, T. W. Cenggoro, A. Budiarto, R. Rahutomo, and B. Pardamean, "Analysis of acoustic features in gender identification model for English and Bahasa Indonesia telephone speeches," *Procedia Comput. Sci.*, vol. 157, pp. 199–204, 2019, doi: 10.1016/j.procs.2019.08.158.
[11]. M. Araya-Salas and G. Smith-Vidaurre, "warbleR: an r package to streamline analysis of animal acoustic signals," *Methods Ecol. Evol.*, vol. 8, no. 2, pp. 184–191, 2017, doi: 10.1111/2041-210X.12624.
[12]. E. Sprengel, M. Jaggi, Y. Kilcher, and T. Hofmann, "Audio based bird species identification using deep learning techniques," *CEUR Workshop Proc.*, vol. 1609, pp. 547–559, 2016.
[13]. C. Y. Koh, J. Y. Chang, C. L. Tai, D. Y. Huang, H. H. Hsieh, and Y. W. Liu, "Bird sound classification using convolutional neural networks," *CEUR Workshop Proc.*, vol. 2380, pp. 9–12, 2019.
[14]. J. Martinsson, "Bird Species Identification using Convolutional Neural Networks," University of Gothenburg and Chalmers University of Technology, 2017.
[15]. M. Lasseck, "Audio-based bird species identification with deep convolutional neural networks," *CEUR Workshop Proc.*, vol. 2125, 2018.
[16]. M. Lasseck, "Bird song classification in field recordings: Winning solution for NIPS4B 2013 competition," *Proc. int. symp. Neural Inf. Scaled ...*, pp. 1–6, 2013, [Online]. Available: http://www.animalsoundarchive.org/RefSys/Nips4b2013NotesAndSourceCode/WorkingNotes_Mario_NipsFormat.pdf.
[17]. G. Sangster, B. F. King, P. Verbelen, and C. R. Trainor, "A New Owl Species of the Genus Otus (Aves: Strigidae) from Lombok, Indonesia," *PLoS One*, vol. 8, no. 2, 2013, doi: 10.1371/journal.pone.0053712.
[18]. S. S. Stevens, J. Volkmann, and E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," vol. 8, no. 3, pp. 185–190, 1937.
[19]. L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," vol. 2, no. 3, pp. 138–143, 2010, [Online]. Available:

http://arxiv.org/abs/1003.4083.
[20]. E. Cakir, S. Adavanne, G. Parascandolo, K. Drossos, and T. Virtanen, "Convolutional recurrent neural networks for bird audio detection," *25th Eur. Signal Process. Conf. EUSIPCO 2017*, vol. 2017-Janua, pp. 1744–1748, 2017, doi: 10.23919/EUSIPCO.2017.8081508.
[21]. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2818–2826, 2016, doi: 10.1109/CVPR.2016.308.
[22]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, vol. 7, no. 3, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[23]. M. Lasseck, "Acoustic bird detection with deep convolutional neural networks," *Classif. Acoust. Scenes Events 2018 Work.*, pp. 143–147, 2018.

Kevin William Gunawan received the bachelor's degree in computer science from Bina Nusantara University. He is currently a computer science master's degree student at Bina Nusantara University. He has previously published a conference paper on an international conference on computer science and computational intelligence in computer vision and deep learning field.

Alam Ahmad Hidayat received the B.Sc. degree in physics from the Institut Teknologi Bandung, Indonesia, in 2014, and the M.Sc. degree in theoretical physics with the University of Bonn, Germany, in 2018. He is currently a Research Assistant with the Bioinformatics and Data Science Research Center (BDSRC), Bina Nusantara University, Indonesia. His research interests include applications of deep learning and statistical models to analyze data from diverse topics, including health sciences

Tjeng Wawan Cenggoro received the bachelor's degree in information technology from STMIK Widya Cipta Dharma, and the master's degree in information technology from Bina Nusantara University. He is currently an AI Researcher whose focus is in the development of deep learning algorithms for application in computer vision, natural language processing, and bioinformatics. He is also a Certified Instructor with the NVIDIA Deep Learning Institute. He led several research projects that utilize deep learning for computer vision, which is applied to indoor video analytics and plant phenotyping. He has published over 20 peer-reviewed publications and reviewed for prestigious journals, such as Scientific Reports and IEEE ACCESS. He also holds two copyrights for AI-based video analytics software.

Bens Pardamean received the bachelor's degree in computer science and the master's degree in computer education from California State University, Los Angeles, and the Doctoral degree in informative research from the University of Southern California (USC). He has over 30 years of global experience in information technology, bioinformatics, and education. After successfully leading the Bioinformatics Research Interest Group, he currently holds a dual appointment as the Director of the Bioinformatics and Data Science Research Center (BDSRC), and an Associate Professor of computer science with Bina Nusantara University, Jakarta, Indonesia