

# Classification of dances using AlexNet, ResNet18 and SqueezeNet1\_0

Khalif Amir Zakry, Irwandi Hipiny, Hamimah Ujir

Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia

## Article Info

### Article history:

Received Apr 5, 2022

Revised Oct 16, 2022

Accepted Nov 15, 2022

### Keywords:

Dance classification

Deep learning

## ABSTRACT

Dancing is an art form of creative expression that is based on movement. Dancing comprises varying styles, pacing and composition to convey an artist's expression. Thus, the classification of any dance to a certain genre or type depends on how accurate or similar it is to what is generally understood to be the specific movements of that dance type. This presents a problem for new dancers to assess if the dance movements that they have just learned is accurate or not to what the original dance type is. This paper proposed that deep learning models can classify dance videos of amateur dancers according to the similar movements of actions of several dance classes. For this study, AlexNet, ResNet and SqueezeNet models was used to perform training on multiple frames of actions of several dance videos for label prediction and the classification accuracy of the models during each training epoch is compared. This study observed that the average classification accuracy of the deep learning models is 94.9669% and is comparable to other approaches used for dance classifications.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Khalif Amir Zakry

Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak

Jalan Datuk Mohammad Musa, 94300 Kota Samarahan, Sarawak, Malaysia

Email: 21020463@siswa.unimas.my

## 1. INTRODUCTION

Dancing is a type of performative art that is based on movement and actions of the performers. Professional dancers in the field of performative arts can identify and distinguish between several different types of dances based on their professional experience. Amateur dancers on the other hand, may have difficulty in distinguishing a type of dance due to the wide variety of dances that are present in the world of performance art. Ballet for example, can easily be assessed by a professional ballerina on how accurate one's performance is to be a representation of that dance type, while an amateur would find it difficult. On another note, deep learning models have been used to classify human actions and movements [1]. Dancing however, unlike previous work that has been done to classify human actions; is slightly different because there exists a metric of how accurate the specific movement of the dancer is to what is generally considered to be a dance style of a specific dance type [2]. Therefore, it is difficult for a non-expert to assess how accurate a dance is to its generally understood dance interpretation.

This paper introduces using deep learning models to classify dances according to several dancers' general interpretation of a class of dance and how accurate the dances are to the deep learning models' understanding of that dance class. We proposed using AlexNet [3], ResNet18 [4] and SqueezeNet1\_0 [5] deep learning models to classify and evaluate the accuracy of dances to its class. Several works have been done to classify human movements such as proposed by Yildirim and Çinar [6], Kumar and Harikiran [7] and Zamri *et al.* [8] that uses deep learning models. However, those works are similar in method whereby the authors utilized singular images of

a human performing an action to train their deep learning models. This method generalises the action in a video into an image classification problem [9]–[11]. Alternative methods for human action recognition by utilizing temporal gradients of action have also been proposed by Hutchison *et al.* [12]. The authors demonstrated that by using temporal structures of motion segments in an activity for action recognition they can achieve an average precision of 72.1% for classification of a sports action dataset [12].

We proposed that instead of using singular images to classify an action, we should use multiple frames of actions in an action sequence to classify dances. Such approach for identifying the quality of other human actions have been demonstrated [14], [15]. Utilizing deep learning models with dance videos has also been performed by several authors. Wang *et al.* [16] proposed the use of a dataset of viral dance videos to predict the virality of a dance video. The authors introduced a relational temporal convolutional network (RTCNN) for performing viral predictions of a dance video by incorporating the capture of temporal dynamics from the appearance of the video. Based on their study, factors such as facial, scenic, and holistic appearance of a dance video is considered an important aspect that in virality prediction of a dance video [16]. In our method, we adapt a different approach from the mentioned works by utilizing a dance video dataset with deep learning models to incorporate multiple frames of a dance video to classify the types of dances and for measuring the accuracy of the dance video to how similar the dance video is to others of its type.

## 2. METHOD

In our proposed method, we used our own dance video dataset with FastAI, a machine learning framework that provides high-level and low-level components for classifying multi-modal datasets [17]. The dataset consists of dance videos that have been obtained from TikTok which is an online public social media platform and from our research participants that were tasked to perform their interpretation of the dance type that we have labelled. The dataset consists of 240 videos and 12 different labelled dance classes, each class representing a single type of dance. Since each video may only attribute to a single class; this means that no video may be multi-classed. 20 human participants were asked to perform each of the 12 types of dances. A reference video of each dance was provided, and the participants were asked to perform their own interpretation of the dance based on the reference video. The dance videos produced are of varying quality with slight variations in color grading, background, and image quality. Several of the videos also include the application of TikTok video filters which drastically changes the visual scene and the visual fidelity of the video by the introduction of noise, as shown in Figure 1 and Figure 2.

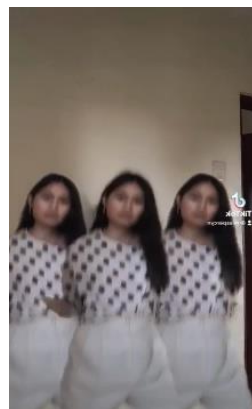


Figure 1. Sample Image from a video with “Clone filter” applied



Figure 2. Sample Image from a video with “Disco filter” applied

For our deep learning classification experiment with FastAI, a set of frames of one-second intervals of each second in a video is extracted by using fast forward moving picture experts group (FFMPEG), an open-source multimedia framework [18]. This generated a total of 17,389 images from the videos that can be categorized into the respective dance class of the original video that the image was extracted from. Using a combination of automatic tool and manual review, images that contain noise or without any subject context such as empty frames were filtered and discarded. Each dance class has different amounts of images due to the difference in video length and due to the automatic and manual removal of frames. See Table 1 for the final number of frames per label category.

Table 1. The number of frames that has been extracted for each label category

Label Category	Number of Frames
All TikTok Mashup	1,490
Big Up's	1,457
Blinding Lights	1,197
<i>Diam Diam Menyukaiku</i>	1,320
Laxed	1,461
Lottery	1,525
Say So	1,138
Slide To The Left	1,395
Supalonely	1,516
<i>Tak Mau Mau</i>	1,684
The Dance Song	1,892
Tokyo	1,314

This extraction method when used provided us with numerous image frames that represents the various action frames of the participant from every dance video in their specific dance class. For example, there would be 1,138 frames of dancing actions that can be derived from the 20 videos in the “Say So” dance class. Sample frames are shown in Figure 3. We then experimented training with the deep learning models; AlexNet, Squeezenet1\_0 and ResNet18 Before training is carried out for each pre-trained model, 20% of the total frames is separated as the validation set while the remaining 80% is used as the training set. To standardize, the training epoch for each pre-trained model is set to a uniform value of 10.

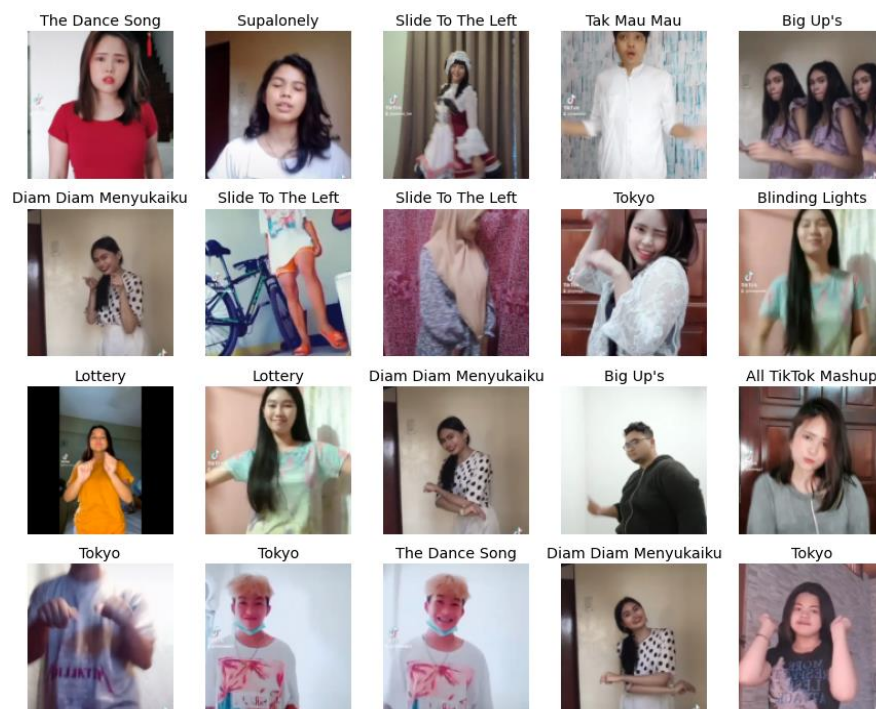


Figure 3. Each image represents a single frame that has been extracted from class of dance

### 3. RESULTS AND DISCUSSION

#### 3.1. Discussion of training results with AlexNet, Squeezenet1\_0 and ResNet18

The line plots shown in Figure 4 show a consistent increase in accuracy for every consecutive epoch of the trained model when predicting using the validation set for Squeezenet1\_0 and ResNet18. This is distinct however from the results for AlexNet. There is a 0.0575% drop in predictive accuracy with the validation set between the 9<sup>th</sup> and 10<sup>th</sup> training epoch for AlexNet. This can be attributed overfitting of the trained model that occurs due the 9<sup>th</sup> training epoch with AlexNet. An indicator of overfitting can be seen when the error rate goes up as the model is by-hearting the data [19].

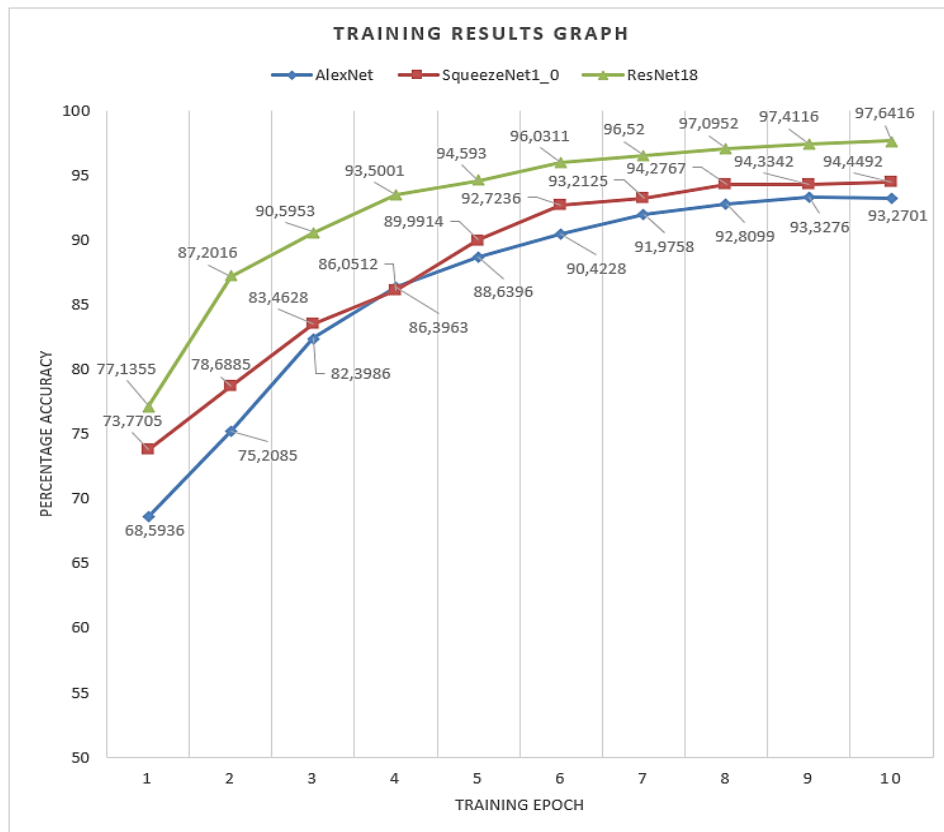


Figure 4. Training result graph

By analysing the total findings, we can consider that the trained model with ResNet18 provides the best final accuracy when validating with its validation set at 97.6416%. The average accuracy of the final epoch for all 3 training models is 94.9669% which is good for predicting the types of dances for image frames in their respective validation sets as experimented. However, as mentioned in [20], these 3 trained models might have difficulty in predicting a similar dataset if they were pre-processed differently before analyses. As mentioned in [21], even though accuracy is a widely used performance metric for evaluating classification models, it might not be the most suitable for some cases. As such we can also utilize confusion matrix another evaluation metric for this experiment.

In [22] a confusion matrix is performance matrix that can summarize the performance of a classifier with respect to some test data. From our experiment, the confusion matrix for all three models, as shown in Figures 5-7, also shows a generally good outcome whereby the ratio of correct predictions far outnumbers the incorrect predictions made by each model with its validation set. We can see this via the visible dark blue diagonal colour contrast from the top left to the bottom right of each confusion matrix. This indicates that the predictive capability of the model when used to predict its training validation set is high.

For AlexNet, the highest occurrence of misclassification is between Supalonely and Say So. For SqueezeNet 1\_0, the highest occurrence of misclassification occurs among Big Up's, All TikTok Mashup, Say So, Lottery and The Dance Song. For ResNet18, the highest occurrence of misclassification is between *Diam Diam Menyukaiku* and *Tak Mau Mau*. We theorise the cause of these misclassifications is due to these dances sharing similar dance motions hence making the frames to be visually similar. Most of the participants had captured the videos for each dance class using the same clothing and background hence motions would be an important information to discriminate between classes. Based off the results of the confusion matrices of all 3 deep learning models, it shows that the models can be used to classify dances. They can classify with a high accuracy what class of dance the image in the validation layer is based on what it has learned from the input layer. This method can be used to create further programmes and applications that enables users such as amateur dancers to assess the accuracy of their dance performance according to the general interpretation by other dances in that dance class.

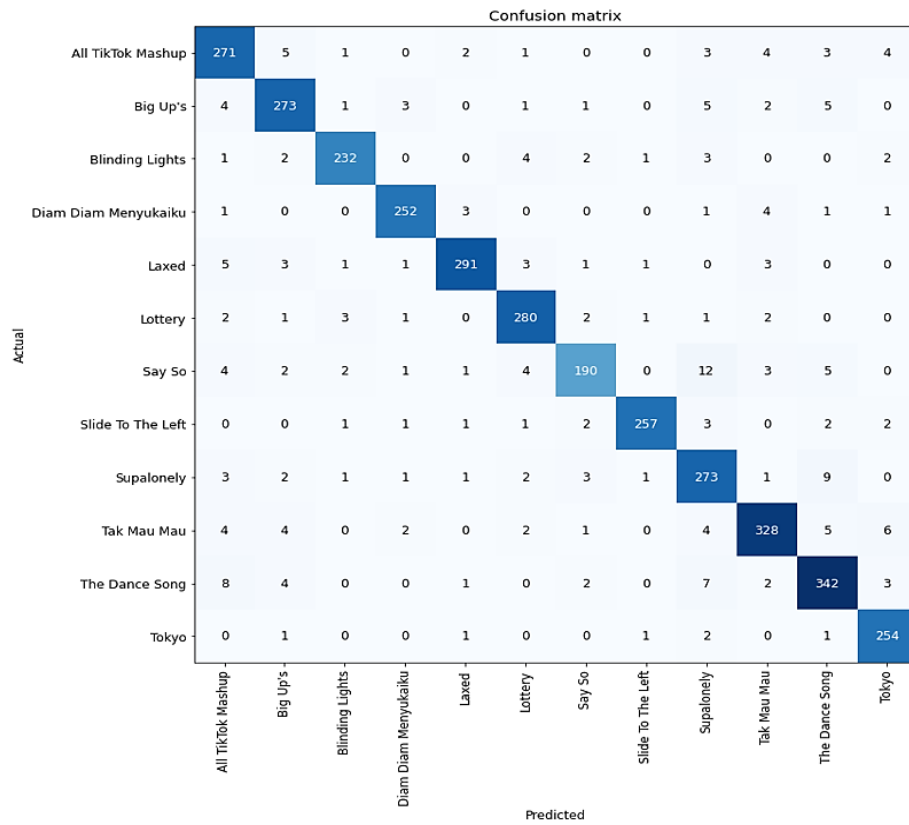


Figure 5. Confusion matrix with AlexNet

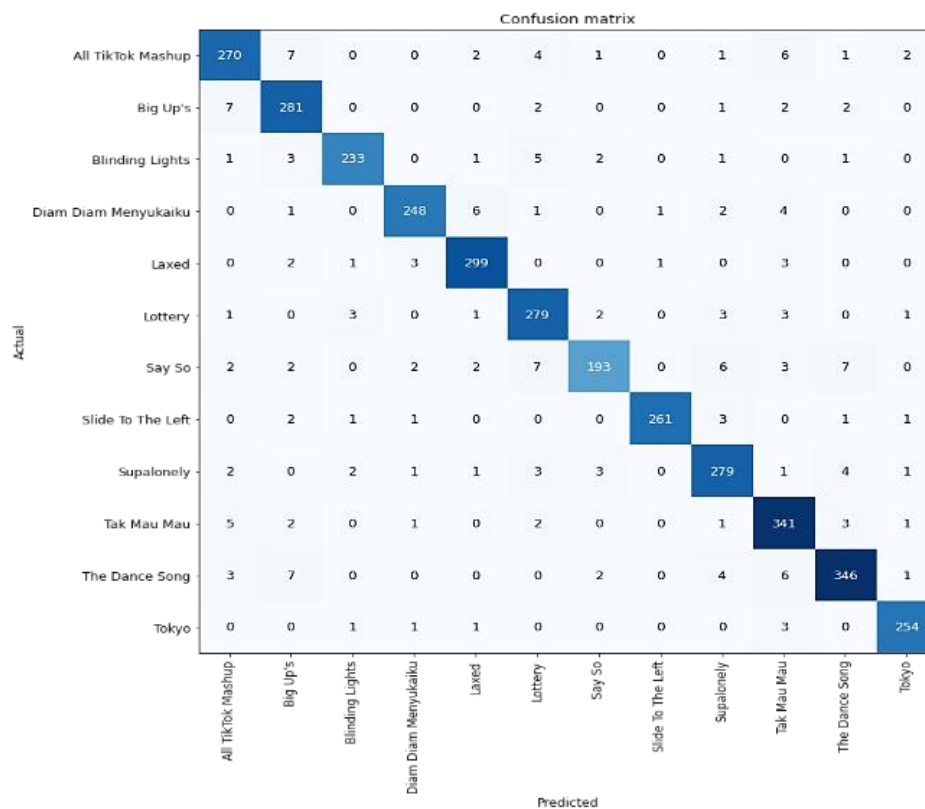


Figure 6. Confusion matrix with Squeezenet1\_0

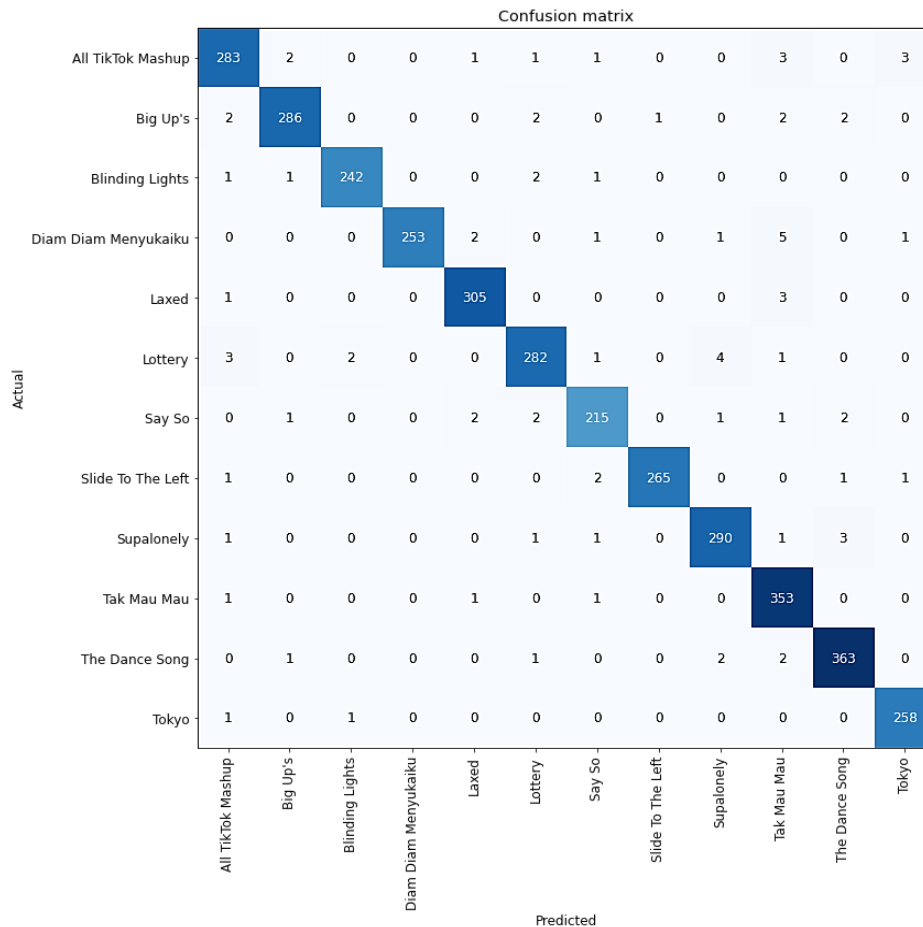


Figure 7. Confusion matrix with ResNet18

### 3.2. Discussion comparing our results with other works done for classifying dances

Kishore *et al.* has also presented a similar work in dance classification by using their own convolutional neural network (CNN) to classify various indial classical dances [23]. The performance of their CNN for classifying Indian dances with different settings of convolutional filter sizes can be compared to our approach of using pre-trained models for classifying modern dances. Kishore *et al.* [23] managed to achieve a 93.33% average accuracy by using their own proposed deep learning CNN model to classify their dance dataset which is only a 1.6469% difference in accuracy than our 94.9669% average accuracy while using AlexNet, ResNet and SqueezeNet1\_0 for our dataset. The difference in approach that is taken between ours and Kishore *et al.* does prove and support the evidence that deep learning models can produce a good average accuracy for classifying dances albeit Kishore *et al.* dataset consists of a different type of dance from ours.

We can also compare our results with the approach proposed by Li [24]. They utilized a deep CNN model based on differential evolution for their approach to classify dances. The accuracy of their DE-CNN approach to classify their dances achieved an average accuracy of 92.75%. This is a 2.2169 difference in accuracy from our approach. This comparison indicates the viability of our approach in being able to achieve marginally better classification accuracy for classifying dances. However, it must be noted that the dataset of dances used in the other works are different.

### 3.3. Discussion on the automatic removal of extremely noisy frames inside the TikTok dataset

Using a tool provided with the FastAI framework, we can identify frames with a high loss based on the outcome of our training. This is how we locate extremely noisy frames for possible manual removal. See Figure 8 for some sample frames with high loss values.

We identified an occurring pattern whereby the top 9 frames with the highest loss for one model can also reappear for the others. For example, Frame A is the frame with the highest loss for the model trained with AlexNet, but it is also coincidentally the frame with the highest loss for the model trained with Squeezenet1\_0 as well. We can consider that the high loss for Frame A can be attributed with the highly



noisy image quality of the frame. These frames must be removed since even when using human judgement, it is impossible to ascertain the dance class for Frame A, Frame B and Frame C whereby little to no useful characteristic that can be used to identify a dance is being shown. Several actions can be taken when dealing with noisy data, in [25] techniques such as ignoring the noise, and filtering noise by removing was used to handle noise. As demonstrated in [26], an increase in noise can reduce the accuracy of classification. As such, noise can and does impact the performance of a machine learning model. We observed that there are several frames that could be excluded for training purposes in this dataset as those frames provide very little relevance to the classification purposes of a dance video. These frames are duly removed from the dataset.

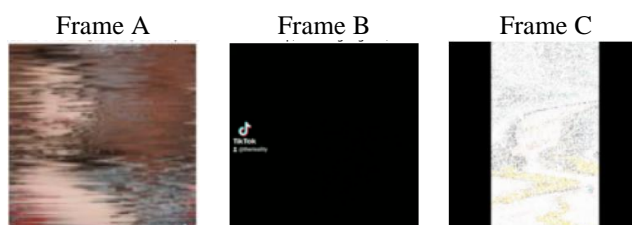


Figure 8. Three sample frames with high loss

#### 4. CONCLUSION

In conclusion, we proposed and experimented the use of deep learning approach for classifying dances using deep learning CNN models to classify dances according to how accurate the dance is with other similarly labelled dances. The results of our approach demonstrated that deep learning models can be used to classify dances and the average accuracy of our approach with 3 different deep learning models is 94.969%. We have also identified potential areas of improvement regarding the utilization of video frames for deep learning applications such as the removal of noisy frames in a dataset. To end, we would like to thank our research participants for their contributions in our research and would also like to acknowledge that this work was funded by the Universiti Malaysia Sarawak's internal research grant; UNIMAS CDRG (F08/CDRG/1820/2019).





#### REFERENCES

- [1] M. S. Srividya, M. R. Anala, and C. Tayal, "Deep learning techniques for physical abuse detection," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 4, pp. 971–981, 2021, doi: 10.11591/IJAI.V10.I4.PP971-981.
- [2] A. L. Kaeppler, "Dance and the Concept of Style," *Yearbook for Traditional Music*, vol. 33, p. 49, 2001, doi: 10.2307/1519630.
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.
- [4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.
- [5] Y. Liu, Z. Li, X. Chen, G. Gong, and H. Lu, "Improving the accuracy of SqueezeNet with negligible extra computational cost," *2020 International Conference on High Performance Big Data and Intelligent Systems, HPBD and IS 2020*, 2020, doi: 10.1109/HPBDIS49115.2020.9130577.
- [6] M. Yildirim and A. Çinar, "Classification of 40 Different Human Movements with CNN Architectures and Comparison of Their Performance," *Turkish Journal of Science & Technology*, vol. 16, no. 1, pp. 103–112, 2021.
- [7] K. V. Kumar and J. Harikiran, "Privacy preserving human activity recognition framework using an optimized prediction algorithm," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 254–264, 2022, doi: 10.11591/ijai.v11.i1.pp254-264.
- [8] N. N. M. Zamri, G. F. Ling, P. Y. Han, and O. S. Yin, "Vision-based Human Action Recognition on Pre-trained AlexNet," *Proceedings-9th IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2019*, pp. 1–5, 2019, doi: 10.1109/ICCSCE47578.2019.9068586.
- [9] W. Yang, Y. Wang, and G. Mori, "Recognizing human actions from still images with latent poses," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2030–2037, 2010, doi: 10.1109/CVPR.2010.5539879.
- [10] N. Ikizler-Cinbis, R. G. Cinbis, and S. Sclaroff, "Learning actions from the web," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 995–1002, 2009, doi: 10.1109/ICCV.2009.5459368.
- [11] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: A study of bag-of-features and part-based representations," *British Machine Vision Conference, BMVC 2010 - Proceedings*, 2010, doi: 10.5244/C.24.97.
- [12] D. Hutchison et al., "Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification", in *Computer Vision – ECCV 2010*, vol. 6312, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 392–405. doi: 10.1007/978-3-642-15552-9\_29.
- [13] J. C. Niebles, C. W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6312 LNCS, no. PART 2, pp. 392–405, 2010, doi: 10.1007/978-3-642-15552-9\_29.





- [14] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8694 LNCS, no. PART 6, pp. 556–571, 2014, doi: 10.1007/978-3-319-10599-4\_36.
- [15] Y. Xu, "A Sports Training Video Classification Model Based on Deep Learning," *Scientific Programming*, vol. 2021, 2021, doi: 10.1155/2021/7252896.
- [16] J. Wang *et al.*, "Will You Ever Become Popular? Learning to Predict Virality of Dance Clips," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 18, no. 2, 2022, doi: 10.1145/3477533.
- [17] J. Howard and S. Gugger, "Fastai: A layered api for deep learning," *Information (Switzerland)*, vol. 11, no. 2, 2020, doi: 10.3390/info11020108.
- [18] X. Lei, X. Jiang, and C. Wang, "Design and implementation of a real-time video stream analysis system based on FFMPEG," *Proceedings-2013 4th World Congress on Software Engineering, WCSE 2013*, pp. 212–216, 2013, doi: 10.1109/WCSE.2013.38.
- [19] S. Afaq and S. Rao, "Significance Of Epochs On Training A Neural Network," *International Journal of Scientific & Technology Research*, vol. 9, no. 6, pp. 485–488, 2020, [Online]. Available: [www.ijstr.org](http://www.ijstr.org).
- [20] K. Aftarczuk, "Evaluation of selected datamining algorithm in medical decision support systems," 2007.
- [21] M. Vakili, M. Ghamsari, and M. Rezaei, "Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification," 2020, [Online]. Available: <http://arxiv.org/abs/2001.09636>.
- [22] T. R. Shultz *et al.*, "Confusion Matrix," *Encyclopedia of Machine Learning*, pp. 209–209, 2011, doi: 10.1007/978-0-387-30164-8\_157.
- [23] P. V. V. Kishore *et al.*, "Indian Classical Dance Action Identification and Classification with Convolutional Neural Networks," *Advances in Multimedia*, vol. 2018, 2018, doi: 10.1155/2018/5141402.
- [24] L. Li, "Dance Art Scene Classification Based on Convolutional Neural Networks," *Scientific Programming*, vol. 2022, 2022, doi: 10.1155/2022/6355959.
- [25] S. Gupta and A. Gupta, "Dealing with noise problem in machine learning data-sets: A systematic review," *Procedia Computer Science*, vol. 161, pp. 466–474, 2019, doi: 10.1016/j.procs.2019.11.146.
- [26] A. T. Saseendran, L. Setia, V. Chhabria, and D. Chakraborty, "Impact of noise in dataset on machine learning algorithms," *Impact of Noise in Dataset on Machine Learning Algorithms*, 2019, doi: 10.13140/RG.2.2.25669.91369.

## BIOGRAPHIES OF AUTHORS







**Khalif Amir Zakry**     holds a BSc (2020) in Computer Science (Software Engineering) from Universiti Malaysia Sarawak (UNIMAS) with his final year thesis titled "Contract Distribution System Framework with Case Study on PC Building Market". He was a system engineer for Sarawak Information Systems, working specifically in field of audiovisual engineering. He is currently pursuing his master education at the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak (UNIMAS). His research areas of interest include Artificial Intelligence and Computer Vision. He can be contacted at email: [21020463@siswa.unimas.my](mailto:21020463@siswa.unimas.my).



**Dr. Irwandi Hipni Mohamad Hipiny**     is currently a Senior Lecturer (DS52) at Universiti Malaysia Sarawak. He holds a BSc (2003) and an MSc (2007) in Computer Science from Universiti Teknologi Malaysia; later graduated with a PhD (2014) in Computer Vision from University of Bristol, UK. Prior to his current appointment, Irwandi was a recipient of Ekpress UTM-MARA scholarship (1999-2003), and a MOSTI's National Science fellow (2003-2007)). Irwandi's current research interests are, but not limited to, Computer vision, Pattern recognition and Animal Re-ID. He can be contacted at email: [mhihipni@unimas.my](mailto:mhihipni@unimas.my).



**Ts. Dr. Hamimah Ujir**     is currently a senior lecturer at the Faculty of Computer Science and Information Technology (FCSIT), Universiti Malaysia Sarawak (UNIMAS). She completed her master's degree (by research) at the Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, Skudai, Johor and obtained her Ph.D. in the School of Electronic, Electrical and Computer Engineering, University of Birmingham, United Kingdom. Her works include 3D physical simulation and 3D static and dynamic facial expression analysis. She also conducted research works on academic quality in higher education. She can be contacted at email: [uhamimah@unimas.my](mailto:uhamimah@unimas.my).