International Journal of e-Collaboration
Volume 19 • Issue 1

Virtual Teaching Assistant for Capturing Facial and Pose Landmarks of the Students in the Classroom Using Deep Learning

Samer Rihawi, Faculty of Art, Computing, and Creative Industry, Universiti Pendidikan Sultan Idris, Malaysia*
Samar Mouti, Al Khawarizmi International College, UAE
Roznim Mohamed Rasli, Faculty of Art, Computing, and Creative Industry, Universiti Pendidikan Sultan Idris, Malaysia
Shamsul Ariffin, Faculty of Art, Computing, and Creative Industry, Universiti Pendidikan Sultan Idris, Malaysia
https://orcid.org/0000-0001-6266-6797

ABSTRACT

This research focuses on the learning challenges that both students and teachers face during the learning process. It addresses the different techniques and methods used for face recognition. The proposed VTA model uses the convolutional neural networks to recognize the identities of the student. It gathers the facial expressions and body poses of each student in the classroom and predicts the attention level of that student, thus determining his/her learning capabilities. This research will help the students achieve their learning objectives by being able to get an accurate and real evaluation of their contribution and attention during the classes. Also, the proposed VTA model helps the teacher get some insight into his/her teaching methodologies during the class as the model will observe and record the attentiveness of the students. This research will have a significant positive impact on student success and on effective lecturing.

KEYWORDS

Artificial Intelligence, *Convolutional Neural Networks*, *Deep Learning*, *Facial Expressions*, *Virtual Assistant*

INTRODUCTION

In schools and colleges, the teachers find it hard to accommodate all the students, overcome language barriers and ensure that the students follow along. Also, the students' attitude toward learning, where many students come to the institution supposedly to learn and gain knowledge. However, this might not be the case all the time due to the emergence of smart devices that the students tend to use instead of focusing on the teacher or simply daydreaming during the classes. Students' responses to course feedback questionnaires that the institution sends at the end of each semester as surveys via emails or online forms to ask them to put their feedback about the courses are a part of the data required

DOI: 10.4018/IJeC.316663 *Corresponding Author

to ensure the quality assurance and standardization in the course delivery, material, and teaching methods. This research automats one of the tools of students' experience to ensure that all the students have an exceptional and distinctive experience while at the college by building a novel VIRTUAL TEACHING ASSISTANT (VTA) model for Capturing Facial and Pose Landmarks of the Students in the Classroom Based on the Deep Learning.

LITERATURE REVIEW

Presently, many methods have been proposed to cater to online learning. To name some include blended learning, flipped learning, virtual and augmented reality, face recognition, gesture detection, chatbot assistance, and so forth. This agrees with the assertion that focuses on the ability of adoption to and smooth accommodation of various types of learners or students in on-campus classrooms and for students with remote or online classes (Bakken et al., 2020). Artificial Intelligence has already been applied to education primarily as a tool that helps develop skills and testing systems and can help fill need gaps in learning and teaching and allow schools and teachers to do more than ever before. Classroom discipline and management have taken a quantum leap in the past century away from the traditional model. The purpose of implementing classroom management strategies is to enhance prosocial behavior and increase student academic engagement (Dahlgren, n.d.). Educational data mining studies have been implemented to analyze student performance and prediction in classroom learning. Predictive modeling falls under AI, which can be used to accommodate all kinds of students. It is flexible enough to help all the students regardless of their learning speeds. It helps the learners move ahead only after they fully understand and grasp all the information they need. It analyzes the information teachers are using to determine whether the quality of the content meets the expected standards. Additionally, it also helps teachers monitor the progress of students on a personal level, thereby suggesting the best ways of teaching (Khan & Ghosh, 2021)(Romero & Ventura, 2013). However, it is also important for the teachers to be able to monitor the students' performance during class times and to be able to know whether they are understanding what the teachers are saying or they are just "daydreaming" as it can help them to or improve or possibly change their teaching techniques and methods, so to attain this objective, the teachers need some tools that can help them monitor the students' interaction during classes by watching their face and body movements and predicting their understanding levels, such as predictive models (Yang, 2000). Machine Learning and Predictive Modeling provide solutions to organizations worldwide for their own needs. However, predictive models with face detection and body gestures can be generated by extracting some dynamic visual processes, as in sign language recognition, where the movements of hands and body can give different meanings (Priya Pedamkar, 2020).

Machine learning has been used in the security field as well such as malware detection, in which a novel monte-carlo simulation-based model was used (Naveed et al., 2020). It uses a simulation based model called Heuristic-based Generative model that generalizes the attack patterns and then predicts any new unknown attacked and then detects and flags them in real-time with a high accuracy.

Methods from the field of machine learning have been implemented in the medical sector, such as tracking several tasks in medical imaging, starting from image reconstruction or processing to predictive modeling, clinical planning, and decision-aid systems (Hatt et al., 2019). Image processing techniques along with predictive models have been used in many applications in different fields, where recent advances in machine learning (ML) are revolutionizing computational approaches by providing principled approaches to feature extraction methods with improved optimization algorithms. For example, DyBM model was applied to human handwriting motion tracking with a UR-5 robot and the results show that the framework significantly improves tracking performance (Kamani et al., 2017)(Agravante et al., 2018). A novel approach for face spoof detection was presented. The novel lay in distinct features derived from scatter and variance measures on the HSI color space. The volumetric measures around the convex hull and geometric description have yielded a compact and effective feature (Nagabhushan, Singh, & Roy, 2017). Face and body detection has been used in education to communicate with the

students to help them overcome their passive attitudes (Azeez & Azeez, 2018). Machine learning (ML) was introduced in the 1980s, it is the study of computer algorithms that improve automatically through experience and by the use of data to give the computers the ability to act intelligently (Mitchell, 1997) (Hutter, 2019). It is seen as a part of Artificial Intelligence (AI). Machine learning algorithms build a model based on sample data, known as "training data" (Kubat, 2017). However, deep learning can be defined as using neural networks to train models. Neural networks consist of multiple layers that can learn from the training data, making it better than the ordinary machine learning process that makes it useful in operations that require powerful computing such as image and video recognition (Marr & Ward, 2019)(Ng, 2018)(Raschka et al., 2020). CNN is a well-known type of Neural Networks, and it is used for image classification. It takes an input image and extracts its features by dividing it into matrices called convolutions and then using the convolutional layers to filter them and then generate a feature map that contains numbers for each pixel of the image, and it can be used later for image classification (Campesato, 2020)(Aggarwal, 2018). A convolutional layer is composed of many independent filters that operate on its input. Those layers slide through all pixels from the entire input image and then from an activation map from which the most relevant regions of the image are extracted, and then the output of each filter is sent to the next layer. The second type of layer is the pooling layers which aggregate information they receive from the filter layers. They shrink the image dimension to a predetermined value by replacing all the information presented into one pool with a single value, mostly by its maximum or average (Cinelli et al., 2018). The figure (1) below describes the general design of the CNN.

Giant computer companies such as Facebook and Google had their own contributions to the field of Artificial Intelligence and Deep Learning by introducing new products and devices such as simulating the camera movements which are currently used in the Google Photos app (Bataeva, 2021), and hand-tracking using the VR device Oculus to improve hand tracking (Wang, 2019).

Face recognition has been used in a wide range of applications using different face recognition methods. CCTV cameras have been installed in different places like shops, malls, and factories to protect against theft and trespassing. Also, cheaper devices such as Raspberry PI have been used along with camera modules and PIR sensors as more feasible alternatives due to lower their power consumption (Hazim Barnouti et al., 2016)(Zakaria, 2017). The technology used is Raspberry PI along with the PI Camera and sensors, which is good for detecting motions. However, it is still impossible to exactly know the meaning of those motions or their sources since no neural networks were used.

HAAR classifier is known as one of the common algorithms for face detection as it uses rectangular features to detect all the angles of the face as shown in the figure (2) below

It uses those features to detect the faces in an image along with the different parts of the face such as eyes, eye brows, nose, and mouth (figure 3) (Mittal, 2020).

Figure 1. Design of CNN (Source: Cinelli et al. 2018)

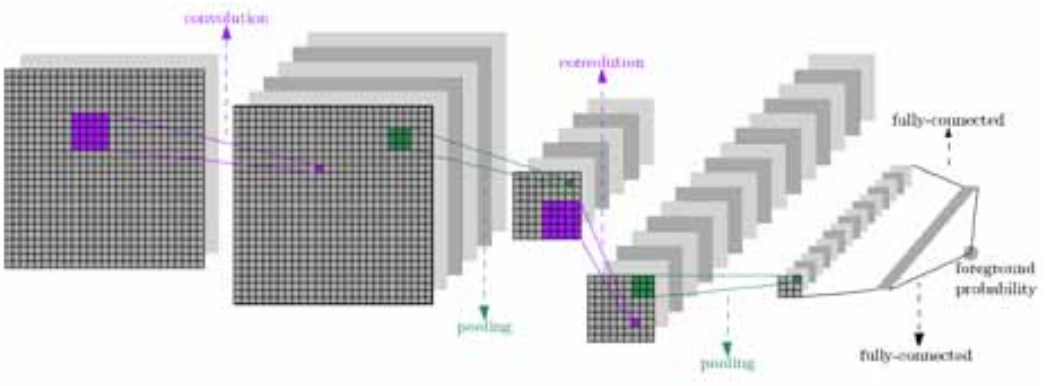


Figure 2. Edge, Line, and Rectangular features detected using HAAR object detector (Source: Mittal 2020)

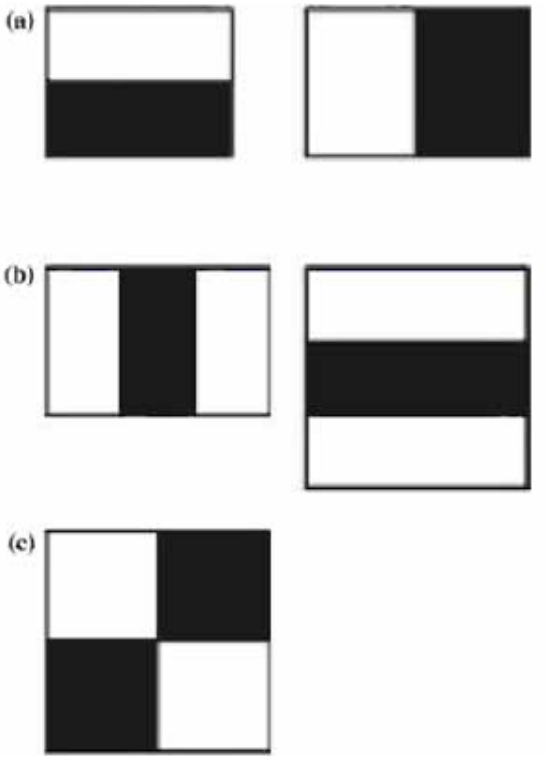
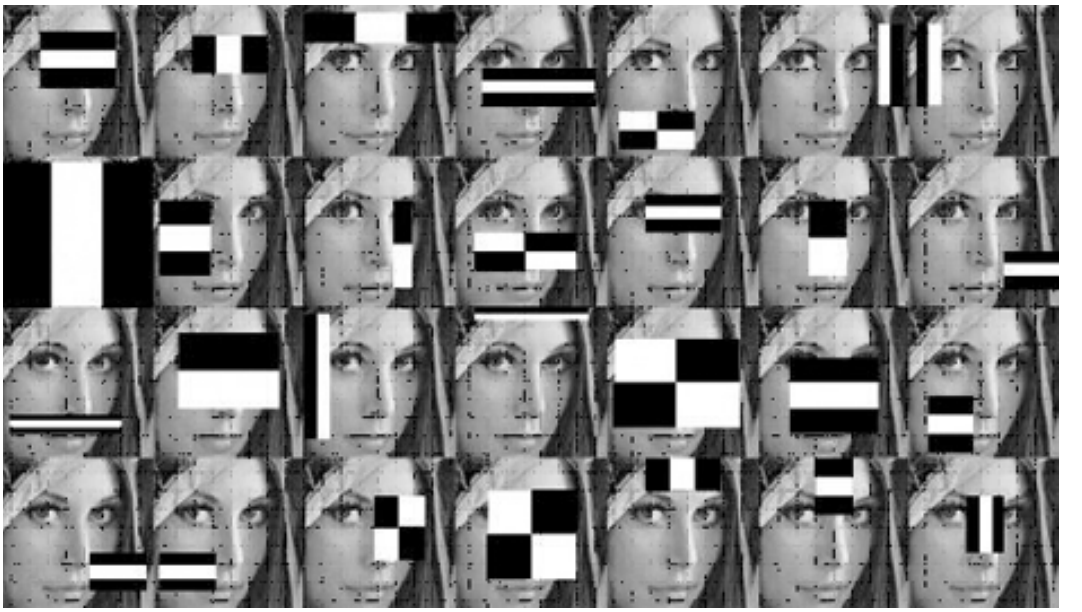


Figure 3. Face Detection using the HAAR classifier (Source: Mittal 2020)



Another study has been conducted to detect the drowsiness of drivers using the Raspberry PI and HAAR Cascade Classifier by using the HAAR classifier to detect the faces and then detecting the eyes and calculating the Eyes Aspect Ratio (EAR) to determine whether the driver was sleepy or not. This research relies heavily on the face detection but not its pose as sleepy drivers tend to hang their heads down, thus their eyes might not be detected. (Kamarudin et al., 2019)

Also, deep learning has been used in Face Detection Systems such as the detection of attendance of students, where the attendance of the students can be marked automatically without interference from the teacher (Fuzail et al., 2014). In this system, the faces of the students are scanned and captured by a digital camera, and the faces are detected using HAAR Cascade Classifier, and then they are compared with a database of the faces of the students enrolled in this class. The system is good for face recognition and thus taking the attendance. However, it is limited in terms of capturing the facial expressions of the students during the class, so the students might be attending the class, but they are not engaged or active.

The proposed model focuses on detecting the engagement levels of the students in the classes by retrieving the faces of the students in the classes using CNN and then capturing their expressions using the mediapipe library provided in Python that will capture the facial landmarks and pose landmarks as both play as significant roles in the proposed model, for example, a bored student will put his head on his hand as he leans to the desk. Additionally, some smart assistants such as Google Home, Alexa, and Siri were designed for other purposes such as interacting with users using speech recognition only without being able to detect their emotions, but the proposed model VTA depends on facial emotion detection of the user.

RESEARCH OBJECTIVES

This research uses CNN to monitor the attention of the students in the class and improve the teaching process by studying their facial expressions, which will allow the teacher to make the correct evaluation of how the students are learning and help them achieve the success they need in their studies. Based on the issues and problems stated earlier, the following research objectives were identified to guide the research as follows: identify the factors which can be used to evaluate the performance of the teacher and students, such as the face and body gestures of the students, their voice tone when answering the questions asked by the teacher, develop a CNN model that can analyze the data obtained during the classes and use them to evaluate the class and predict whether the teaching methodology should be changed or not to improve the effectiveness of lecturing, and evaluate the overall performance of the students based on their attention during the classes.

This research will answer the following points: the cultural and social backgrounds of the participants and how it can affect the learning process, the right timing and duration for the classes to keep the students active, the face gestures to detect, the learning aspects of the study, how to predict whether the student is reacting well or poorly in the class.

THE PROPOSED VIRTUAL TEACHING ASSISTANT (VTA)

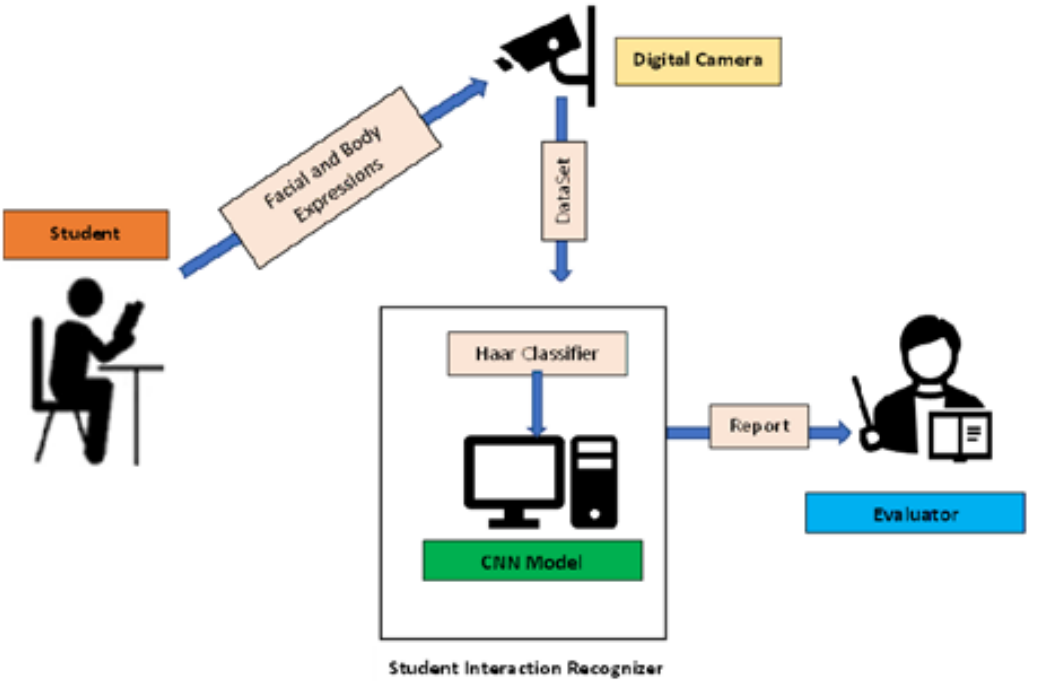
The VTA model captures the faces of the students of the classes using a camera installed in the class and then recognizes the identities of the students and their facial expressions using the VTA model and then sends the predictions to an evaluator who will learn about the performance of the students in the class. The Architecture of the VTA model is shown in Figure 4 below:

The VTA algorithm is described in Figure 5 below:

EXPERIMENTAL RESULTS

The experiments were done by using Python as programming along with OpenCV and Keras which are used for image processing and creating and training CNN models. The software used is Jupyter Notebook.

Figure 4. Architecture of the VTA Model



The experiments went through 3 different phases: training phase, capture the facial and pose expressions, and predicting the engagement level.

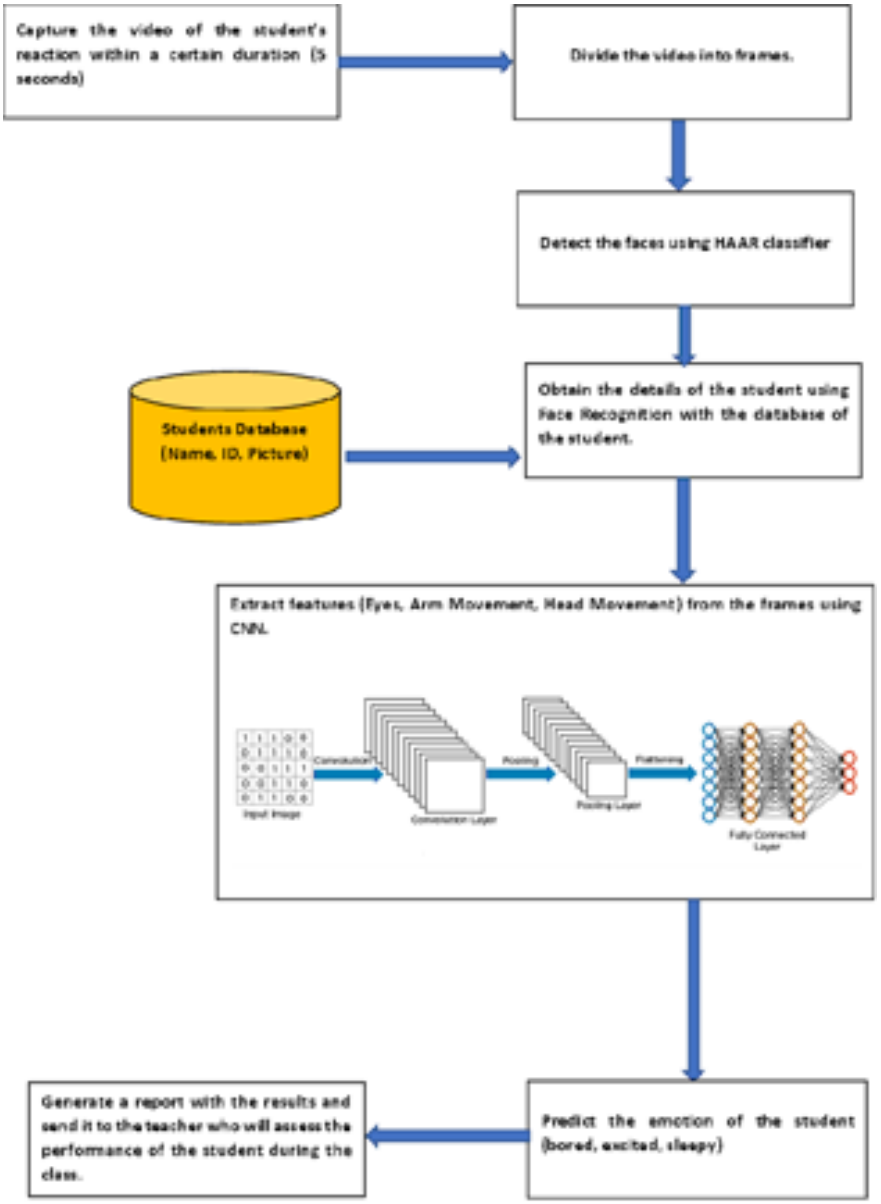
Training Phase

In this phase, the VTA model should be trained to recognize the faces of the students by using a CNN with a dataset of the faces of each student as shown below (figure 6)

The training is done as shown in the code below:

```
‘’’Initializing the Convolutional Neural Network’’’
classifier= Sequential()
 ‘’’ STEP--1 Convolution
# Adding the first layer of CNN
classifier.add(Convolution2D(32, kernel_size=(5, 5), strides=(1,
1), input_shape=(64,64,3), activation=’relu’))
 ‘’# STEP--2 MAX Pooling’’’
classifier.add(MaxPool2D(pool_size=(2,2)))
 classifier.add(Convolution2D(64, kernel_size=(5, 5), strides=(1,
1), activation=’relu’))
 classifier.add(MaxPool2D(pool_size=(2,2)))
 ‘’’# STEP--3 FLattening’’’
classifier.add(Flatten())
 ‘’’# STEP--4 Fully Connected Neural Network’’’
classifier.add(Dense(64, activation=’relu’))
 classifier.add(Dense(OutputNeurons, activation=’softmax’))
 ‘’’# Compiling the CNN’’’
classifier.compile(loss=’categorical_crossentropy’, optimizer =
```

Figure 5. Algorithm of the VTA Model




```
'adam', metrics=["accuracy"])
 # Starting the model training
classifier.fit(training_set,epochs=1000,validation_data=test_set,)
```

Capturing the Facial and Pose Expressions

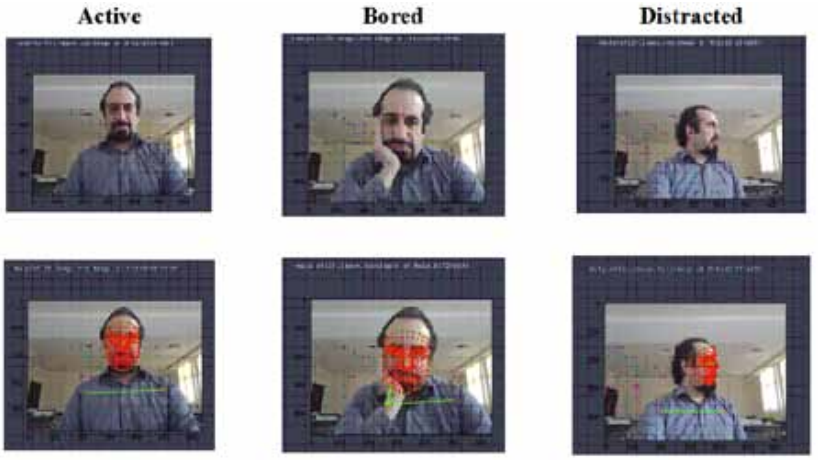
In this phase, the model must be trained to capture the facial and pose landmarks. The mediapipe library in python has been used. There are three main expressions that the model was trained on shown in the table below (figure 7):

The following code can be used to capture the face and body landmarks

Figure 6. Training Set

```
import cv2
import numpy as np
import os
from matplotlib import pyplot as plt
import time
import mediapipe as mp
mp_holistic=mp.solutions.holistic
mp_drawing=mp.solutions.drawing_utils
def mediapipe_detection(image,model):
  image=cv2.cvtColor(image,cv2.COLOR_BGR2RGB) #Color Conversion BGR
2 RGB
  image.flags.writeable=False # Image is no longer writeable
  results=model.process(image)    # Make prediction
  image.flags.writeable=True  # Image is now writeable
  image=cv2.cvtColor(image,cv2.COLOR_RGB2BGR) #Color Conversion RGB
w BGR
  return image,results
```

Figure 7. Facial and Pose Landmarks



```
def draw_landmarks (image,results):
  mp_drawing.draw_landmarks (image,results.face_landmarks,mp_
holistic.FACE_CONNECTIONS) # Draw face connections
  mp_drawing.draw_landmarks (image,results.pose_landmarks,mp_
holistic.POSE_CONNECTIONS) # pose connection
  mp_drawing.draw_landmarks (image,results.left_hand_landmarks,mp_
holistic.HAND_CONNECTIONS) # Draw left hand connections
  mp_drawing.draw_landmarks (image,results.right_hand_landmarks,mp_
holistic.HAND_CONNECTIONS) #Draw right hand connections
```

The training was done over 1000 epochs, so the loss is dropped to almost zero to guarantee the accuracy of the model as the accuracy was low when experimenting it over less than 1000 epochs

The training was done over 200, 500, and 1000 epochs as shown in the table below (figure 8) to test the accuracy and the loss of the model

PREDICTING THE ENGAGEMENT LEVEL

By using the trained model, the teacher can see the name of the student and determine his attention span in the class as the results are displayed on the screen and sent to the evaluator.

Figure 8. The epoch categorical accuracy and the epoch categorical loss

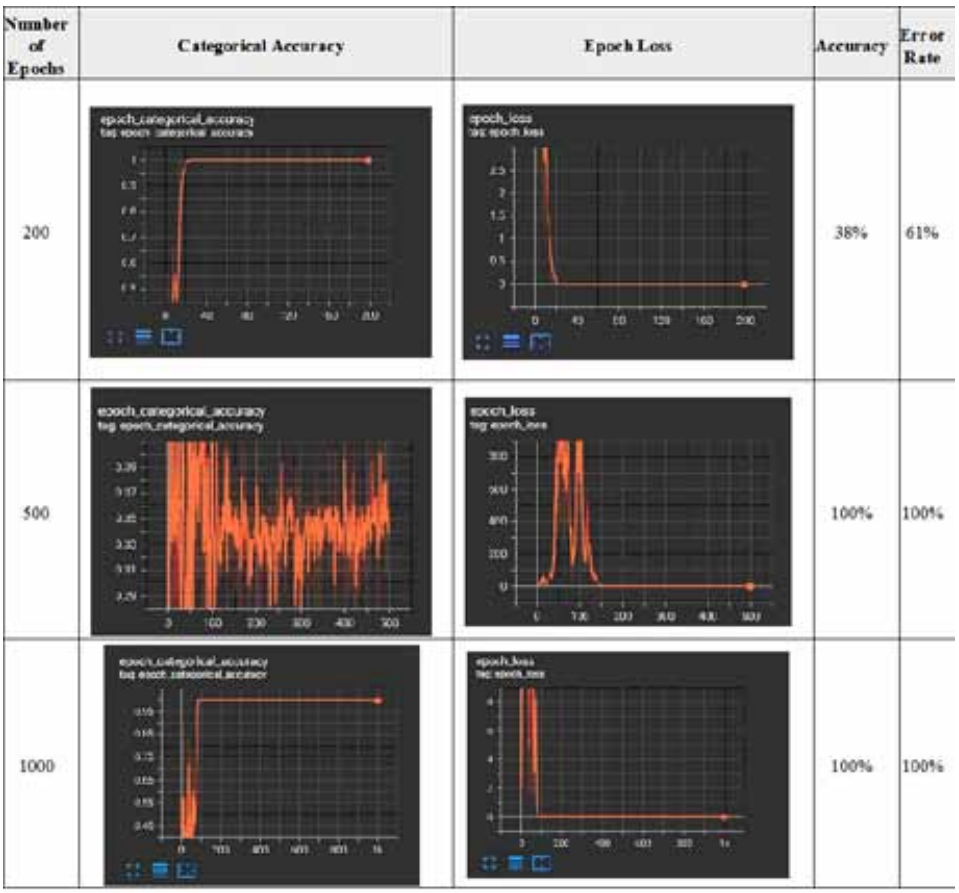


Figure 9. Prediction of the VTA Model

A performance test has been conducted to evaluate the VTA efficiency (figure 9). The accuracy of each sample was calculated as the following: The training was done for 100 epochs and took around 2 minutes with an accuracy of 98%.

CONCLUSION AND RECOMMENDATIONS

In this research, a Virtual Teaching Assistant (VTA) model for Capturing Facial and Pose Landmarks of the Students in the Classroom Based on the Deep Learning is proposed to help in achieving the learning outcomes for students and improving the teaching methods. This research will help the teacher focus more on teaching and then analyze the students' performance later with the help of the data recorded by the model. The VTA model can help the schools/institutions know more about the overall attention and understanding of the students during the classes. This model will use Convolutional Neural Networks (CNN) to detect the faces of the students and extract the required features such as facial gestures and then use these features to predict the attention of the students in the class. Class management is essential for successful teaching, and it is one of the biggest challenges that teachers face nowadays.

In future research, we will develop the model to recognize the identity of the students using the biometric tools such as the iris identity without having to train the model using the faces of the students one by one by integrating the model with the registration system that might contain the identity details of the students in each class.

REFERENCES

Aggarwal, C. C. (2018). Neural Networks and Deep Learning. In Neural Networks and Deep Learning. doi:10.1007/978-3-319-94463-0

Agravante, D. J., De Magistris, G., Munawar, A., Vinayavekhin, P., & Tachibana, R. (2018). *Deep Learning with Predictive Control for Human Motion Tracking*. Cornell University.

Azeez, R. A., & Azeez, P. Z. (2018). Incorporating Body Language into EFL Teaching. *Koya University Journal of Humanities and Social Sciences*, *1*(1), 36–45. Advance online publication. doi:10.14500/kujhss.v1n1y2018.pp36-45

Bakken, J. P., Varidireddy, N., & Uskov, V. L. (2020). Smart Universities: Gesture Recognition Systems for College Students with Disabilities. *Smart Innovation*. *Systems and Technologies*, *188*, 393–411. Advance online publication. doi:10.1007/978-981-15-5584-8_34

Bataeva, A. (2021). *Google AI neural network simulates camera movement*. https://neurohive.io/en/applications/google-ai-neural-network-simulates-camera-movement/

Campesato, O. (2020). *Artificial Intelligence*. Machine Learning, and Deep Learning.

Cinelli, L., Chaves, G., & Lima, M. (2018). *Vessel Classification through Convolutional Neural Networks using Passive Sonar Spectrogram Images*. 10.14209/sbrt.2018.340

Dahlgren, R. L. (n.d.). *From martyrs to murderers : Images of teachers and teaching in Hollywood films*. Academic Press.

Fuzail, M., Muhammad, H., Nouman, F., Mushtaq, M. O., Raza, B., Tayyab, A., & Waqas Talib, M. (2014). Face Detection System for Attendance of Class Students. *International Journal of Multidisciplinary Sciences and Engineering*, *5*(4).

Hatt, M., Parmar, C., Qi, J., & El Naqa, I. (2019). Machine (Deep) Learning Methods for Image Processing and Radiomics. *IEEE Transactions on Radiation and Plasma Medical Sciences*, *3*(2), 104–108. Advance online publication. doi:10.1109/TRPMS.2019.2899538

Hazim Barnouti, N., Sameer Mahmood Al-Dabbagh, S., & Esam Matti, W. (2016). Face Recognition: A Literature Review. *International Journal of Applied Information Systems*, *11*(4), 21–31. Advance online publication. doi:10.5120/ijais2016451597

Hutter, F. (2019). Automated Machine Learning. Springer. doi:10.1007/978-3-030-05318-5

Kamani, M. H., Safari, O., Mortazavi, S. A., Mehraban Sang Atash, M., & Azghadi, N. M. (2017). Using an image processing based technique and predictive models for assessing lipid oxidation in rainbow trout fillet. *Food Bioscience*, *19*, 42–48. Advance online publication. doi:10.1016/j.fbio.2017.05.005

Kamarudin, N., Jumadi, N. A., Mun, N. L., Keat, N. C., Ching, A. H. K., Mahmud, W. M. H. W., Morsin, M., & Mahmud, F. (2019). Implementation of haar cascade classifier and eye aspect ratio for driver drowsiness detection using raspberry Pi. *Universal Journal of Electrical and Electronic Engineering*, *6*(5), 67–75. Advance online publication. doi:10.13189/ujeee.2019.061609

Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies*, *26*(1), 205–240. Advance online publication. doi:10.1007/s10639-020-10230-3

Kubat, M. (2017). An Introduction to Machine Learning. In An Introduction to Machine Learning. doi:10.1007/978-3-319-63913-0

Marr, B., & Ward, M. (2019). *Artificial intelligence in practice : how 50 successful companies used artificial intelligence to solve problems*. Academic Press.

Mitchell. (1997). *Machine Learning textbook*. McGraw Hill.

Mittal, A. (2020). *Haar Cascades, Explained*. https://medium.com/analytics-vidhya/haar-cascades-explained-38210e57970d

Nagabhushan, P., Singh, S. K., & Partha Roy, B. R. (2017). Computer Vision and Image Processing. Springer.

Naveed, M., Alrammal, M., & Bensefia, A. (2020). HGM: A Novel Monte-Carlo Simulations based Model for Malware Detection. *IOP Conference Series. Materials Science and Engineering*, *946*(1), 012003. Advance online publication. doi:10.1088/1757-899X/946/1/012003

Ng, A. (2018). *Machine Learning Yearning: Technical Strategy for AI Engineers in the Era of Deep Learning* [Draft Version]. https://gallery.mailchimp.com/dc3a7ef4d750c0abfc19202a3/files/5dd91615-3b3f-4f5d-bbfb-4ebd8608d330/Ng_MLY01_13.pdf

Pedamkar. (2020). *Machine Learning vs Predictive Modelling*. Academic Press.

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. Information, 11(4). doi:10.3390/info11040193

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *3*(1), 12–27. Advance online publication. doi:10.1002/widm.1075

Wang, S. (2019). *Using deep neural networks for accurate hand0-tracking on Oculus Quest*. https://ai.facebook.com/blog/hand-tracking-deep-neural-networks/

Yang, M.-H. (2000). Hand gesture recognition and face detection in images. ProQuest Dissertations and Theses.

Zakaria, R. (2017). Smart Motion Detection : Security System Using Raspberry Pi. *Journal of the Engineering Research Institute, 30.*