International Journal of E-Services and Mobile Applications
Volume 15 • Issue 1

Bi-Model Engagement Emotion Recognition Based on Facial and Upper-Body Landmarks and Machine Learning Approaches

Haifa F. Alhasson, Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia
Ghada M. Alsaheel, Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia
Noura S. Alharbi, Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia
Alhatoon A. Alsalamah, Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia
Joud M. Alhujilan, Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia
Shuaa S. Alharbi, Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia*
https://orcid.org/0000-0003-2121-0296

ABSTRACT

Customer satisfaction can be measured using facial expression recognition. The current generation of artificial intelligence systems heavily depends on facial features such as eyebrows, eyes, and foreheads. This dependence introduces a limitation as people generally prefer to conceal their genuine emotions. As body gestures are difficult to conceal and can convey a more detailed and accurate emotional state, the authors incorporate upper-body gestures as an additional feature that improves the predicted emotion's accuracy. This work uses an ensemble machine-learning model that integrates support vector machines, random forest classifiers, and logistic regression classifiers. The proposed method detects emotions from facial expressions and upper-body movements and is experimentally evaluated and has been found to be effective, with an accuracy rate of 97% on the EMOTIC dataset and 99% accuracy on MELD dataset.

KEYWORDS

Body Pose, Classification Accuracy, Computer Vision, Face Recognition, Gesture, Image Classifiers, Machine Learning, Supervised Learning

INTRODUCTION

In recent years, automatic human behavior analysis has received much attention, especially in social signal processing. Several studies have shown that Machine Learning (ML) classifiers can understand emotion by using personal experiences like annoyance, aversion, disconnection, engagement, excitement, and pleasure as triggers for emotions (Mehta et al., 2019). To improve the event quality, it

DOI: 10.4018/IJESMA.330756 *Corresponding Author

is important to measure participants' emotions as this provides more accurate clues for event providers than self-reports, which may cause biased decisions or poor judgments of the entire event. However, relying solely on self-report surveys to measure these emotional experiences has been criticized for its inherent limitations, including participants' difficulty in pinpointing specific reasons for their attitudes, their hesitancy to provide socially undesirable answers, and their tendency to rationalize their responses (Ciuk et al., 2015). As a result of automatic emotion detection, experience satisfaction can be measured more holistically.

Although emotion recognition has been well studied in recent years, and its performance is significantly enhanced by deep learning, one main challenge still exists. Most human emotion recognition methods through computer vision depend on facial expressions (Lee et al., 2019). This dependence has a limited ability where individuals prefer to conceal their genuine emotions, but body gestures are difficult to conceal and can convey more detailed and accurate emotional states.

This paper proposes a machine learning-based bi-modal emotion detection system comprising facial and upper-body posture to overcome this limitation. Multiple studies have proved empirically and theoretically the advantage of incorporating various modalities in emotion recognition instead of using only one method (Wu et al., 2019; Chen et al., 2013; Filntisis et al., 2019). Complicated human emotions may be fully implied by incorporating key features from various modalities, for example, facial features and body movement.

RELATED WORKS

Assessing emotions within the context of experience satisfaction is crucial, given that leisure satisfaction encompasses diverse cognitive, physical, and emotional experiences, making it a multidimensional construct (Mansfield et al., 2020). Emotions are reactions to stimuli that are personally relevant, and they can be expressed at three levels, including phenomenology (the subjective experience of the emotion), behavior (the actions associated with the emotion), and physiology (the bodily changes that occur during the emotional response). Thus, emotions can be measured through behavioral observation to capture actions and physiological measurements to capture bodily changes (Ekman, 2016).

Bi-Modal Emotion Recognition

While advancements in emotion recognition have been substantially propelled by machine learning, a key hurdle remains unresolved. Most techniques employed in computer vision for detecting human emotions predominantly rely on facial expressions. It is common for individuals to mask their genuine emotions; however, body gestures, which are more challenging to hide, can provide a more comprehensive and precise depiction of emotional states.

BlazePose (Bazarevsky et al., 2020) is a lightweight convolutional neural network architecture for human pose estimation that is adapted for real-time inference on mobile devices and runs at over 30 frames per second. They overcome the limitation of using Non-Maximum Suppression (NMS) in literature, making multiple, ambiguous boxes satisfy the intersection over union (IOU) threshold. Alternatively, it makes a bounding box around a relatively rigid body part. They used BlazeFace (Bazarevsky et al., 2019), BlazeHand (Bazarevsky et al., 2020), and CoCo (Kreiss et al., 2019).

Machine Learning Classifiers in Measurement of Satisfaction

Kaur et al. (2019) proposed a supervised two-tier ensemble method for determining a nation's Better Life Index score. They used stacked generalization based on a novel approach that integrates different ML approaches to produce a meta-machine-learning model that further aids in optimizing prediction accuracy. They combined three of the top four models as a hybrid model to improve the regression performance, where the top four models are decision trees, support vector regressions, neural networks, and random forests. The ensemble model was a considerably more accurate predictor of a nation's

life satisfaction than the base models. Predicting the life satisfaction score of a country, the model was around 90% accurate.

Image-Based Machine Learning Classifiers in Measurement of Satisfaction

Facial-recognition technology reduces the risk of bias in social desirability appearing in responses (Li et al., 2015; Poels & Dewitte, 2006; Morin, 2011). Moreover, Gonzalez-Rodrıguez et al. (2020) aim to study and analyze individuals' immediate emotional responses to a heritage place experience, focusing on the hypothesis that the emotional responses are good indicators of the perception of the quality of the service provided. They confirm that the information obtained from facial-expression recognition demonstrated that it is as valid an instrument. (Kosti et al., 2017) aim to address the problem of recognizing emotional states in context. Specifically, they present the EMOTIC database, a collection of non-controlled images of people in non-controlled environments. Annotations are made based on the apparent emotional states of the subjects depicted in the images have discussed the importance of considering the person's scene in the problem of automatic emotion recognition in the wild using EMOTIC database presented a CNN-based three-stream deep hybrid framework that combines the proposed visual feature type with the features extracted from the entire image and the primary human subject. This approach aims to optimize the integration of the proposed feature with the visual cues from the entire image and the main subject (Wu et al., 2022) proposed a hierarchical relation-based emotion recognition method using scene graphs inspired by humans' advanced reasoning patterns. In the scene, entities are labeled, and their relationships are described abstractly.

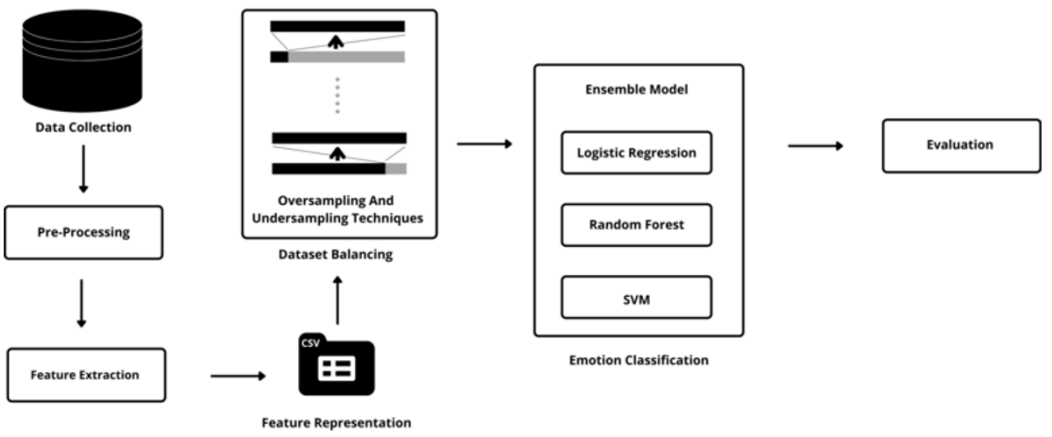
METHODOLOGY

We based our work on the approach Kaur et al. (2019) used in stacked generalization, seeking to optimize emotion prediction. We developed a bi-modal emotion detection system that employs ML algorithms to accurately detects a subject's emotional state. The proposed emotion detection model consists of several stages: data collection, pre-processing, landmark feature extraction, dataset balancing, and ML classification. Our proposed emotion detection system is depicted in Figure 1.

Data Collection

We use two standard datasets for training and validating our proposed system. This allows us to detect the relationship between emotions, facial expressions, and upper body gestures in real-

Figure 1. The proposed bi-modal emotion detection system



world settings and develop algorithms that can accurately detect emotions from video or images. The datasets are EMOTIC dataset (Kosti et al., 2017) and MELD (Multimodal EmotionLines Dataset) (Poria et al., 2018).

The EMOTIC dataset (Kosti et al., 2017) consists of images of individuals in unrestricted situations annotated based on their perceived emotional states. The EMOTIC dataset includes two types of emotional representation: (1) a collection of 26 discrete categories and (2) the continuous dimensions of valence, arousal, and dominance. It consists of 23,185 images, obtained from a combination of manual collection from the internet using a Google search engine and from two public benchmark datasets, COCO (Lin et al., 2014) and Ade20k (Zhou et al., 2019). For our proposed model, we have selected to concentrate on six specific emotions based on our target, including "Aversion," "Engagement," "Excitement," "Pleasure," "Annoyance," and "Disconnection" as these more accurately describe a person's emotional state during leisure activities. An example of the selected emotion can be found in Figure 2.

The MELD dataset (Poria et al., 2018) consists of video data collected from TV series. It contains approximately 13,000 utterances, each annotated with emotion labels. Accordingly, we have selected six emotions based on our target, including "Anger," "Disgust," "Fear," "Joy," "Sadness," "Surprise," and "Neutral" as shown in Figure 3.

Proposed Model

- **Preprocessing:** Data preprocessing is an essential step in ML that involves preparing the dataset for model training. In this pipeline, we covert the MELD dataset's videos into a set of images. One of the common challenges in many ML tasks is a class imbalance, where one or more classes have significantly fewer instances than the others. This can result in poor performance of the ML model, particularly for the underrepresented classes, as the model may be biased towards the majority class. Further details can be found below in the data balancing section.
- **Feature Extraction:** The simultaneous inference of several dependent neural networks (Bazarevsky et al., 2020; Zhang et al., 2020) in real-time perception that combines face landmarks and body landmarks into a semantically compatible end-to-end solution is a uniquely challenging problem (Chen et al., 2013). To address this challenge, MediaPipe Holistic model (Neilblaze, 2020) provides rapid and accurate solutions, as shown in Figure 4, which employs MediaPipe Facial Mesh model (Joefernandez, 2020) to estimate the human face and BlazePose's (Bazarevsky

Figure 2. The six emotions selected for emotion detection from the EMOTIC dataset

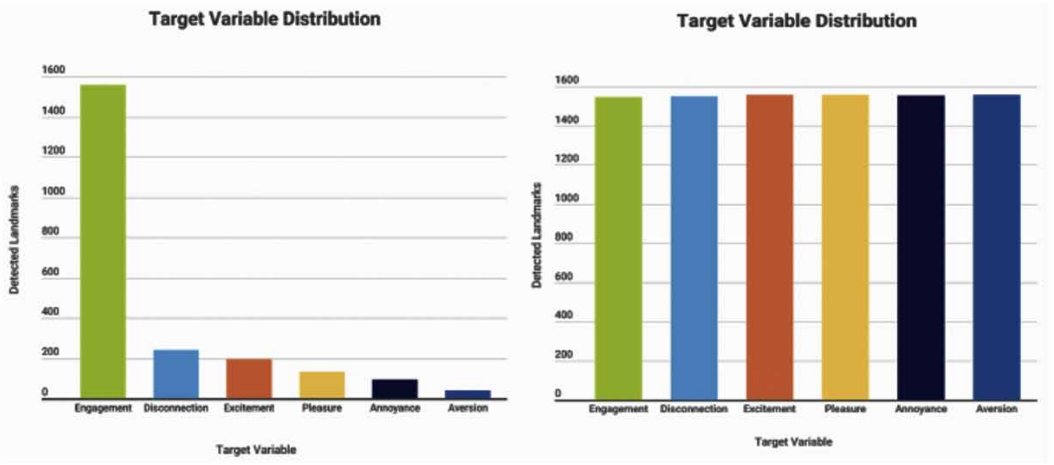


Figure 3. Selected frames from MELD dataset in seven basic emotions from left to right: Anger, disgust, fear, joy, sadness, surprise, and neutral

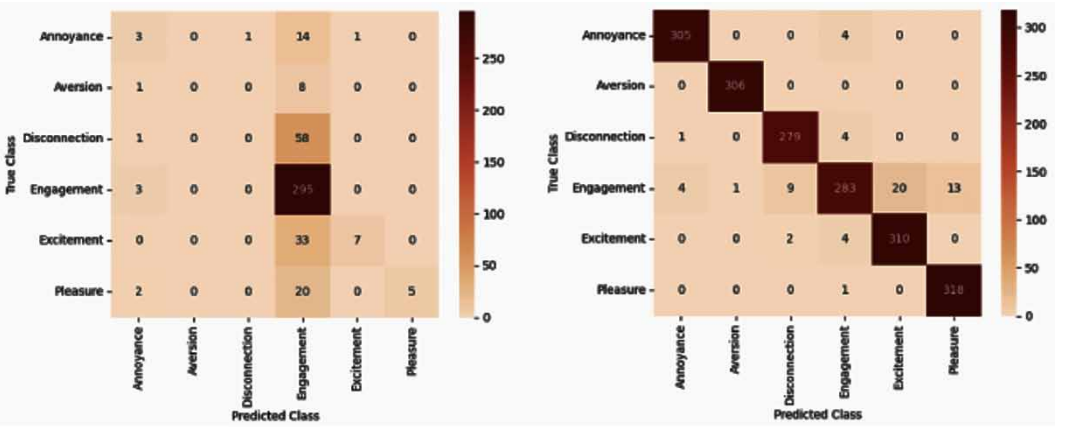
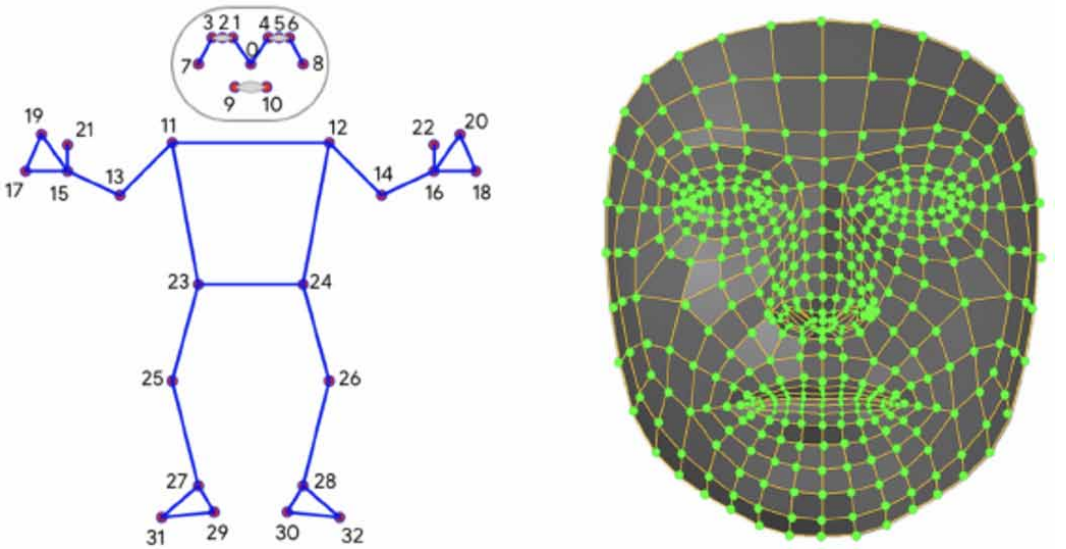


Figure 4. MediaPipe holistic Model



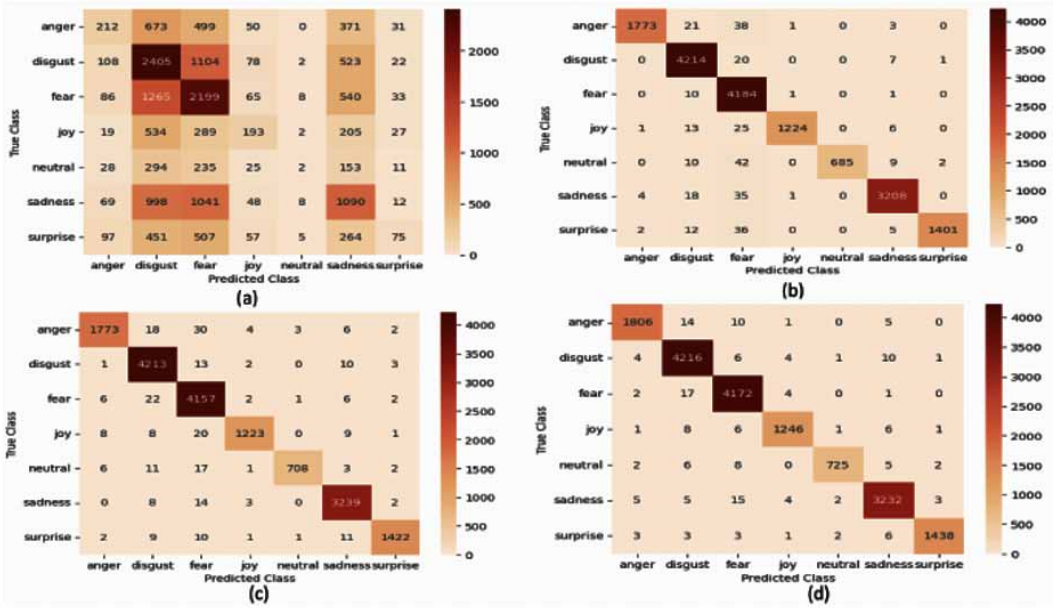
et al., 2020) proposed detector and landmark model to estimate the body pose. The model derives ROI for the hands and face, improves these regions using a re-crop model, and applies task-specific models to the ROI to detect landmarks for the hands and face. The model combines all the landmarks to yield over 540+ landmarks, allowing for a highly detailed representation of the subject's facial and body features. The extracted landmarks will be more effective in any environment if a threshold value is defined for the visibility of other objects. A CSV file is created, which is used to store the coordinate values, which are sent to the ML algorithm for the classification of emotions after they have been detected and obtained. The model is trained using various machine learning classifiers from the scikit-learn package, specifically SVM, Logistic Regression, and Random Forest. Moreover, training was carried out using an ensemble model.

- **Facial Expression Recognition:** MediaPipe Holistic uses the Facial Mesh model to estimate the location of 468 3D facial landmarks in real-time to capture a great range of facial expressions, as

illustrated in Figure 4. Machine learning is used to determine the 3D facial surface from a single camera input without a depth sensor. A face transformation matrix and a triangular face mesh are typical 3D primitives in the face transform data. Procrustes Analysis, a lightweight statistical analysis technique, is used to develop a robust algorithm. Correctly cropping the face reduces the need for data enhancements like rotations, translations, and scale adjustments. Instead, the network can focus its resources on improving coordinate prediction. It is possible to create crops based on facial landmarks from the previous frame, and the face detector is only employed when the landmark model cannot identify a face's presence.

- **Body Gesture Recognition:** The BlazePose model utilized by MediaPipe Holistic is capable of accurately estimating thirty-three 3D landmarks of the human body in real-time as illustrated in Figure 5, including vital points such as the nose, shoulders, elbows, wrists, hips, knees, and ankles. To achieve this, the model uses a two-step detector-tracker machine learning process. In the first step, the pipeline applies a detector to identify the person's region of interest (ROI) within the frame. Once the ROI is determined, the tracker utilizes the cropped ROI frame as input and predicts the pose landmarks within it. This two-step process improves the speed and accuracy of the landmark estimation, making it ideal for real-time applications where processing speed is critical.
- **Classification Phase:** We use a set of most popular ML classifiers like: Support Vector Machines (SVM) (Cortes & Vapnik, 1995), Random Forest Classifier (RF) (Ho, 1995) and Logistic Regression (LR) (Cox, 1958; Sammut & Webb, 2011). The results of all these classifiers will be combined in the ensemble model. Ensemble Modeling (EM) (Rokach, 2010; Sammut & Webb, 2011) is a popular machine-learning technique that combines several models' predictions to improve the system's overall performance. The idea behind the ensemble model is that multiple models are better than one, and by combining the predictions of different models, the final prediction is more accurate and less prone to errors.

Figure 5. Mediapipe landmarks where (a) illustration of various coordinates in the human body using Mediapipe, and (b) illustration MediaPipe face mesh model with 468 vertices



One of the popular implementations of the ensemble model in scikit-learn is the Voting Classifier. It combines the predictions from multiple models by taking a majority vote. It can be used for both binary and multi-class classification problems. Figure 1 shows the detailed structure of the proposed model.

Dataset Balancing

When working with the EMOTIC dataset, we encountered a significant imbalance in the distribution of emotions represented in the data, as shown in Figure 6. The class imbalance issue has been highlighted in previous studies (Wu et al., 2019), where certain classes had substantially higher volumes than others. To address this issue, we applied a combination of over-sampling and under-sampling techniques using the SMOTETomek (Ghafourian et al., 2022) approach for balancing the dataset. As shown in Figure 7, the target variable count before implementing the SMOTEKomek technique was found to be imbalanced, with "Engagement" having the highest count followed by "Disconnection,"

Figure 6. The confusion matrix of the ensemble model applied on the EMOTIC dataset before and after preprocessing: (a) Confusion matrix before balancing, (b) Confusion matrix after balancing

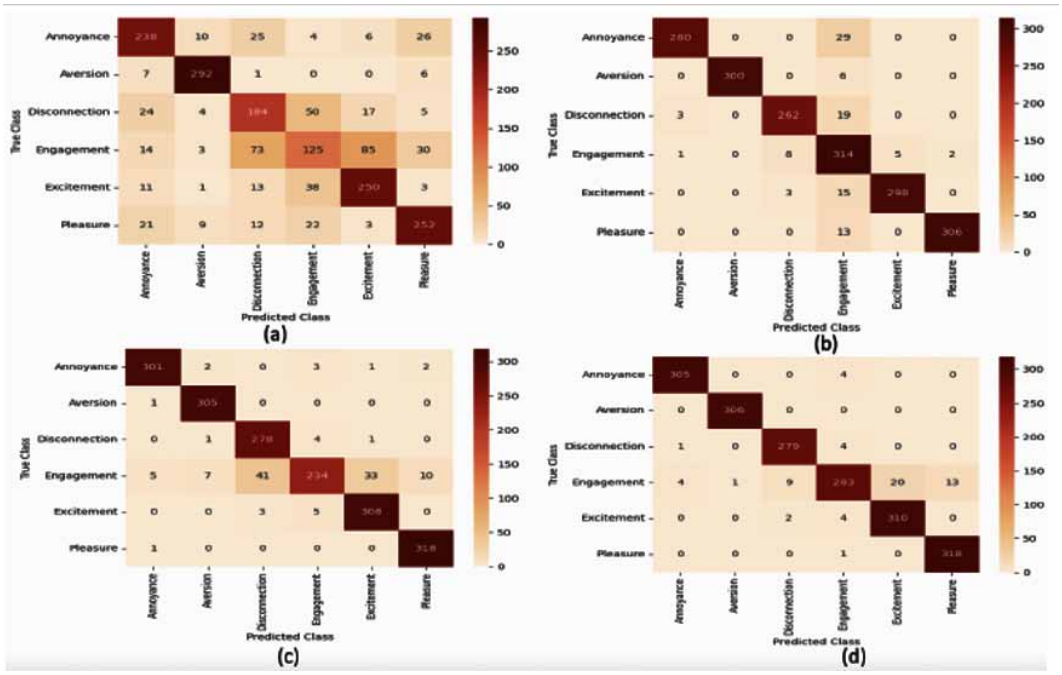


Figure 7. The distribution of emotion classes in the EMOTIC dataset before and after balancing: (a) Target variable count before balancing, (b) target variable count after balancing

"Excitement," "Pleasure," "Annoyance," and "Aversion." The SMOTETomek technique combines oversampling and undersampling techniques to generate synthetic samples for the minority classes and remove noisy samples from the majority classes, respectively. The resulting target variable count after implementing the SMOTETomek technique is shown in Figure 7(b), where all classes have a count of 1544-1558.

Table 1 displays the enhancement of the accuracy of the LR, RF, and SVM classifiers, as well as the Ensemble model, following the integration of the SMOTETomek technique. This approach effectively balanced the class distribution in the EMOTIC dataset, leading to a more accurate performance of the machine learning algorithms.

Moreover, to assess the impact of the SMOTETomek technique on the ensemble model, we compared its confusion matrix before and after applying the technique. Before implementing the SMOTETomek technique, the confusion matrix showed low accuracy for most classes, while "Engagement" had the highest accuracy. However, "Annoyance," "Aversion," and "Disconnection" had very low accuracies, resulting in an overall accuracy of only 0.69 for the ensemble model.

We observed that after applying the SMOTETomek technique, there was a significant improvement in accuracy for all classes, resulting in an overall accuracy of 0.97. All classes achieved accuracies above 0.90, with "Annoyance," "Aversion," and "Disconnection" achieving the highest accuracies. These results demonstrate the effectiveness of the SMOTETomek technique in improving the accuracy of the ensemble model for all classes.

RESULTS AND DISCUSSIONS

Performance Metrics

To evaluate the performance of the ensemble model, we used a set of performance metrics, including Overall accuracy, Average precision, recall, and F1 score based on the following four values: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The number of accurately predicted data points out of all the data points is called accuracy; ideally, it must be close to one. The evaluation was performed based on six different emotions of "Annoyance," "Aversion," "Disconnection," "Engagement," "Excitement," and "Pleasure." The ensemble model's evaluation findings for the EMOTIC dataset and the MELD data set are shown in Tables 2 and Table 3, respectively. The model achieved high precision, recall, and F1-score for all six emotions, ranging from 0.90 to 1.00.

Qualitative Analysis of the Findings

In addition to the quantitative analysis presented previously, we also performed a qualitative analysis of the emotion detection results through a confusion matrix, as shown in Figures 8 and 9, to calculate classification performance metrics, such as accuracy, precision, recall, and F1 score, To facilitate this analysis, we used a confusion matrix to visualize the distribution of emotions that the ensemble model correctly or incorrectly detected.

Table 1. Effect of data balancing on emotion detection accuracy: A comparison on the EMOTIC dataset

ML Alg.	Before SMOTETomek	After SMOTETomek
LR	0.67	0.71
SVM	0.65	0.93
RF	0.67	0.94
Ensemble	0.68	0.97

Table 2. Quantitative evaluation comparison of the EMOTIC dataset

ML Model	Evaluation Matrix			
	F1-Measure	Recall	Average Precision	Overall Accuracy
Logistic Regression	0.72	0.72	0.71	0.719
Random Forest	0.94	0.94	0.94	0.935
SVM	0.94	0.94	0.95	0.944
Ensemble Model	0.97	0.97	0.97	0.966

* Bold font indicates the best achieved value.

Table 3. Quantitative evaluation comparison of the MELD dataset

ML Mode	Evaluation Matrix			
	F-Measure	Recall	Average Precision	Overall Accuracy
Logistic Regression	0.36	0.25	0.32	0.719
Random Forest	0.98	0.98	0.99	0.983
SVM	0.98	0.97	0.99	0.980
Ensemble Model	0.99	0.99	0.99	0.989

* Bold font indicates the best achieved value.

Figure 8. Confusion matrix evaluation of EMOTIC dataset: (a) Logistic regression, (b) SVM, (c) random forest, and (d) ensemble model

As can be seen from the confusion matrix Figure 8 a, b, c, and d of the comparison of the three models and the ensemble model, the ensemble model achieved the highest accuracy in detecting all six emotions, with the majority being correctly detected. For instance, annoyance was correctly detected in 305 out of 309 instances, while aversion was correctly detected in all instances.

Disconnection was correctly detected in 279 out of 284 instances, and engagement was correctly detected in 283 out of 330 instances. Excitement was correctly detected in 310 out of 316 instances, and pleasure was correctly detected in 318 out of 319 instances.

Figure 9. Confusion matrix evaluation of the MELD dataset: (a) Logistic regression, (b) SVM, (c) random forest, and (d) ensemble model

The confusion matrix also shows that some emotions were more challenging to detect than others. For instance, engagement and excitement had a higher number of misclassified instances compared to the other emotions. This could be attributed to the fact that these emotions have a higher degree of overlap with each other and other emotions, making them more challenging to differentiate accurately.

Our decision to choose balanced EMOTIC over MELD to train our model was due to the latter's unreliability as can be seen from the confusion matrix in Figure 9 a, b, c, and d. Our results are shown through confusion matrices 8a, 8b, 8c, 8d, illustrating how the EMOTIC dataset has enabled our model to detect emotions accurately. The lesser misclassifications in each emotion category demonstrate that the balanced EMOTIC dataset has enabled the model to distinguish better between different emotions, resulting in more reliable detection of emotions. In contrast, the inconsistency of MELDs accuracy can vary significantly between different emotion classifiers, making it an unreliable choice for our work.

In conclusion, these findings highlight the potential for using ensemble models trained on the balanced EMOTIC dataset, Figure 8d, to improve the accuracy and effectiveness of bi-modal emotion detection in various domains. The interpretation of our results indicates that our model's bi-modal approach significantly improved detecting emotions' accuracy and effectiveness, particularly in dynamic scenarios where facial expressions are insufficient. Most recent research on emotion detection has focused on using facial expressions to detect emotions, while studies focusing on bi-modal detection are lacking (Kosti et al., 2017; Schindler et al., 2008). In contrast to previous studies, ours contributes to previous research by highlighting the significance of combining facial and upper body cues in detecting emotions, which provides a more comprehensive view of individuals' emotional states. Thuseethan et al. (2021) used of both cues allows the model to better capture the complexities of human emotions and improve the accuracy of emotion detection, especially in dynamic environments where facial expressions alone may be insufficient.

CONCLUSION

The bi-modal emotion detection model has been shown to be effective in measuring leisure satisfaction and providing insights into the emotions associated with higher satisfaction levels. The model can

potentially improve leisure satisfaction and quality of life by identifying interventions that increase positive emotions during leisure activities. Our bi-model emotion detection model uses a machine learning ensemble approach. Three classifiers are included: SVM, Random Forest, and Logistic Regression. A combination of the outputs of the three classifiers produced a 97% and 99% accuracy on EMOTIC and MELD datasets, respectively. Based on these results, future research can explore the use of bi-modal emotion detection to measure other aspects of quality of life.

AUTHOR NOTES

The authors would like to thank the Deanship of Scientific Research, Qassim University, for funding the publication of this project.

REFERENCES

Chen, S., Tian, Y., Liu, Q., & Metaxas, D. N. (2013). Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Image and Vision Computing*, *31*(2), 175–185. doi:10.1016/j.imavis.2012.06.014

Ciuk, A., Troy, M., & Jones. (2015). Measuring emotion: Self-reports vs. physiological indicators. *SSRN Electronic Journal*.

Cortes, V., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. doi:10.1007/BF00994018

Cox, R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B. Methodological*, *20*(2). doi:10.1111/j.2517-6161.1958.tb00292.x

Ekman, P. (2016). What scientists who study emotion agree about. *Perspectives on Psychological Science*, *11*(1), 31–34. doi:10.1177/1745691615596992 PMID:26817724

Filntisis, P. P., Efthymiou, N., Koutras, P., Potamianos, G., & Maragos, P. (2019). Fusing body posture with facial expressions for joint recognition of affect in child-robot interaction. *IEEE Robotics and Automation Letters*, *4*(4), 4011–4018. doi:10.1109/LRA.2019.2930434

Ghafourian, S., Sharifi, R., & Baniasadi, A. (2022). Facial emotion recognition in imbalanced datasets. *Computer Science and Information Technology*.

Gonz'alez-Rod'rıguez, M. R., Díaz-Fernández, M. C., & Gómez, C. P. (2020). Facial-expression recognition: An emergent approach to the measurement of tourist satisfaction through emotions. *Telematics and Informatics*, *51*, 101404. doi:10.1016/j.tele.2020.101404

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd international conference on document analysis and recognition*.

Joefernandez. (2020). *MediaPipe Face Mesh*. https://github.com/google/mediapipe/blob/master/docs/s olutions/face_mesh.md

Kabalevsky, I., Grishchenko, K., Raveendran, T., Zhu, F., Zhang, M., & Grundmann. (2020). *Blazepose: On-device real-time body pose tracking*. arXiv.

Kaur, M., Dhalaria, P. K., Sharma, J. H., & Park, J. H. (2019). Supervised machine-learning predictive analytics for national quality of life scoring. *Applied Sciences (Basel, Switzerland)*, *9*(8), 1613–1613. doi:10.3390/app9081613

Kosti, R., Alvarez, J. M., Recasens, A., & Lapedriza, A. (2017). Emotion recognition in context. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1667–1675.

Kreiss, S., Bertoni, L., & Alahi, A. (2019). Pifpaf: Composite fields for human pose estimation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11– 977. doi:10.1109/CVPR.2019.01225

Lee, J., Kim, S., Kim, S., Park, J., & Sohn, K. (2019). Context-aware emotion recognition networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10–143.

Li, S., Scott, N., & Walters, G. (2015). Current and potential methods for measuring emotion in tourism experiences: A review. *Current Issues in Tourism*, *18*(9), 805–827. doi:10.1080/13683500.2014.975679

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *Computer Vision-ECCV 2014: 13th European Conference*, 740–755.

Mansfield, N., Daykin, T., & Kay, T. (2020). Leisure and wellbeing. *Leisure Studies*, *39*(1), 1–10. doi:10.1080/02614367.2020.1713195

Mehta, D., Siddiqui, M. F. H., & Javaid, A. Y. (2019). Recognition of emotion intensities using machine learning algorithms: A comparative study. *Sensors (Basel)*, *19*(8), 1897. doi:10.3390/s19081897 PMID:31010081

Morin, C. (2011). Neuromarketing: The new science of consumer behavior. *Society*, *48*(2), 131–135. doi:10.1007/s12115-010-9408-1

Neilblaze. (2020). *MediaPipe Holistic*. https://github.com/google/mediapipe/blob/master/docs/solutions/holistic.md

Poels, K., & Dewitte, S. (2006). How to capture the heart? reviewing 20 years of emotion measurement in advertising. *Journal of Advertising Research*, *46*(1), 18–37. doi:10.2501/S0021849906060041

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). *Meld: A multimodal multi-party dataset for emotion recognition in conversations*. arXiv.

Rokach. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*, 1–39.

Sammut, G. I., & Webb, G. I. (Eds.). (2011). *Encyclopedia of machine learning*. Springer Science & Business Media. doi:10.1007/978-0-387-30164-8

Schindler, K., Gool, L. V., & Gelder, B. D. (2008). Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural Networks*, *21*(9), 1238–1246. doi:10.1016/j.neunet.2008.05.003 PMID:18585892

Thuseethan, S., Rajasegarar, S., & Yearwood, J. (2021). Boosting emotion recognition in context using non-target subject information. *2021 International Joint Conference on Neural Networks (IJCNN)*. doi:10.1109/IJCNN52387.2021.9533637

Wu, J., Zhang, Y., & Ning, L. (2019). The fusion knowledge of face, body and context for emotion recognition. *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 108–113. doi:10.1109/ICMEW.2019.0-102

Wu, S., Zhou, L., Hu, Z., & Liu, J. (2022). Hierarchical context-based emotion recognition with scene graphs. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15. Advance online publication. doi:10.1109/TNNLS.2022.3196831 PMID:36018874

Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C. L., & Grundmann, M. (2020). *Mediapipe hands: On-device real-time hand tracking*. arXiv.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2019). Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, *127*(3), 302–321. doi:10.1007/s11263-018-1140-0

Haifa F. Alhasson received a BSc in Computer Science from Qassim University, an MSc in Computer Science from King Saud University, KSA, and her PhD in Computer Science from Durham University, UK. She is currently working as Assistant Professor at Computer College, Qassim University, KSA. Her interdisciplinary research focuses on image processing and machine learning. In particular, her research aims to understand better machine learning in object detection and recognition required for variable tasks.

Shuaa S. Alharbi received a BSc and MSc in Computer Science from Qassim University, KSA, and her PhD also in Computer Science from Durham University, UK. She is currently working as Assistant Professor at Computer College, Qassim University, KSA. Her interdisciplinary research focuses on machine learning and image processing in biology and medical domains. In particular, she is interested in using deep learning to analyze medical images and improve the accuracy of disease diagnosis, which is a rapidly growing area of interest.