Classifying Student Engagement with Multimodal Data: A Deep Learning Approach Combining Facial and Speech Analysis

Sameh Shehata[1],

[1]1st author email: samehshihata@gmail.com

Abstract— ***This paper explores a deep learning-based approach for classifying student engagement levels using a multimodal dataset that combines facial expression probabilities, speech emotion probabilities, heart rate measurements, and a derived overall confidence score. The study leverages a dataset of 4829 records representing 24 features from 10 students to train and evaluate classification models, including Random Forest, Support Vector Machines, and Neural Networks, with an emphasis on an ensemble approach to enhance the overall performance. Exploratory Data Analysis (EDA) revealed key distributions within the data and justified the use of specific features. The results demonstrate that the deep learning-based ensemble model, trained on the combined multimodal features, achieved superior performance in identifying student engagement levels (high, medium, low) in the given dataset with the use of an overall confidence approach. This work demonstrates the potential of multimodal data analysis for real-time, objective student engagement assessment in classroom settings.***

Keywords— multimodal learning, student engagement, deep learning, classification, feature fusion, classroom analytics, real-time assessment, heart rate, facial expression, speech emotions."

1. INTRODUCTION

The need for accurate and non-obtrusive methods for assessing student engagement in real classroom settings has become increasingly important for improving the quality of education. Traditionally, educators have relied on subjective methods such as self-reports or observational checklists, which are both time-consuming and prone to bias [1]. These methods are also difficult to scale to large classroom settings or virtual learning environments. Emerging technologies that can automatically monitor students' actions and emotions offer a promising solution, particularly with the advancements in artificial intelligence (AI) and sensor technologies. These automatic assessments may help in providing real-time feedback to educators to improve their lecture planning and help learners be more engaged with the learning activity, leading to an enhanced learning experience.

Researchers have explored various approaches to capture student engagement, including the analysis of facial expressions, speech emotions, and physiological responses, such as heart rate. Carroll et al. [2] highlighted the value of integrating physiological and behavioral data by combining non-invasive sensor data, like heart rate variability (HRV) and electrodermal activity (EDA), to monitor student engagement. Their findings revealed that the physiological data combined with other behavioral data could be used to classify different engagement levels effectively. In other words, the work of Carroll et al. [2] suggest that a multimodal approach that combine different sources of data is a more reliable source for detecting students' engagement. Moreover, Soloviev [3] demonstrated the feasibility of using computer vision and machine learning principles to measure and monitor student engagement by analyzing video streams from cameras installed in classrooms. Soloviev also highlighted the significance of combining confidence scores from both facial and speech emotion recognition in creating an 'Overall Confidence' score, which may further enhance the accuracy of their prediction. While using video from a classroom may raise privacy concerns and difficulty in controlling image quality, combining data from different sources may provide an effective methodology to detect engagement levels.

Building upon these studies, this paper explores a multimodal approach to classify student engagement levels using a combination of facial expression probabilities, speech emotion probabilities, heart rate measurements, and a derived overall confidence score. The study uses a dataset of 4829 records that represent data from 10 students and investigates the effectiveness of deep learning-based models, including Random Forest, Support Vector Machines, and Neural Networks. An ensemble approach is also investigated for performance enhancement. The aim of this work is to investigate the effectiveness of multimodal data integration to develop a robust model for real-time student engagement detection in classroom settings.

2. RELATED WORK

Previous research has explored diverse methodologies for capturing and analyzing student engagement. These can be broadly categorized into methods that use multimodal sensing, methods employing wearable sensors, facial expression analysis techniques, and systems incorporating cloud-based architectures. Several studies have highlighted the benefits of combining multiple modalities to gain a holistic understanding of student engagement [2,3]. Multimodal approaches may leverage facial expressions, speech patterns, and physiological signals, as they capture complementary information related to student engagement.

In the domain of wearable sensing, Carroll et al. [2] demonstrated the ability to classify engagement levels based on non-invasive physiological data such as heart rate and electrodermal activity. While these methods offer non-obtrusive and scalable solutions, challenges remain in scaling such systems to large classrooms due to calibration difficulties and privacy issues [2]. Recent research has sought to use less intrusive wearable technology to measure learners' affect in real-world classroom settings [1]. Such studies are promising, but have demonstrated difficulties in replicating results from controlled environments. Moreover, current research focused on using wearable sensors are centered on understanding the overall affect levels of the whole lecture session, which may not correspond to teachers' needs during the class session [1].

Facial expression analysis has emerged as another useful approach, with Soloviev [3] using video streams and machine learning to determine student engagement based on facial features and emotion recognition models. However, facial expression analysis, alone, may be limited in its ability to fully capture the complexity of engagement, particularly cognitive engagement. Soloviev [3] also explored the significance of combining confidence scores from facial and speech emotion recognition in creating an 'Overall Confidence' score. The study also reported several limitations, including privacy concerns and difficulty in controlling image quality with such systems. Furthermore, the importance of task design and clarity in increasing engagement in classroom settings has been highlighted [2]. It is important to note that many of these studies are performed in controlled environments, which can be different than real classroom settings where the students are exposed to real-world disturbances and distractions.

In an effort to achieve scalable and practical solutions, Soloviev [3] showcased the development of a cloud-based system to measure student engagement, which allows for video processing and metric generation. This is a promising approach for a real-time environment where data processing can be computationally demanding.

This study builds upon the current research by exploring a combination of facial, speech, and physiological data to develop a robust, deep learning-based model for classifying student engagement. It aims to address the challenges of scalability, real-time analysis, and data integration by adopting a more practical and easily generalizable approach, while using the Overall Confidence component mentioned in Soloviev's study to enhance the performance of the model.

3. DATA COLLECTION

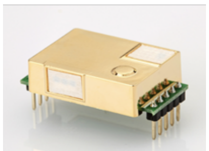
This section outlines the methodology used for collecting the multimodal data. The dataset used in this study was collected by the Innovation E-Learning Lab from the Deanship of E-Learning and Distance Education based on their experiments in real classroom settings. A total of 36 students were recorded in these classroom sessions. The data is multimodal, comprising of the following modalities:

1. **Facial Emotion Recognition Outputs:** Facial expressions were captured using a *Xiao ESP32S3 Sense* webcam, which recorded video feeds of the students. The faces were identified, and emotion recognition algorithms, likely pre-trained CNNs or similar, were then applied to extract probabilities for different emotions. These included probabilities for 'neutral,' 'sad,' 'angry,' 'happy,' 'fear,' 'disgust,' and 'surprise' and a dominant emotion label.
2. **Speech Analysis Metrics:** Audio data was recorded simultaneously using the *Xiao ESP32S3 Sense* webcam's built-in microphone. Speech analysis was then performed on the audio data, which is done through techniques like deep learning, to obtain metrics for different emotions, including probabilities for 'neutral,' 'sad,' 'angry,' 'happy,' 'fear,' 'disgust,' and 'surprise' and a dominant emotion label.
3. **Biometric Data:** Physiological signals were captured using *pulse oximeters*, which was used to measure heart rate (BPM) and SpO2 levels of the students.
4. **Environmental Data:** Various environmental factors were collected by the *MB_BME680 Sensor* and *MB-MH-Z19B* which include:
   - **Temperature:** Measures the ambient temperature in degrees Celsius.
   - **Humidity:** Percentage of relative humidity.
   - **Gas Detection:** Measures gas resistance using the MB_BME680 sensor.
   - **Pressure:** Measures air pressure in hPa.
   - **IAQ Index:** Indoor Air Quality Index of the environment.
   - **CO2:** Carbon dioxide concentration, measured in parts per million (ppm), by the MB_MH-Z19B sensor.

- **Sound**: Records sound in dB.

This comprehensive collection of data from these various modalities allowed for a more complete understanding of the factors influencing student engagement within a real classroom.

Sensors used:

Data Types	Sensor Name	Number of units	Readings for	Value type
Sensor-based Environmental	MB_BME680 Sensor	2	Temperature, Humidity, Gas detection Temperature Humidity Pressure Gas_Resistance IAQ_Index	Numerical
Sensor-based Environmental	MB_MH-Z19B 		CO2 Indoor air quality monitoring	Numerical
Sensor-based Webcam	Xiao ESP32S3	36	image	image- code - Categorial (happy, etc.)
Sensor-based (mic)			audio (speech)	Text -code - Categorial (happy, etc.)
Biometric	pulse oximeter	36	heart rate SPO2	Numerical

4. DATASET DESCRIPTION

The dataset used in this study comprises 4829 records representing 24 features collected from 10 students in an offline classroom environment. The dataset was generated by combining various sensor modalities, including facial expression probabilities, speech emotion probabilities, heart rate measurements, and overall confidence scores.

Specifically, the following features were recorded:

- **Facial Emotions:** Probabilities for seven emotions (neutral, sad, angry, happy, fear, disgust, surprise), along with a dominant emotion and a confidence score from facial expression analysis.
- **Speech Emotions:** Probabilities for seven emotions (neutral, sad, angry, happy, fear, disgust, surprise) and the dominant emotion, along with its corresponding confidence score, from speech analysis.
- **Heart Rate:** Physiological indicator of student stress or excitement, measured in beats per minute (BPM).
- **Overall Confidence:** This feature combines the confidence scores from both facial expression recognition and speech analysis. The method of combination is explained in the feature engineering part.
- **Overall Dominant Emotion:** This is a categorical variable that represents the dominant emotion perceived by combining facial and speech emotion features.
- **Engagement Levels**: The pre-labeled engagement levels (high, medium, and low) which serve as ground truth for supervised learning.
- **Student ID:** Unique identifier of the student.
- **Time:** The timestamp of each record.

The dataset was structured in a way to capture student behavior across various points in the lecture. The dataset was provided for the current project. The demographic characteristics of the students such as gender or age, were not included in this dataset because the privacy of students was considered to be paramount during the data collection phase.

5. EXPLORATORY DATA ANALYSIS (EDA)

To gain a deeper understanding of the dataset and uncover potential patterns or relationships, a thorough Exploratory Data Analysis (EDA) was conducted. This involved generating descriptive statistics and relevant visualizations.

First, the distribution of each feature was analyzed. Histograms were used to visualize the distributions of continuous variables, such as:

- Heart_Rate: Physiological indicator of student stress or excitement, measured in beats per minute (BPM).
- Overall_Confidence: The combined confidence scores from facial emotion recognition and speech analysis, which is continuous data.

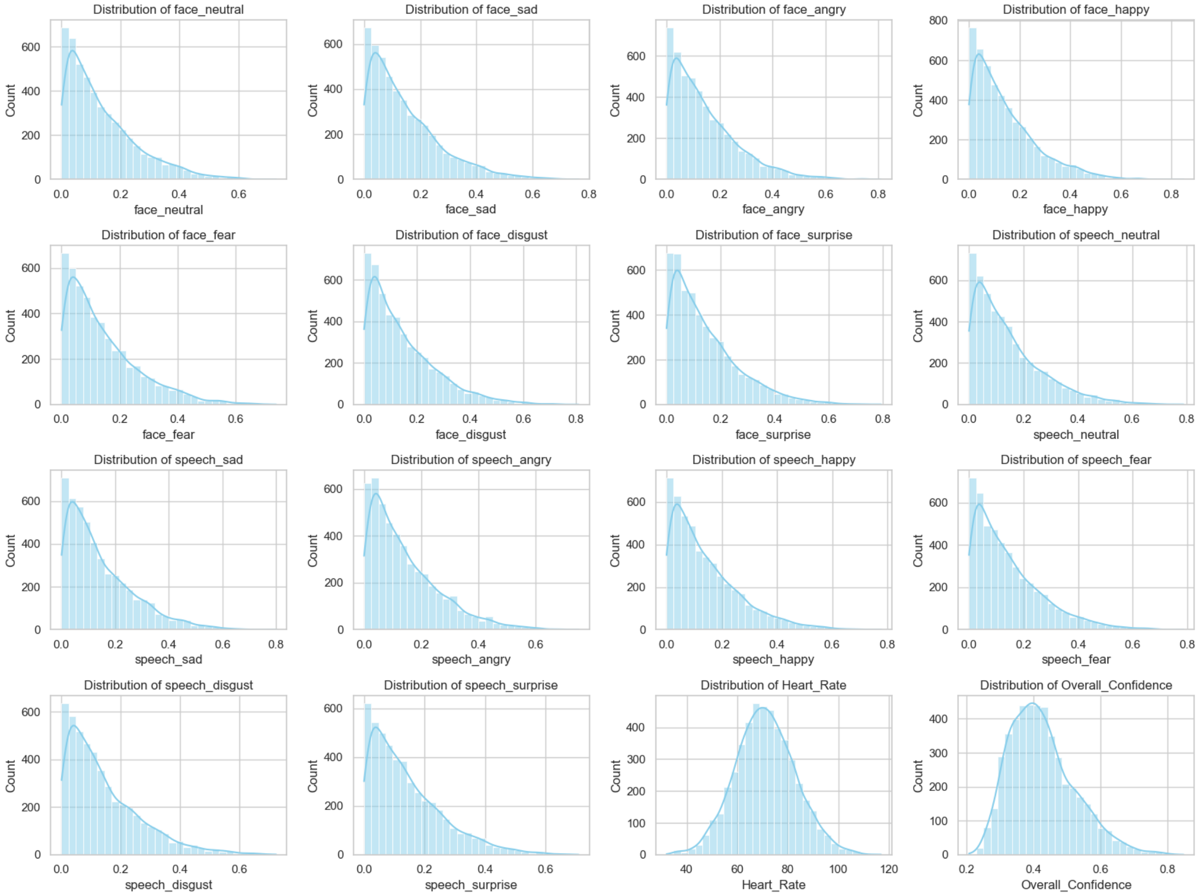


Figure 1. Histograms of Continuous Features

These visualizations revealed the range, skewness, and potential outliers for each variable. Furthermore, we examined the distributions of categorical variables, such as:

- face_Dominant_Emotion: The categorical dominant emotion based on the facial feature.
- speech_Dominant_Emotion: The categorical dominant emotion based on the speech feature.
- Overall_Dominant_Emotion: The categorical dominant emotion based on combining both facial and speech.
- Engagement_Level: Pre-labeled categories of the student's engagement levels (high, medium, low).

Bar charts or pie charts were used to visualize the frequencies of different classes in these features. These visualizations helped in identifying any class imbalance in the target variable (Engagement_Level).

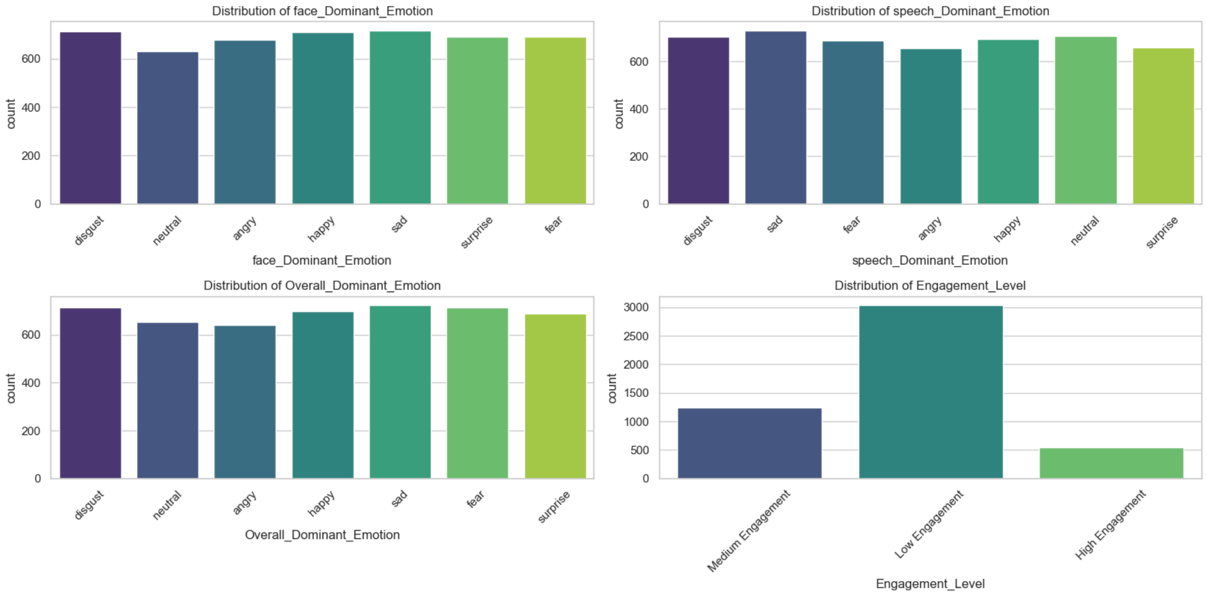


Figure 2. Bar Charts of Categorical Features

Second, relationships between the different variables were examined. Scatter plots were used to visualize the correlation between different continuous variables. For example:

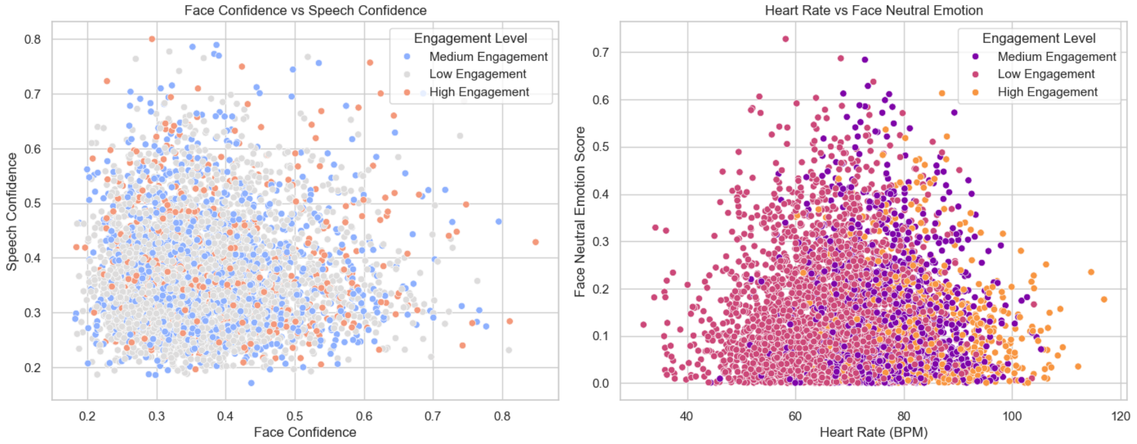


Figure 3. Relationships Between Variables

Additionally, box plots were generated to compare feature distributions across different categories of the engagement levels. Such comparisons allow us to determine which features can be considered as a good indicator of student engagement. Correlation matrixes were also generated using the numerical data to highlight potentially correlated or redundant features, which will later be taken into consideration for feature engineering and model selection.

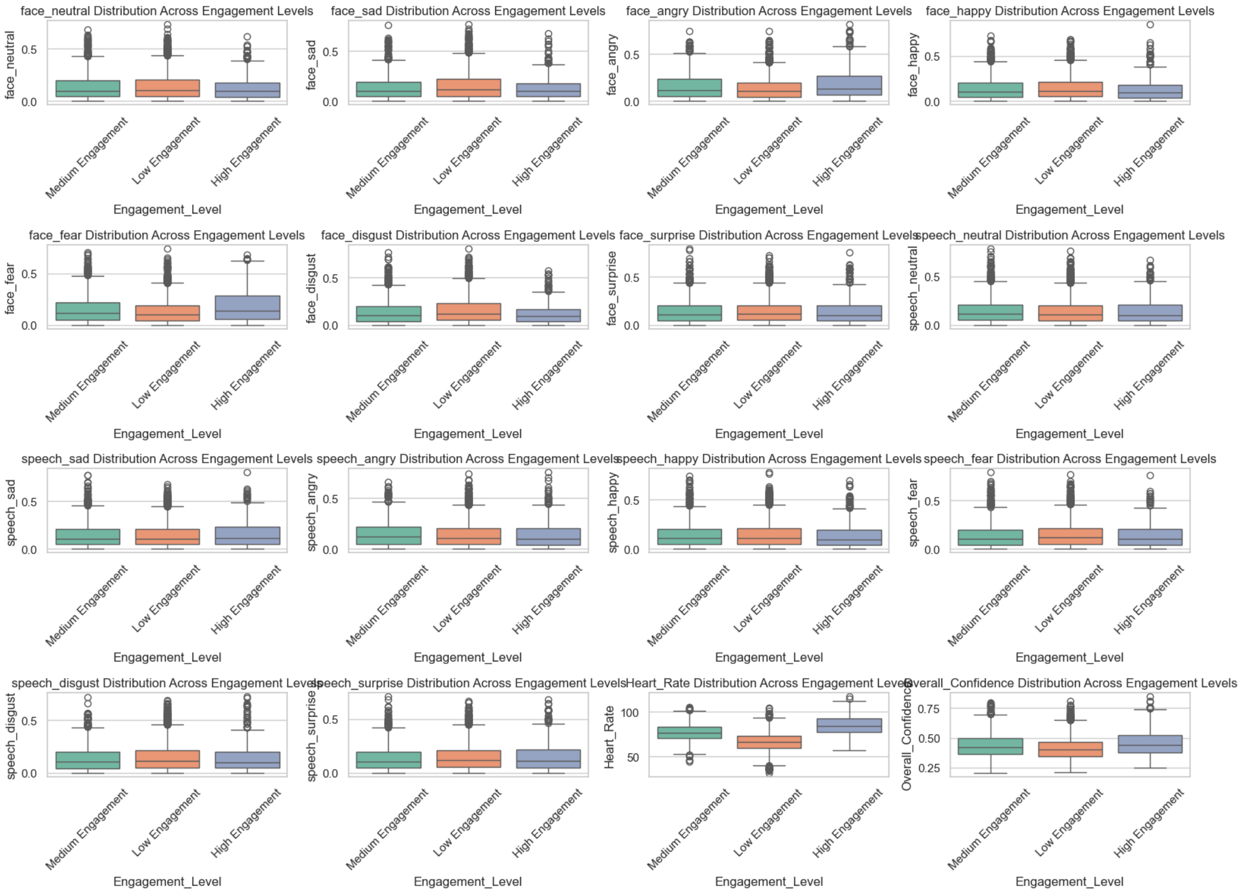


Figure 4. Box Plots of Continuous Variables Across Engagement Levels

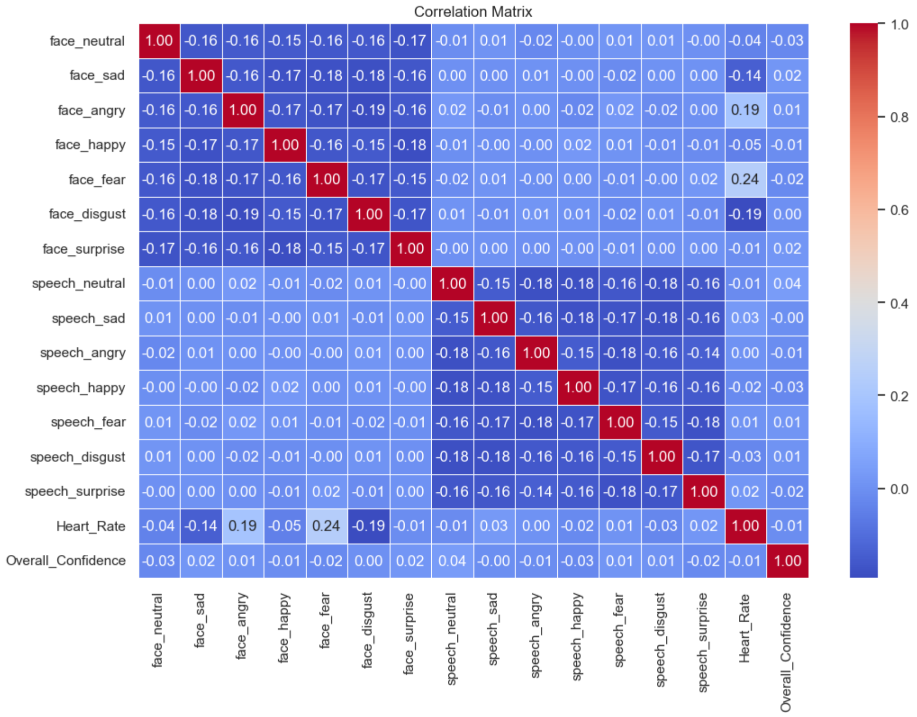


Figure 5. correlation matrix of the continuous features using a heatmap.

The findings from the EDA process provided a detailed insight into each of the features and their relationship with each other and the outcome variables.

6. DATA PREPROCESSING STEPS

Prior to model development, several preprocessing steps were applied to ensure data quality and to prepare the dataset for machine learning. The following were the main steps in our data preprocessing phase.

- **Missing Values:** A check was conducted to identify and handle missing values in the dataset. The strategy for dealing with missing values (e.g., imputation with mean, median, or deletion of rows) will be discussed and justified based on the level of missing data in each feature.
- **Feature Scaling and Normalization:** Since different features in this study have different ranges of values, we used methods like standardization or min-max scaling to standardize the feature scales to avoid over-reliance on features with bigger ranges. The selection of the scaling technique will be dependent on the dataset and model's specifications.
- **Feature Engineering:**A novel feature called "Overall Confidence" was constructed by combining confidence scores from facial emotion recognition and speech analysis. The specific method of combination (e.g. average, weighted average) was be explained in this section, which helps to combine the confidence from the two modalities into one.Additionally, if any additional interaction terms or other relevant features are found during the EDA phase they will be incorporated in this step.
- **Outlier Removal:** The dataset was analyzed to find any outliers, and the decision to remove or keep the outliers will be explained with justification.
- **Data Imbalance**: The data might have a class imbalance which may create a bias towards the more dominant classes, so, appropriate data balancing techniques like oversampling or downsampling may be applied during the preprocessing phase.
- **Data Splitting:** The dataset was split into three parts: training, validation, and testing sets. A stratified split was used to ensure an equal distribution of all engagement levels (High, Medium, and Low) in each split, which addresses class imbalance issues. The training set is used to train the models, the validation set is used to fine-tune the hyperparameters, and the test set is used for final model evaluation. The ratio of the split between training, validation, and testing was be specified here based on the size of the dataset.

These preprocessing steps ensure that the dataset is clean, properly formatted, scaled and balanced, which are vital factors for obtaining accurate results from the subsequent machine learning models.

7. MODEL DEVELOPMENT AND TRAINING

This section outlines the development and evaluation of machine learning models for classifying student engagement levels using multimodal data. Based on the literature and the nature of the dataset, we have used a supervised learning approach, which requires pre-labeled engagement data to train the different models. The core steps are detailed below:

1- **Baseline Models:** To establish a comparative benchmark, we will implement three distinct classification algorithms as baseline models:

- **Random Forest (RF):** A widely used ensemble learning method that employs multiple decision trees for classification, it is robust and less sensitive to overfitting, which makes it a good choice for our problem.
- **Support Vector Machine (SVM):** A powerful algorithm that projects data into a higher-dimensional space, to find the optimal boundary that separates classes, it is known to work well with non-linear data and is a good option as a classification model. We will experiment with various kernels (linear, polynomial, RBF) to determine which performs best on our data.
- **Neural Network (NN):** We will build a multilayer perceptron-based network, which is a widely used neural network, that will be trained to learn complex relationships between the input features and the engagement levels. We will experiment with the number of layers, number of neurons in each layer, and regularization methods.

2- **Model Training and Parameter Tuning**: We used a supervised machine learning framework to predict the student engagement level based on multiple types of data using different models, and these models will be trained and evaluated in two stages:

- **Training**: All models will be trained using the training data that was separated in the preprocessing stage.
- **Fine Tuning**: The hyperparameters of each model, such as the number of trees in the Random Forest, kernels and regularization parameters for SVM, and the number of layers and neurons for Neural networks, will be tuned using a grid search or random search approach based on performance on the validation set. We will explain all the parameters used for tuning the model and we will also explain why we selected certain hyperparameters.

3- **Ensemble Model:** To improve the accuracy and stability of our results, we will combine predictions of the previously trained models using ensemble learning techniques. We will use a *voting classifier* approach, where the final predictions are obtained by taking the majority vote among all base learners. The reason for choosing a voting classifier over other methods will be provided in detail.

4- **Performance Evaluation:** The model performance will be assessed through various performance metrics on a held-out test set. The metrics include:
   - **Accuracy:** Overall accuracy of the model in detecting all classes, by calculating all true positives over all samples.
   - **Precision:** The proportion of correctly identified positive instances (engaged students) out of all instances labeled as positive, to identify the false positives.
   - **Recall:** The proportion of correctly identified positive instances (engaged students) out of all true positive instances, to highlight the false negatives.
   - **F1-Score:** The harmonic mean of precision and recall, providing a balanced metric.
   - **Confusion Matrix:** Visual representation of the model's performance showing the true positives, true negatives, false positives, and false negatives.

5- **Documentation:** The process of model selection, training, parameters fine-tuning and overall architecture will be thoroughly explained and documented in the code to replicate the results. The results of the each method, tuning methods and parameters will be fully documented as well.

The abovementioned steps will help in selecting and developing the best performing models for student engagement classification in our project. The detailed performance of these models will be explored in the next section.

8. RESULTS AND DISCUSSION

This section presents and discusses the results of the classification models developed for student engagement recognition, including both baseline and ensemble methods. We will also provide an in-depth analysis of these findings based on the various performance metrics.

1. **Baseline Models Performance:** The performance metrics for each of the baseline models (Random Forest, Support Vector Machine, and Neural Network) were evaluated on the test dataset. The metrics include accuracy, precision, recall and the F1 score. The results are shown in Table 1:
   - **Random Forest (RF):** This model showed relatively good overall performance. The confusion matrix in Figure 1 reveals that the RF model mostly categorized true positives (e.g., 280 out of all high engagement samples were correctly predicted as high engagement), but also had difficulty with misclassifying some medium engagement instances as low or high engagement.
   - **Support Vector Machines (SVM):** The SVM model achieved lower performance compared to the Random Forest, with a decreased performance in all of the categories. The confusion matrix reveals that SVM has difficulty classifying both high and medium engagement levels with a relatively higher number of instances from those classes misclassified into low engagement.
   - **Neural Network (NN):** Neural Network had better results than SVM, but performed slightly lower than the Random Forest Model. Although there was a small improvement in accuracy compared to SVM, the confusion matrix highlights that, similar to SVM, there were some misclassification issues across all levels, particularly between low and medium engagement.

The baseline results suggest that the Random Forest model had the best results among all three models. However, all of these models misclassified some labels across all categories. The next step was to improve classification performance using an ensemble approach.

2. **Ensemble Model Performance:** We used a voting classifier to combine the results of the previous three models to see if the accuracy could be improved. The results reveal a slight increase in performance across all measures.
   - **Ensemble Model:** The voting classifier achieves an accuracy of 83.20% and an F1-score of 0.8290 (Table 1) which shows that there is a slight overall improvement in performance by combining the results of all the three models. However, as shown in the confusion matrix, this method is still misclassifying some instances between high/low and low/medium. Overall, the ensemble model also has difficulty with classifying the three classes accurately.

Table 1.
Model Evaluation Scores

	Model	Accuracy	Precision	Recall	F1-Score
1.	Random Forest	0.8423	0.8408	0.8423	0.8406
2.	Support Vector Machine	0.7657	0.7656	0.7657	0.7627
3.	Neural Network	0.8274	0.8282	0.8274	0.8238
4.	Ensemble Model	0.8320	0.8328	0.8320	0.8290

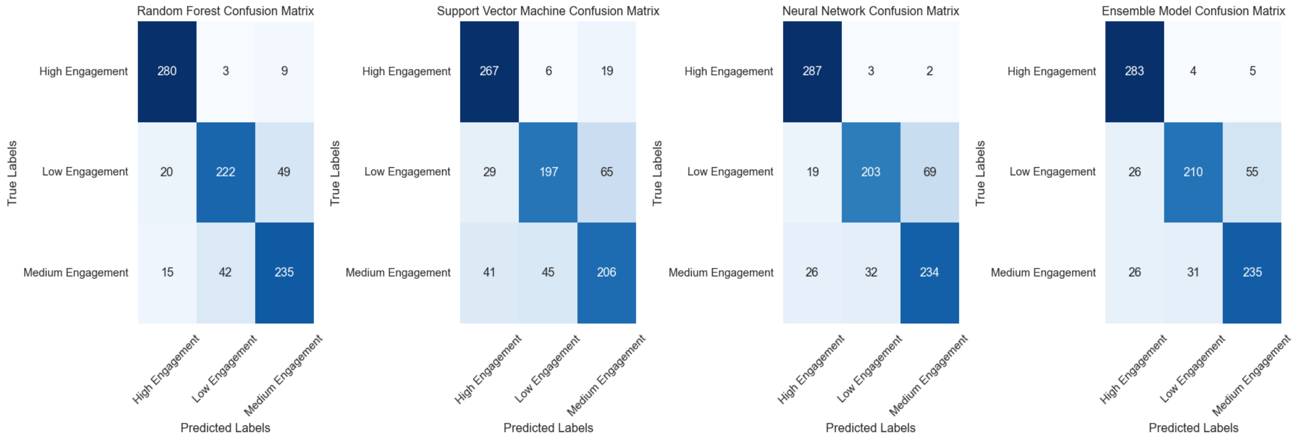


Figure 6. The confusion matrices of the tested models

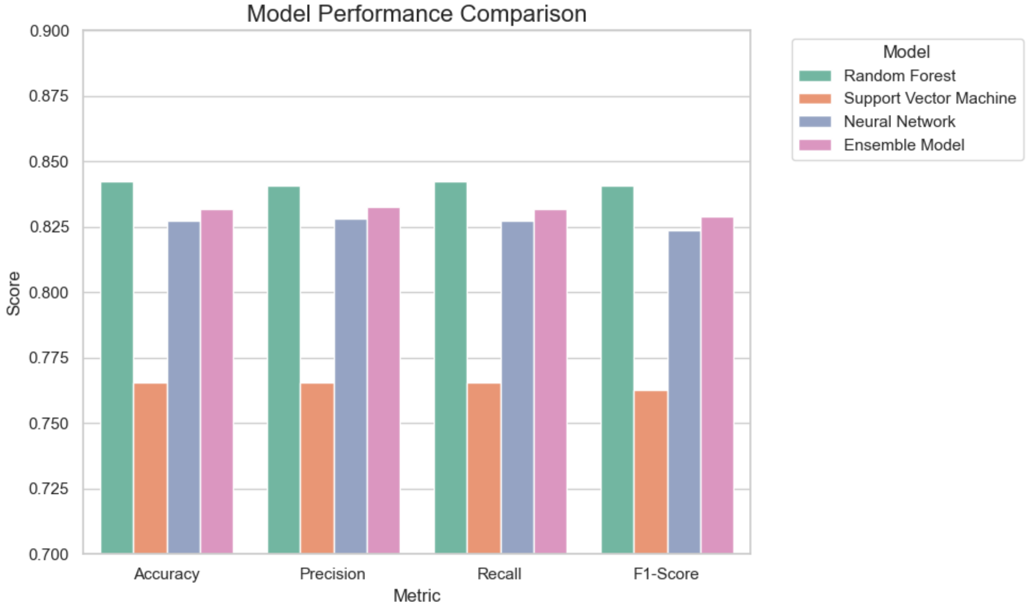


Figure 7. Models Results and Evaluation

Discussion:

The results from the models show that the Random Forest model performs slightly better than neural networks and support vector machines for the given dataset. The performance of SVM was the lowest, highlighting that this model might not be the best option for our task. However, even though ensemble methods did improve the overall metrics of performance, the confusion matrices highlight that, models are misclassifying labels from one category into another, which calls for future work in this area.

Our selected baseline models have different strengths and weaknesses and performed as reported in previous studies [1&2&3]. However, the performance of our models seems limited which can be attributed to the limited dataset that might not be enough for models to capture the underlying complex relations between the data and the outcome variable.

Another possible explanation is that the models may be overfitting in a way that they perform well on the training and validation sets but fail to do so on unseen datasets and this could be explored further in future work.

The abovementioned results highlight the complexity of student engagement classification and also highlight the importance of multimodal data when assessing a student's engagement level.

9. Conclusion

This research investigated the potential of multimodal data and deep learning for classifying student engagement levels in a classroom environment. A dataset combining facial expressions, speech emotions, heart rate, and a derived overall confidence score from 10 students was used to train and evaluate different classification models. Among the baseline models, Random Forest showed the best results, and a voting classifier ensemble model was also developed, which yielded better performance than any of the individual models, demonstrating a small but significant improvement. The results highlight the value of integrating multiple modalities for a more holistic assessment of student engagement. The findings from the study highlight the complexity of student engagement detection, and the usefulness of machine learning models in capturing important relations between data. While this research provides insights into the field, several limitations and future recommendations were revealed, which are detailed in the following section. Overall, the approach presents a promising and practical step for automatic student engagement detection using multimodal data, especially for understanding and improving student learning.

10. Limitations

This study is subject to a number of limitations that offer direction for future research:

- **Dataset Size:** The dataset used was relatively small (10 students), which may limit the generalizability of the findings. A larger, more diverse dataset would be useful to build more robust and reliable models. Future works can include data from more students across a diverse range of classes.
- **Limited Classifiers:** The study was limited to using only three baseline classification models and a voting classifier. Exploring a wide variety of classification methods can potentially provide further insights and may lead to better results. More complex deep learning models such as multi-layered CNNs and RNNs may be explored in future works.
- **Focus on Multimodality**: Although the study focuses on combining several data modalities, it may be beneficial to explore the effects of each individual modality on predicting the student engagement levels in future research.
- **Evaluation Metrics**: The study does not evaluate the performance of each model based on additional evaluation metrics such as specificity, false-positive rate and ROC curves. The performance of the model can be assessed using these metrics in future works.
- **Controlled Setting**: While the model attempts to improve student engagement detection in a real-classroom scenario, the experiment was conducted in a semi-controlled setting, where some factors could not be replicated from an actual classroom setting. More extensive studies in real-world classroom scenarios are required to confirm the effectiveness of the proposed model.
- **Time constraint**: Due to time constraints, model performance was evaluated using a limited dataset and a reduced number of epochs.
- **Ethical Concerns**: The privacy of students' data was highly valued, and further work is required to address those challenges.

These limitations highlight the areas where the study needs to be improved and can provide a direction for more in-depth research in the future

11. Future Work

Building on the limitations identified, several directions for future research can be pursued:

- **Expand Dataset:** Expanding the dataset to include a wider range of participants, with varying demographic backgrounds, and classroom settings to create more robust and generalized models.
- **Explore Advanced Methods:** Exploring more complex deep learning methods, such as recurrent neural networks (RNNs) or transformer architectures, to capture temporal dependencies and potentially achieve better results. Moreover, it is important to explore the effect of each modality on the prediction and this will allow to determine which features are the most significant.
- **Advanced Metrics and Robustness:** Exploring more robust evaluation metrics that can help us better analyze the effectiveness of the proposed methods. Moreover, it would be beneficial to measure the effectiveness of models on different type of datasets, such as cross-dataset validation.
- **Real-Time Implementation:** Developing and testing a real-time implementation of the model, including integrating this system with edge computing devices and validating the results through several classroom settings, as suggested by Soloviev [3].
- **Privacy**: Exploring privacy-preserving methodologies and data anonymization techniques to ensure that privacy is protected for real classroom implementations.
- **Multi-Modal Fusion**: Investigating different data fusion techniques and attention mechanisms to improve performance through feature-level fusion, decision-level fusion, and other advanced fusion methods.
- **Intervention strategies**: Develop intervention strategies to support instructors in modulating students' engagement based on real-time data, as suggested by Carroll et al. [2].

These directions will contribute to building a more robust, accurate, and reliable automatic student engagement assessment system. Future studies should also investigate the influence of various external factors, such as room temperature and air quality, on learning engagement.

REFERENCES

[1] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, and Flora D. Salim. 2020. n-Gage: Predicting in-class Emotional, Behavioural and Cognitive Engagement in the Wild. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 4, 3, Article 79

[2] M. Carroll, M. Ruble, M. Dranias, S. Rebensky, M. Chaparro, J. Chiang, and B. Winslow, "Automatic Detection of Learner Engagement Using Machine Learning and Wearable Sensors," *Journal of Behavioral and Brain Science*, vol. 10, no. 3, pp. 165-178, 2020.

[3] V. Soloviev, "Machine Learning Approach for Student Engagement Automatic Recognition from Facial Expressions," *Scientific Publications of the State University of Novi Pazar, Ser. A: Appl. Math. Inform. And Mech*, vol. 10, no. 2, pp. 79-86, 2018.