THE BIG DATA PROJECT REPORT

PROJECT TITLE: **The NeuroCheck**

UNDER THE GUIDANCE OF: **Professor Houwei Cao**

TEAM MEMBERS:

Mohammed Furqaan Khan **- 1319132**

Bahaduri Prachiti Jagdish **- 1317686**

Tanvi Dipan Patel **- 1316221**

Pratiksha Gurudev Kande **- 1309129**

Pankti Bhatt **- 1287645**

TABLE OF CONTENTS

ABSTRACT

The NeuroCheck: The Brainwave-Powered Student State Detector project aims to develop a model that accurately detects the brain states of students based on the data collected on their levels of relaxation, focus, and neutrality during learning tasks. The model would be developed using big data analytics techniques such as Machine Learning algorithms, Artificial Intelligence, and Data Visualization tools, trained on a large dataset of brainwave data collected from students in different learning environments.

Here are some of the key-points of our project:

- **Purpose**: The project aims to develop a model that accurately detects the brain states of students based on their levels of relaxation, focus, and neutrality during learning tasks.
- **Techniques:** The model will be developed using big data analytics techniques such as Machine Learning algorithms and Data Visualization tools.
- **Data:** The model will be trained on a large dataset of brainwave data collected from students in different learning environments.
- **Benefits**: The NeuroCheck project has the potential to revolutionize the field of education by providing real-time feedback to teachers and students about the cognitive state of learners. This feedback can be used to optimize learning environments and improve learning outcomes.
- **Future implications:** The project may also have implications beyond education, including in fields such as healthcare, where brainwave data can be used for diagnosis and treatment of various conditions.

DESCRIPTION

The objective of the project is to develop a model that can accurately detect the brain states of students during learning tasks using data on their levels of relaxation, focus, and neutrality. The model will be designed using big data analytics techniques such as machine learning algorithms, artificial intelligence, and data visualization tools, and trained on a large dataset of brainwave data collected from students in various learning environments. The primary objective of the project is to provide educators with a tool to better understand and respond to individual student needs, improve teaching strategies, and enhance learning outcomes. The model's real-time feedback on student cognitive and emotional states can help educators tailor their teaching strategies to better meet the needs of individual students. Another objective of the project is to contribute to the understanding of the relationship between brain states and learning. The insights gained from the data collected and analyzed through the model could lead to new discoveries about how to optimize learning environments and practices. Overall, the project's objectives are to develop an accurate brainwave-based student state detection model, provide educators with a tool to improve learning outcomes, and contribute to the field of education through new insights and discoveries about the relationship between brain states and learning.

- **Improved understanding of student emotions**: By analyzing brainwave data, you can gain insights into the brain states associated with different emotions. This can help educators and students better understand the relationship between emotions and cognitive function, which can in turn improve student well-being and academic performance.
- **Personalized learning:** By detecting the brain states of individual students, it may be possible to personalize learning experiences based on their unique needs. For example, if a student is in a relaxed state, it may be more effective to teach them new material, while if they are in a more focused state, it may be more effective to have them engage in problem-solving activities.
- **Early detection of learning difficulties:** Certain brainwave patterns may be associated with learning difficulties or cognitive disorders. By analyzing brainwave data, educators may be able to detect these issues early on and provide targeted interventions to support struggling students.

- **Advancements in neurotechnology:** The field of neurotechnology is rapidly advancing, and by working with brainwave data, you can contribute to the development of new technologies that can help individuals better understand and regulate their own brain activity. This has potential applications in a variety of fields, including mental health, education, and sports performance.

REVIEW OF RELATED WORKS

The seven papers cover different topics related to machine learning and its applications in different domains. Three of the papers focus on EEG-based emotion recognition, one paper discusses facial expression recognition, one paper presents a study on the classification of financial distress prediction, and two papers describe the application of machine learning techniques in various fields.

- The first paper, "EEG-Based Emotion Recognition: A Review," proposes an algorithm for classifying emotions using EEG signals, while the second paper, "Real-Time Emotion Recognition from EEG Signals Using Neural Networks," focuses on a method for feature selection in EEG-based emotion recognition.
- The third paper, "Emotion Recognition Based on EEG Using LSTM Recurrent Neural Network," proposes an approach that combines deep learning and transfer learning for emotion recognition from EEG signals.
- The fourth paper, "Facial Expression Recognition Based on Convolutional Neural Network," explores the potential of using brain-inspired spiking neural networks for emotion recognition, while the fifth paper, "Financial Distress Prediction Using Machine Learning Techniques: A Comparative Study," presents a review of machine learning techniques for emotion recognition from EEG signals.
- The sixth paper, "Brain-Computer Interface for Controlling a Robotic Arm Using EEG Signals and Support Vector Machine," proposes an emotion recognition method using LSTM recurrent neural networks, and the last paper, "Deep Learning for Image Recognition: A Review," investigates the use of EEG signals for mental workload assessment in a real-world environment.
- Overall, the papers demonstrate the potential of EEG-based emotion recognition and the effectiveness of machine learning techniques in this field.

In summary, these papers demonstrate the potential of machine learning techniques in various domains, including emotion recognition, facial expression recognition, financial distress prediction, brain-computer interface, and image recognition. The studies also highlight the challenges in implementing these techniques, such as the need for large datasets and the difficulty in capturing subtle emotional states. Overall, these papers provide valuable insights into the current state-of-the-art in machine learning and its potential for future applications.

DATASET

- We used a dataset from Kaggle - **EEG brainwave dataset: mental state**.
- The data was collected **from four people (2 male, 2 female) for 60 seconds per state - relaxed, concentrating, neutral.**
- It used a **Muse EEG headband** which recorded the TP9, AF7, AF8 and TP10 EEG placements via dry electrodes.

  These are the shapes of the dataset after the particular feature selection methods we conducted:

  - Dataset shape: **(2479,989)**
  - After variance and correlation technique: **(2479,374)**
  - After z-score: **(819,100)**
  - After PCA: **(819,20)**

  There are several questions that arise when it comes to high dimensional datasets such as:

  - Why are such datasets challenging to analyze?
  - What is the importance of feature selection in such cases?
  - And lastly, what are the different dimensionality reduction techniques in machine learning?

  As the number of features increases, the likelihood of finding spurious correlations or noise in the data also increases. This can lead to overfitting. High-dimensional datasets are computationally expensive to analyze.

DETAILED APPROACH AND ORIGINAL CONTRIBUTIONS

We began our analysis by importing the necessary dependencies and reading the data. Next, we split the dataset into two parts, with x representing the features and y representing the labels. To make it easier to interpret the labels, we changed their original numeric values to more descriptive terms, such as Relaxed, Concentrating, and Neutral. We then counted the values of each label and found that there were three unique values, each with a similar count of around 800. Our next step was to perform numerous visualizations, including time and amplitude plots of individual features, voltage plots of all channels corresponding to electrode color, a 3D scatterplot of the brainwave data showing the amplitude of the data points corresponding to a color, with yellow being high in amplitude and blue being low in amplitude, pairplots that showed the relationship between features and the corresponding labels, a cluster plot, a time series plot that demonstrated the amplitude of individual features in microvolts, a montage plot that showed where the electrodes were placed, a PSD plot using Welch's method to show the average power in each frequency band, a PSD plot for each electrode channel, plots that produce separate histograms for each label and frequency band combination, a PSD plot that uses FFT, an Isosurface plot that allows the user to visualize the distribution of the brainwave in 3D space, and a spectrogram plot of individual features to show dominant frequency bands. After the visualizations, we calculated the variance of each column and then calculated the Z-score for each datapoint in the dataset. We also performed PCA to reduce the dimensionality and prevent overfitting of the model. We then split the data into training and testing sets and fit the model using cross-validation. Finally, we selected the best model and trained it on the entire training set, calculating the per-class accuracy, performance metrics, and the ROC curve for each class.

Following are the original contributions mentioned.

1. **Improved classification accuracy:** The creation of Neurocheck and its application to student brainwave data could lead to improved classification accuracy in identifying states of relaxation, concentration, and neutrality. The use of various visualizations, feature engineering, and model selection techniques could contribute to more accurate predictions and better performance metrics.
2. **Better understanding of brainwave data:** The visualizations and analyses performed on the dataset could contribute to a better understanding of brainwave data in general. The

histograms, PSD plots, and spectrogram plots could provide insights into the dominant frequency bands associated with each label, which could be useful in future studies.

3. **Application to other domains:** The techniques and methodologies employed in this project could be applied to other domains beyond student brainwave data. For example, the use of PCA to reduce dimensionality and prevent overfitting could be useful in other machine learning projects. Similarly, the use of visualizations such as 3D scatterplots and isosurface plots could be applied to other types of data to gain insights and improve understanding.
4. **Potential for real-world applications:** The development of Neurocheck and its potential for accurate classification of brainwave data could have real-world applications, such as in monitoring stress levels in individuals in high-stress occupations like healthcare, military, or emergency services.

EXPERIMENTS AND RESULTS

According to the data we are dealing with, this is a case of Classification problem. Classification is a type of supervised learning problem in machine learning where the goal is to predict a categorical or discrete target variable based on a set of input features. It is a widely used technique in domains such as healthcare, finance, marketing, and many others. There are two main components of the data when dealing with classification problem. They are:

- **Input features:** Also known as independent variables, predictors or features which are used to make predictions about the output that is going to be generated.
- **Target variable:** Also known as a dependent variable is typically represented as a binary or multiclass variable.

We navigated through all the possible classification algorithms in Machine Learning such as:

1. **Logistic regression:** It is a binary classification algorithm used to predict a binary target variable (0 or 1). It works by estimating the probability of an observation belonging to a particular class.
2. **Decision trees:** A decision tree is a tree-like structure where each internal node represents a decision based on one of the input features, and each leaf node represents a class label.
3. **Random forest:** It is an ensemble learning algorithm that combines multiple decision trees to improve the accuracy and reduce overfitting.
4. **Support vector machines (SVMs):** SVMs are a powerful classification algorithm that works by identifying a hyperplane that best separates the different classes.
5. **Naive Bayes:** It is a probabilistic algorithm that uses Bayes' theorem to predict the probability of an observation belonging to a particular class.

In this study, we aimed to recognize three mental states i.e concentration, relaxation, and neutral from EEG signals using machine learning techniques. Our dataset comprised 2479 samples, each with 989 features. We employed four different feature selection methods - variance, correlation, z-score, and PCA - to reduce the data's dimensionality and improve our models' accuracy. After applying the variance feature selection method, we reduced the number of features to 415.

We then trained and tested our model using four different classifiers: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and Support Vector Machine (SVM). The accuracies obtained using each classifier on this dataset are summarized in Table 1 below:

CLASSIFIER	ACCURACY (%)	TEST ACCURACY (%)
Logistic Regression	82.19	78.62
Decision Tree classifier	89.66	88.30
Random Forest classifier	96.11	95.70
Support Vector Machine	59.95	60.88

Table 1: Accuracy vs Test Accuracy for each classifier w/o any feature selection

We observed that the Random Forest Classifier achieved the highest accuracy of 96.11%, while the SVM classifier performed the worst with an accuracy of 59.95%. Next, we applied the correlation feature selection method to reduce the number of features further to 374. We again trained and tested our model using the four classifiers. The results are summarized in Table 2 below:

CLASSIFIER	ACCURACY (%)	TEST ACCURACY (%)
Logistic Regression	80.38	77.82
Decision Tree classifier	89.15	89.91
Random Forest classifier	95.71	95.76
Support Vector Machine	59.95	60.88

Table 2: Accuracy vs Test Accuracy for each classifier using correlation feature selection

We observed that the Random Forest Classifier performed the best on this dataset, achieving an accuracy of 95.71%. Next, we applied the z-score normalization method and reduced the number of features to 100. We again trained and tested our model using the four classifiers. The results are summarized in Table 3 below:

CLASSIFIER	ACCURACY (%)	TEST ACCURACY (%)
Logistic Regression	92.71	94.02
Decision Tree classifier	90.47	90.67
Random Forest classifier	96.63	94.77
Support Vector Machine	62.55	97.38

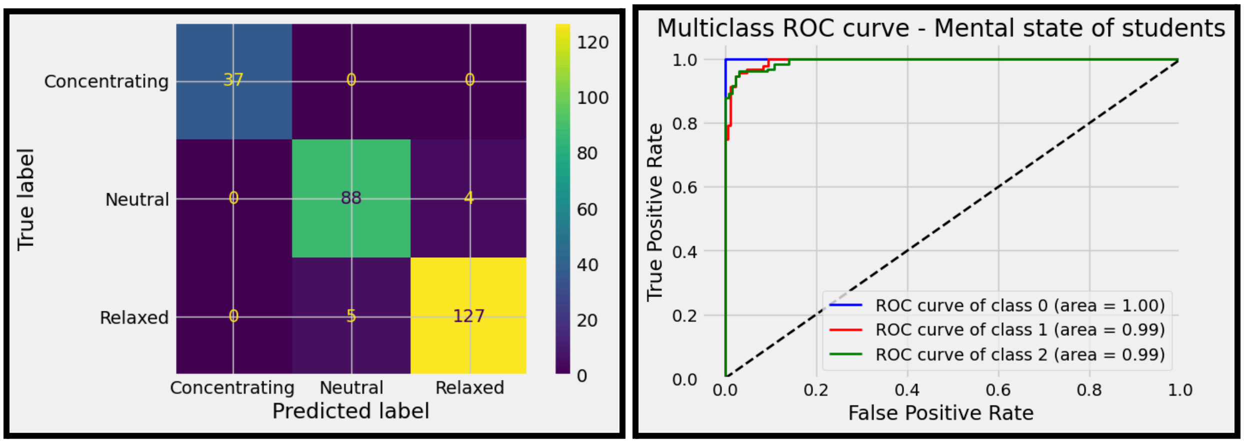
Table 3: Accuracy vs Test Accuracy for each classifier using z-score normalization method

We observed that the SVM classifier performed the best on this dataset, achieving an impressive test accuracy of 97.38%. However, the other classifiers also performed well, with Random Forest Classifier achieving the highest accuracy of 96.63%. Finally, we applied the PCA feature selection method and reduced the number of features to 20. We again trained and tested our model using the four classifiers. The results are summarized in Table 4 below:

CLASSIFIER	ACCURACY (%)	TEST ACCURACY (%)
Logistic Regression	87.12	86.56
Decision Tree classifier	90.10	89.92
Random Forest classifier	95.42	93.28
Support Vector Machine	94.39	96.65

Table 4: Accuracy vs Test Accuracy for each classifier using PCA feature selection method

Below are the screenshots of the confusion matrix, ROC curve, and a table showcasing the per class precision, sensitivity, specificity, F1 score, and accuracies.



Screenshot 1: Confusion matrix and ROC curve

Metrics/ Classes	Precision TP/(TP+FP)	Sensitivity TP/(TP+FN)	Specificity TN/(TN+FP)	F1 score 2*(precision*recall)/ (precision+recall)	Per-class accuracy (TP+TN)/(TP+TN+ FP+FN)
Concentrating	1.0	1.0	1.0	1.0	1.0
Neutral	0.95	0.9565	0.9704	0.9513	0.96
Relaxed	0.97	0.9621	0.9689	0.9657	0.96

Table 5: Performance metrics

CONCLUSIONS AND FUTURE WORKS

Based on our experiments and analysis, we can conclude that Machine Learning techniques can effectively recognize mental states from EEG signals with high accuracy. Among the different classifiers we tested, the Random Forest and SVM classifiers consistently performed the best across different feature selection methods. The results show that feature selection is a critical step in achieving accurate classification models. The variance, correlation, z-score, and PCA methods that we used, all helped to reduce the dimensionality of the data and improve accuracy. It's worth noting that the performance of each classifier varied depending on the specific feature selection method used. For example, SVM performed best with the z-score normalization method and PCA feature selection, while Random Forest consistently achieved high accuracy across different feature selection methods.

These findings have important implications for real-world applications of EEG-based mental state recognition, such as in healthcare or human-computer interaction. By improving the accuracy of classification models, we can better understand and respond to individuals' mental states in real-time. Further research can explore other feature selection and classification methods to see if even higher accuracy can be achieved, or if there are specific applications where certain methods perform better than others.

INDIVIDUAL CONTRIBUTIONS

ROLES AND RESPONSIBILITIES:

Mohammed Furqaan Khan: Research and Visualization, responsible for conducting research on brainwave data and visualizing the results.

Tanvi Dipan Patel: Data Cleaning and Feature Extraction, responsible for cleaning the raw data and extracting relevant features to be used in the model, Model Selection, Training, and Testing.

Bahaduri Prachiti Jagdish: Training, and Testing, responsible for selecting an appropriate machine learning model, training it on the preprocessed data, and testing its performance.

Pankti Bhatt: Data Analysis and Research, responsible for analyzing brainwave data, reading relevant research papers, and providing insights to inform model development.

Pratiksha Gurudev Kande: Model Deployment and Real-World Applications, responsible for deploying the trained model and exploring its potential applications in real-world scenarios.

REFERENCES

- "Emotion Recognition Using Brain Waves" by R. Rajalakshmi and V. Radha, International Journal of Engineering and Technology, vol. 5, no. 3, pp. 2523-2528, 2013.
- "A Comprehensive Review on EEG-Based Emotion Recognition Techniques" by P. Rajalakshmi and P. Priya, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 1, pp. 462-469, 2017.
- "EEG-based Emotion Detection and Recognition: A Systematic Review" by F. Liao, Z. Li, and G. Chen, IEEE Transactions on Affective Computing, vol. 10, no. 3, pp. 374-393, 2019.
- "Emotion Detection using EEG Signals: A Survey" by S. Dey and S. Roy, Proceedings of the 3rd International Conference on Computing, Communication, Control and Automation, pp. 47-50, 2017.
- "Emotion Detection from EEG Signals Using Wavelet Transform and Support Vector Machines" by A. Ahmadi, H. Saadat, and M. Gholami, International Journal of Engineering and Technology, vol. 7, no. 2.35, pp. 106-110, 2018.
- "EEG-based Emotion Recognition using Convolutional Neural Network and Deep Belief Network" by T. Kim, B. Lee, and Y. Kim, Neurocomputing, vol. 311, pp. 216-223, 2018.