Machine learning algorithms for age prediction based on linear and non-linear parameters of electroencephalogram data

Dinmukhamed Sadibekov[1], *Ruslan* Zhulduzbaev[1], *Nurbek* Merkibek[1], *Manzura* Zholdassova[2], *Altyngul* Kamzanova[3], *Gaukhar* Datkhabayeva[3], and *Almira* Kustubayeva[2,3*]

[1]Kazakh British Technical University, Almaty, 050000, Kazakhstan
[2]Brain Institute, al-Farabi Kazakh National University, Almaty, 050040, Kazakhstan
[3]Department of Biophysics, Biomedicine, and Neuroscience, al-Farabi Kazakh National University, Almaty, 050040, Kazakhstan

Abstract. Gaining insights into cognitive and behavioral changes during childhood and adolescence requires a fundamental understanding of the developmental trajectory of the human brain. This research aimed to predict the age of children using linear and non-linear measures of baseline electroencephalogram (EEG) data. EEG is a method that records the electrical activity of the brain, providing valuable insights into its functioning. Participants were 182 children between 7 to 20 years old. Peak alpha and entropy were correlated with age. Various machine learning models were implemented, with Decision Trees yielding the best results. The Decision Trees model achieved strong correlation between predicted and actual age. The study demonstrated the stability of age prediction error over time, suggesting individual brain maturational levels. The findings highlight the potential of EEG data for accurate age prediction, providing insights into brain maturation patterns. This research contributes to tracking neurodevelopment and understanding brain function across age groups, including typically developing children.

1 Introduction

Understanding the developmental trajectory of the human brain is crucial for gaining insights into cognitive and behavioral changes during childhood and adolescence. Electroencephalogram (EEG) recordings provide a non-invasive and cost-effective measure of brain activity, offering potential for age prediction in children and adolescents. This study aims to investigate the use of EEG data for predicting the age of individuals aged 7 to 20 years old.

The analysis includes two key measures: peak alpha and entropy. Peak alpha represents the dominant frequency within the alpha band (8-12 Hz) and provides insights into the strength of alpha oscillations. Entropy, on the other hand, captures the complexity and predictability of EEG signals.

*Corresponding author: almkusto@kaznu.kz.

Linear measure

Chiang et al. (2011) examined age trends and sex differences in alpha rhythms, including split alpha peaks. Their study shed light on the variations in alpha oscillations across different age groups and genders. Edgar et al. (2015) investigated resting-state alpha activity in individuals with autism spectrum disorder and its associations with thalamic volume, providing insights into the potential role of alpha oscillations in this condition.

Furthermore, the APF has been linked to cognitive abilities (Kamzanova et al., 2012, 2020, Kustubayeva et al., 2013, 2022). Grandy et al. (2013) found that individual APF was related to latent factors of general cognitive abilities, indicating a potential association between alpha oscillations and cognitive performance. Clark et al. (2004) demonstrated that spontaneous APF predicted working memory performance across different age groups, highlighting the relevance of alpha rhythms in cognitive functioning.

By considering the alpha peak frequency as a measure, this study extends the analysis beyond linear metrics and explores the frequency-specific characteristics of EEG signals. The statistical analysis aims to uncover age-related patterns and assess the potential of the alpha peak frequency for age prediction, cognitive abilities, and neurodevelopmental processes.

Non-linear measure

Higher entropy indicates more complex, less predictable signals. Entropy has been utilized in various studies to investigate brain connectivity and distinguish between different neurological conditions.

Entropy analysis has also been applied to seizure prediction and classification. Li et al. (2007) conducted predictability analysis of absence seizures using permutation entropy. Piryatinska et al. (2017) focused on binary classification of multichannel EEG records based on the complexity of continuous vector functions.

Previous research has examined the relationship between EEG measures and age-related patterns. Studies by Anderson et al. (2001), Barriga-Paulino et al. (2011), and Benniger et al. (1984) have investigated developmental changes in EEG rhythms and their correlation with age in children and young adults. These studies contribute to our understanding of brain maturation during different developmental stages.

In this paper, various popular machine learning algorithms, including Linear Regression, Support Vector Machines (SVM), Decision Trees, Naive Bayes, Gradient Boosting Machines, Random Forest, and K-Nearest Neighbors (KNN) were utilized to explore the relationship between age and peak alpha across channel networks.

These algorithms are a general choice to explore high-dimensional data and analyze the multidimensional nature of EEG channel networks with relation to age. Some like Linear Regression can be a good fit for observing linear relationship between variables, while others like SVM and Decision Trees can capture more complex interactions between features, making them a valuable tool for classification tasks. The latter, for example, constructs a model by recursively partitioning the data based on feature values, resulting in a tree-like structure that facilitates interpretation.

Overall, machine learning algorithms, including the Decision Trees algorithm, have demonstrated success in various domains. According to Bishop (2006), the effectiveness of Decision Trees and their interpretability in capturing patterns in complex datasets is well recognized and has been widely studied.

Thus, by extracting peak alpha and entropy measures from EEG recordings and employing the machine learning algorithms, this research aims to track the neurodevelopmental trajectories in healthy youth. The analysis seeks to uncover age-related patterns and assess the potential of EEG data for accurate age prediction.

2 Materials and Methods

2.1 Participants

For this study one hundred and eighty participants aging between 7 to 20 years old (mean age 13.87 years, standard deviation 4.01, 88 females, 92 male) were recruited through social network advertisement from Almaty city area and Almaty city universities. All participants were informed and signed consent, adhering to the ethical guidelines established for research involving human subjects. The study protocol received formal approval from the Ethics Committee of the Al Farabi Kazakh National University, ensuring ethical compliance during the research process. Data was collected during the first half of day (between 08.00 and 13.00 hours).

2.2 EEG recordings

Baseline EEG recording was done using an eego™ mylab (ANT Neuro, Welbergweg, Netherlands) according to the 10-20% international recording system with 64 electrodes in the following situations: open eyes (2 min); closed eyes (2 min); completing cognitive tasks (70 min). Sampling rate during recording was 4096 Hz and impedance was below 5 kOhm during the experiment. This article presents results of EEG data during the closed eyes only.

2.3 EEG pre-processing

For this stage, EEGLAB toolbox based on MATLAB software was used to implement preprocessing pipeline, which encompassed resampling, filtering, component analysis, feature extraction and data segmentation. Primarily, the goal was to gather appropriate and meaningful data to train machine learning models further.

First, the data was downsampled from 4096 Hz to 512 Hz to achieve balance between data size and computational efficiency. After that, bandpass filtering was applied, resulting in finite impulse response (FIR) filtering between 4–30 Hz, which included the theta, alpha and beta band activities. Finally, artifact removal with runica ICA algorithm (Independent Component Analysis) served to identify and subsequently remove independent components associated with eye movements and other frequently encountered artifacts. In cases where necessary, manual artifact removal was also employed.

2.4 Statistical analysis

Following the preprocessing stage, Python with MNE library was used to further group each subject data into five predefined channel networks:

Parietal and occipital channels (P7, P3, Pz, P4, P8, POz, O1, O2, P5, P1, P6, PO5, PO3, PO4, PO6, PO7, PO8);

Frontal and central channels (F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, C3, Cz, C4, F5, F1, F2, FC3, FCz, FC4, C5, C1, C2);

Right brain channels (F2, F4, F8, FC2, FC6, C4, T8, M2, CP2, CP6, P4, P8, O2, AF4, AF8, F6, FC4, C2, C6, CP4, P2, P6, PO4, PO6, FT8, TP8, PO8);

Left brain channels (Fp1, F7, F3, FC5, FC1, M1, T7, C3, CP5, CP1, P7, P3, O1, AF7, AF3, F5, F1, FC3, C5, C1, CP3, P5, P1, PO5, PO3, FT7, TP7, PO7);

Mid-line channels (Fpz, Fz, Cz, Pz, POz, FCz, Oz).

Linear value

Power spectral density for each network was computed using MNE's compute_psd method in the range 8-12 Hz. Index which corresponds to the maximum value of power spectral density was used to extract the peak alpha.

Individual peak alpha values were correlated with age using Pearson coefficients.

Non-Linear metrics

The signal was filtered into theta, alpha, and beta bands. Entropy was calculated using the MNE's mne_features.univariate.compute_samp_entropy function, specifically utilizing Sample Entropy (SampEn) per network (Richman et al., 2000).

2.5 Machine learning models

For this stage, Python and sklearn library toolkit was used. Multiple machine learning models were utilized on the dataset, such as Linear Regression, Support Vector Machines (SVM), Decision Trees, Naive Bayes, Gradient Boosting Machines, Random Forest, and K-Nearest Neighbors (KNN). A variety of models was used with peak alpha values in our study.

Multiple evaluation metrics were employed to gauge the efficacy of the predictive model in this investigation such as Mean Squared Error (MSE), Mean Absolute Error (MAE), the coefficient of determination (R^2) and the correlation coefficient (r).

3 Results and Discussion

3.1 Dynamical changes of general peak alpha with age

Observations indicate that there is a positive association between age and peak alpha frequency in the EEG baseline signal for children aged 7 to 20 years. As children mature, their brains tend to exhibit higher dominant frequencies within the alpha band (8-12 Hz).

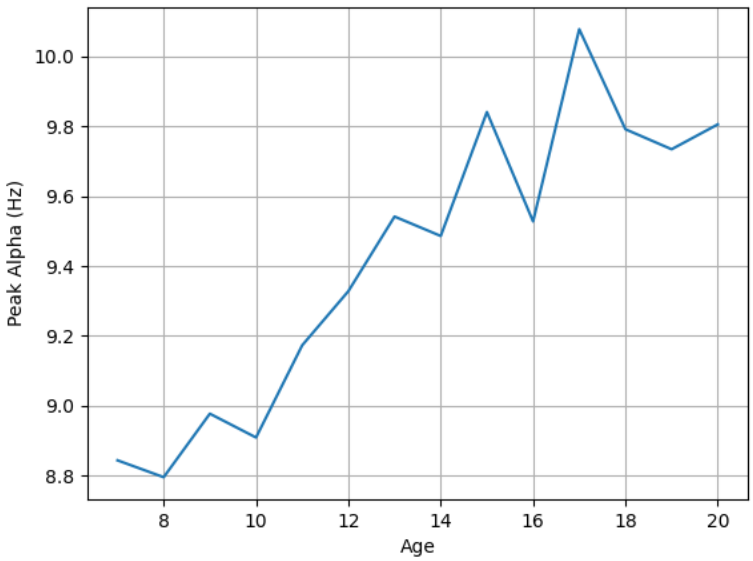

Fig. 1. Peak alpha tendency with age. X-coordinate for age, Y-coordinate for peak alpha (Hz).

3.2 Correlation between alpha peaks in different networks and age

Significant positive correlations were observed between age and peak alpha frequency in the Parietal occipital network, Frontal central network, Right hemisphere, Left hemisphere, and Midline regions (see Table 1).

Table 1. Correlations between peak alpha and age.

	Parietal_ occipital	Frontal_ central	Right_ hemisphere	Left_ hemisphere	Midline
Pearson correlation	0,291**	0,384**	0,389**	0,362**	0,343**
P-value	0,000	0,000	0,000	0,000	0,000

Assuming that correlation is significant at the 0,01 level (2-tailed), this metric was chosen to be tested with machine learning algorithms.

3.3 Correlation between alpha peaks in different networks and age

Investigation of the relationship between age and mean entropy, showed only one significant correlation with Parietal occipital network (see Table 2).

Table 2. Correlations between mean entropy and age.

	Parietal_ occipital	Frontal_ central	Right_ hemisphere	Left_ hemisphere	Midline
Pearson correlation	0,147**	0,044	0,084	0,094	0,038
P-value	0,048	0,558	0,262	0,209	0,610

3.4 Age prediction based on ML model

Age prediction on peak alpha values across networks using each model is presented in Table 3.

Table 3. Machine Learning model comparison table.

Model	**R^2**	**Correlation (r)**	**MSE**	**MAE**
Linear Regression	0.159	0.399	13.230	3.034
SVR	0.103	0.389	14.115	2.969
Decision Trees	0.748	0.865	3.960	0.981
K-Nearest Neighbors (KNN)	0.252	0.516	11.764	2.789
Naive Bayes	0.170	-	23.714	3.626
Random Forest	0.650	-	5.497	1.766
Gradient Boosting Machines	0.566	-	6.830	2.010

Note: Correlation coefficient (r) is not available for Naive Bayes, Random Forest, and Gradient Boosting Machines.

In addition to quantitative metrics, the analysis was augmented by generating scatter plots that visually depicted the predicted values in relation to the actual values (Figure 2).

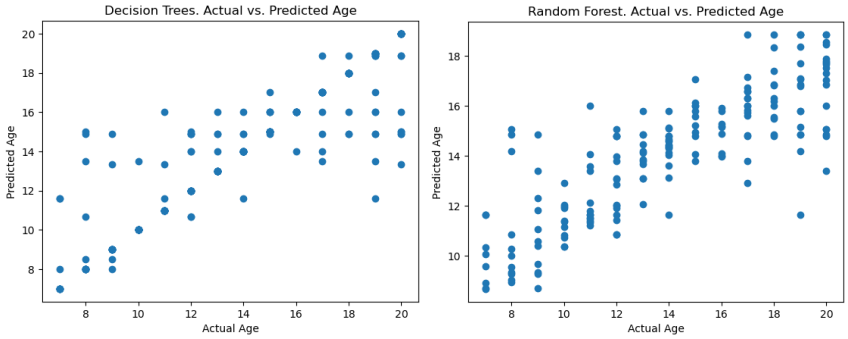

Fig. 2. Actual vs. Predicted Age results in Decision Trees and Random Forest models.

The Decision Trees model demonstrated superior performance compared to other machine learning models, with a coefficient of determination (R^2) of 0.748, a correlation coefficient (r) of 0.865 between Actual Age and Predicted Age, and low Mean Squared Error (MSE) and Mean Absolute Error (MAE) values. This highlights the efficacy of the Decision Trees algorithm in accurately forecasting age based on the provided dataset.

Overall this result is particularly relevant as it explores the potential application of machine learning algorithms in predicting age-related changes. Compared with the study conducted by Zoubi et al. (2018), which primarily focused on the support vector regression (SVR) model, there are interesting insights into the performance of different machine learning models for age prediction. Zoubi et al. (2018) achieved promising results with the SVR model, demonstrating its effectiveness in accurately estimating age based on brain features. Their study reported a coefficient of determination (R^2) of 0.37, a mean absolute error (MAE) of 6.87 years, and a root mean squared error (RMSE) of 8.46 years. While Zoubi et al. (2018) demonstrated the potential of the SVR model, current research suggests that the Decision Trees model outperforms other machine learning models in accurately forecasting age in specific research contexts.

Current research could further be improved to study connectivity between brain networks. According to Farber et al. (2014), age-related differences in connectivity during the preparation for recognizing fragmented images, provide valuable insights into the neural mechanisms underlying attention and working memory. Their findings indicated variations in connectivity patterns between adults and children, suggesting age-related differences in the functional organization of the brain during focused attention.

4 Conclusion

In conclusion, the current study explored machine learning algorithms to predict age using different EEG parameters. The most informative was peak alpha relationship with age with Decision Trees model yielding most significant results.

While entropy was initially considered as a potential feature for age prediction, its correlation with age was found to be weak. Consequently, peak alpha, representing the dominant frequency within the alpha band, was selected as the primary metric for predicting age.

The Decision Trees model achieved a high coefficient of determination (R^2) of 0.748, a strong correlation coefficient (r) of 0.865 between Actual Age and Predicted Age, and low Mean Squared Error (MSE) and Mean Absolute Error (MAE) values. These findings highlight the efficacy of the Decision Trees algorithm in capturing age-related patterns in the provided dataset.

Moving forward, it is important to acknowledge that these results are specific to the dataset used in this study. Further validation and replication using diverse datasets with larger sample sizes are necessary to generalize the findings. Additionally, exploring alternative machine learning algorithms and incorporating other relevant features may provide further insights into the prediction of age based on EEG data.

Overall, this study contributes to the growing body of research on EEG-based age prediction and highlights the potential of peak alpha measures in tracking neurodevelopment in typically developing individuals and those with neurodevelopmental disorders.

Acknowledgement

Research was supported by research grant from Ministry of Education and Science of Kazakhstan to AMK (grant AP08856595).

Author's contribution

Conceptualization: A.Ku.; Methodology: A.Ku., M.Z., R.Z and D.S.; Investigation and Resources: A.Ku., M.Z., A.Ka., G.D.; Review & Editing: A.Ku.; Writing: D.S and R.Z.; Formal Analysis & Machine Learning: D.S, M.N., R.Z.; Supervision: A.Ku.

References

1. A.K.I. Chiang, C.J. Rennie, P.A. Robinson, S.J. van Albada, & C.C. Kerr, Age trends and sex differences of alpha rhythms including split alpha peaks. Clin. Neurophysiol. **122,** 1505–1517 (2011). https://doi.org/10.1016/j.clinph.2011.01.040
2. J. Edgar, K. Heiken, Y. Chen, & J. Herrington, Resting-state alpha in autism spectrum disorder and alpha associations with thalamic volume. J. Autism Dev. Disord. **45**, 795–804 (2015). https://doi.org/10.1007/s10803-014-2236-1
3. A. Kamzanova, A. Kustubayeva, G. Matthews, Diagnostic monitoring of vigilance decrement using EEG workload indices. Proc. Hum. Factors Ergon. Soc. 203 – 207 (2012). https://doi.org/10.1177/0018720814526617
4. A. Kamzanova, G. Matthews, A. Kustubayeva, EEG Coherence Metrics for Vigilance: Sensitivity to Workload, Time-on-Task, and Individual Differences. Appl. Psychophys. Biof. **45(3),** 183 – 194 (2020). https://doi.org/10.1007/s10484-020-09461-4
5. A.M. Kustubayeva, A. Tolegenova, G. Matthews, EEG-brain activity in different strategies of emotions' self-regulation: Suppression and reappraisal. Psikholog. Zh., **34(4),** 58 – 68 (2013).
6. A. Kustubayeva, M. Zholdassova, G. Borbassova, G. Matthews, Temporal changes in ERP amplitudes during sustained performance of the Attention Network Test. Int. J. Psychophysiol. **182,** 142 – 158 (2022) https://doi.org/10.1016/j.ijpsycho.2022.10.006
7. T.H. Grandy, M. Werkle-Bergner, C. Chicherio, M. Lövdén, F. Schmiedek, & U. Lindenberger, Individual alpha peak frequency is related to latent factors of general cognitive abilities. Neuroimage, **79,** 10–18 (2013). https://doi.org/10.1016/j.neuroimage.2013.04.059
8. C. Richard Clark, M.D. Veltmeyer, R.J. Hamilton, E. Simms, R. Paul, D. Hermens, & E. Gordon, Spontaneous alpha peak frequency predicts working memory performance across the age span. Int. J. Psychophysiol. **53,** 1–9 (2004). https://doi.org/10.1016/j.ijpsycho.2003.12.011

9. X. Li, G. Ouyang, D.A. Richards, Predictability analysis of absence seizures with permutation entropy. Epilepsy Res. **73(3),** 232-241 (2007). https://doi.org/10.1016/j.eplepsyres.2007.08.002
10. A. Piryatinska, B. Darkhovsky, A. Kaplan, Binary classification of multichannel-EEG records based on the is an element-of-complexity of continuous vector functions. Comput. Methods Programs Biomed. **152,** 131-139 (2017). https://doi.org/10.1016/j.cmpb.2017.09.001
11. V. Anderson, E. Northam, J. Hendy, & J. Wrennall, Developmental neuropsychology: A clinical approach. (Routledge, London, 2018)
12. C. Barriga-Paulino, A. Flores, & C. Gomez, Developmental changes in the EEG rhythms of children and young adults analyzed by means of correlational, brain topography, and principal component analysis. J. Psychophysiol. **25(3),** 143–158 (2011). https://doi.org/10.1027/0269-8803/a000052
13. C. Benniger, P. Matthis, & D. Scheffner, EEG development of healthy boys and girls: Results of a longitudinal study. Electroencephalogr Clin Neurophysiol, **57(1),** 1–12 (1984). https://doi.org/10.1016/0013-4694(84)90002-6
14. C.M. Bishop, Pattern recognition and machine learning. J. Chem. Inf. Model. **53,** 049901 (2006). https://doi.org/10.1117/1.2819119
15. O.A. Zoubi, C.K. Wong, R.T. Kuplicki, H. Yeh, A. Mayeli, H. Refai, et al. Predicting age from brain EEG signals—a machine learning approach. Front. Aging Neurosci. **10,** 184 (2018). https://doi.org/10.3389/fnagi.2018.00184
16. D.A. Farber, R.I. Machinskaya, A.V. Kurgansky, et al. Functional organization of the brain in the period of preparation for recognizing fragmented images in seven- to eight-year-old children and adults. Hum. Physiol. **40,** 475–482 (2014). https://doi.org/10.1134/S036211971405003X