Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 47 (2026), No. 5-B, pp. 2043–2052
DOI : 10.47974/JIOS-2295

Age prediction on EEG signals via hybrid feature engineering approach

Wanus Srimaharaj [†]
Department of Information Technology
The International College
Payap University
Chiang Mai 50000
Thailand

Neeta N. Thune *
Department of Electronics and Telecommunication Engineering
Marathwada Mitra Mandal's College of Engineering
Savitribai Phule Pune University (SPPU)
Pune 411052
Maharashtra
India

Supansa Chaising [§]
Sustainability and Entrepreneurship Research Center
School of Management
Mae Fah Luang University
Chiang Rai 57100
Thailand

A. B. Kanwade [‡]
Department of Electronics and Telecommunication Engineering
Marathwada Mitra Mandal's College of Engineering
Savitribai Phule Pune University (SPPU)
Pune 411052
Maharashtra
India

This research was conducted during the NREUP at Michigan State University and it was sponsored by NSF grants DMS-1156582 and DMS-1359016.

[†] *E-mail:* `wanus_s@payap.ac.th`
* *E-mail:* `neeta.thune@gmail.com (Corresponding Author)`
[§] *E-mail:* `supansa.cha@mfu.ac.th`
[‡] *E-mail:* `archanakanwade@mmcoe.edu.in`

G. S. Gawande [@]
Department of Electronics and Telecommunications Engineering
MIT-ADT University
Loni Kalbhor
Pune 412201
Maharashtra
India

Abstract

This study proposes a Gaussian Process Regression (GPR) model with a custom composite kernel for age prediction from normal electroencephalogram (EEG) signals. Using the TUH Abnormal EEG Corpus from the Temple University EEG dataset, the model combines Radial Basis Function (RBF) and Matérn kernels to capture both global and local variations in EEG features. The prediction framework demonstrates a mean absolute error of 5.252 years and shows higher accuracy for middle-aged individuals. The composite kernel enables flexible adaptation to EEG signal characteristics across age ranges. The approach highlights differences in prediction performance across age groups and suggests clinical applicability for EEG-based age estimation. The findings demonstrate the potential of combining feature engineering and kernel-based learning for age prediction from normal EEG signals.

Subject Classification: *Primary 93A30, Secondary 49K15*

Keywords: *Age prediction, Deep learning, Electroencephalogram, EEG, Signal processing.*

1. Introduction

Electroencephalography (EEG) has transitioned from a diagnostic tool to a data source for computational modeling. Recent methods focus on normal EEG signals for age estimation, highlighting the importance of effective feature extraction and selection to model the relationship between EEG patterns and chronological age [1]. Elimination of correlated features has been shown to enhance predictive accuracy, confirming EEG features as viable age indicators, despite varied reported performance metrics [1]. Algorithmic approaches provide more consistent age estimation than clinician judgment, while accounting for confounding variables like sex and body mass index [2, 3].

Deep learning and feature engineering have both been applied to improve EEG-based age prediction. Convolutional neural networks (CNNs), functional connectivity, and hybrid frameworks have achieved success in brain-computer interfaces and brain age modeling such as BrainAGE [2, 4, 6]. However, challenges remain due to the limited interpretability of deep learning outputs and difficulty in modeling age-related EEG dynamics. Some models perform close to random estimation, indicating limitations in generalizing across populations [5, 7].

[@] *E-mail:* `gawandeg25@gmail.com`

This study proposes a hybrid method integrating engineered features with Gaussian Process Regression (GPR), employing a composite kernel. The method enhances interpretability and adaptability compared to deep learning alone. GPR's probabilistic structure supports nonlinear modeling of multichannel EEG and outputs predictive uncertainty. Combining RBF and Matérn kernels enables the model to address both global trends and localized EEG variations. Normal EEG signals from the TUH Abnormal EEG Corpus were used due to dataset size and age coverage [8]. The method seeks to improve estimation accuracy and assess model certainty. It also examines prediction variance across age groups, particularly lower performance in younger (<20) and older (>70) individuals. The framework supports age prediction using structured features and multivariate kernels, aiming to improve clinical utility and model transparency.

2. Literature Reviews

The BrainAGE framework, originally designed for MRI, was adapted to EEG for estimating brain age gaps, achieving an R^2 of 0.37 using a stacked model on 468 participants from the Tulsa-1000 study [12]. EEG-based age prediction using deep convolutional networks and data augmentation improved performance by 13% on the TD-BRAIN dataset, indicating its clinical relevance [13]. Spatially-informed benchmarks on M/EEG from four cohorts also showed higher predictive accuracy [14].

Bidirectional LSTM models classified age and gender with 93.7% and 97.5% accuracy, identifying beta band features as essential for brain-computer interface applications [15]. EEGNet and conventional machine learning on multi-hospital EEG data emphasized EEG's role as a brain health biomarker [16]. Raw EEG trained with LSTM yielded 90% classification accuracy across six age groups [17].

A hybrid model using spectral and temporal features improved age prediction for individuals aged 30–50 years [18]. Deep learning models in children achieved low MAE (4.82 months) [19], and CNN-MLP models incorporating sex outperformed CNN-only classifiers for cognitive differentiation [20]. BLSTM models supported EEG's utility in brain maturation tracking across youth [21].

Despite these advances, generalizability, age coverage, and interpretability remain concerns. The current study addresses these by integrating feature engineering with a composite-kernel GPR model applied to normal EEG recordings from the TUH Abnormal EEG Corpus.

3. Methodology

3.1 *Generalized Dynamic Formula for Gaussian Process Regression*

The GPR formulation handle multiple variables and inputs by let $X = \{x_1, x_2, \ldots, x_N\}$ denote the input EEG dataset, where $x_i \in R^d$ is a d-dimensional feature vector, and $y = \{y_1, y_2, \ldots, y_N\}$ represent the corresponding target age values to predict the target y_* for a new input x_*.

3.2 *Prior Distribution*

A Gaussian process defines a prior distribution over functions $f(x)$ with added Gaussian noise and Gaussian process as Equation (1) and (2).

$$y = f(X) + \epsilon, \epsilon \sim N(0, \sigma^2 I) \tag{1}$$

$$f(x) \sim GP(m(x), k(x, x')) \tag{2}$$

Where $m(x)$ is the mean function (commonly zero) and $k(x,x')$ is the covariance function.

3.3 *Kernel (Covariance) Function*

A composite kernel is adopted as Equation (3).

$$k(x, x') = \theta_1 \exp\left(-\frac{||x-x'||^2}{2\theta_2^2}\right) + \theta_3\left(1 + \frac{\sqrt{5}||x-x\prime||}{\theta_4} + \frac{5||x-x\prime||^2}{3\theta_4^2}\right) exp\left(\frac{-5||x-x\prime||}{\theta_4}\right) \tag{3}$$

The first term is the RBF kernel, the second is the Matérn kernel ($v = \frac{5}{2}$). The hyperparameters θ_1, θ_2, θ_3, θ_4 define scale and smoothness.

3.4 *Joint Distribution and Inference*

Given a test point x_*, the joint distribution of y and $f_* = f(x_*)$ as Equation (4).

$$\left(\frac{y}{f_*}\right) \sim N\left(0, \left(\frac{K(X,X)+\theta_4^2 I}{K(x_*,X)} \frac{K(X,x_*)}{K(x_*,x_*)}\right)\right) \tag{4}$$

where $K(X, X)$ is the covariance matrix between training points, $K(X, x_*)$ is the covariance between the training points and the test point, $K(x_*, x_*)$ is the covariance between the test point and itself.

3.5 *Posterior Distribution*

The posterior mean μ_* and variance σ_*^2 are presented as Equation (5) and (6).

$$\mu_* = K(x_*, X)(K(X, X) + \sigma_n^2 I)^{-1} y \tag{5}$$

$$\sigma_*^2 = K(x_*, x_*) - K(x_*, X) K(X, X) + \sigma_n^2 I)^{-1} K(X, x_*) \tag{6}$$

This yields a predictive distribution for each new input, incorporating uncertainty.

3.6 *Dynamic Nature with Multiple Variables*

The kernel function $k(x, x')$ dynamically adapts based on the input features x. Each EEG signal channel (FP1, FP2, etc.) can be treated as one of the dimensions in the input vector x_i. The complexity of the inversion $(K(X, X) + \sigma_n^2 I)^{-1}$ *is* $O(N^3)$, where N is the number of data points for moderate-sized datasets as Equation (7).

$$f(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i) \tag{7}$$

where $k(x, x_i)$ is the kernel function that measures the similarity between the test point x and training points x_i to support multiple variables of EEG signal data.

3.8 *Data Descriptions*

The data utilized in this study is derived from the Temple University EEG Corpus, specifically a segment of the TUH Abnormal EEG Corpus [21]. While the primary intent of the original dataset was to differentiate between normal and abnormal EEG signals. Each file includes time-series data from multiple standard 10–20 system EEG channels. A typical recording contains signals from 12 channels including FP1-REF, FP2-REF, C3-REF, C4-REF, CZ-REF, F3-REF, F4-REF, F7-REF, F8-REF, O1-REF, O2-REF, and T3-REF. Each row corresponds to a 10 ms interval, resulting in 1000 data points for a 10-second segment. The dataset comprises 1297 patients with verified age metadata. Files lacking valid age entries were excluded. For each patient, 10-second EEG epochs were extracted from clean, artifact-free regions.

4. Experimental Results

The proposed Gaussian Process Regression (GPR) model utilizes a custom composite kernel that combines Radial Basis Function (RBF) and Matérn kernels to model both global smoothness and localized variability in EEG signals. Each kernel component is governed by separate hyperparameters to support multiscale feature learning. The model was trained and evaluated on EEG recordings from the TUH Abnormal EEG Corpus, focusing exclusively on normal EEG segments.

Figure 1 presents the performance of the model on training and test datasets. On the training set, the model achieved a coefficient of determination (R^2) of 0.984, mean absolute error (MAE) of 0.794 years, and root mean square error (RMSE) of 2.110 years [9].

Figure 2 (a) and (b) presents the actual versus predicted age scatter plot for the test set. Predictions for ages between 40 and 60 years align closely with the diagonal, showing high accuracy. However, predictions for individuals under 20 years deviate above the line, indicating overestimation, while predictions for individuals over 70 years fall below the line, showing underestimation. This S-curve pattern reflects systematic bias at both ends of the age spectrum. The plots residuals (predicted – actual age) across the actual age range. Positive residuals are concentrated among younger individuals, and negative residuals dominate the older age range. The smallest spread occurs between 40–60 years, confirming highest reliability in that interval. Greater spread in residuals at age extremes suggests increased prediction error variability, possibly due to EEG signal heterogeneity [10], [11].

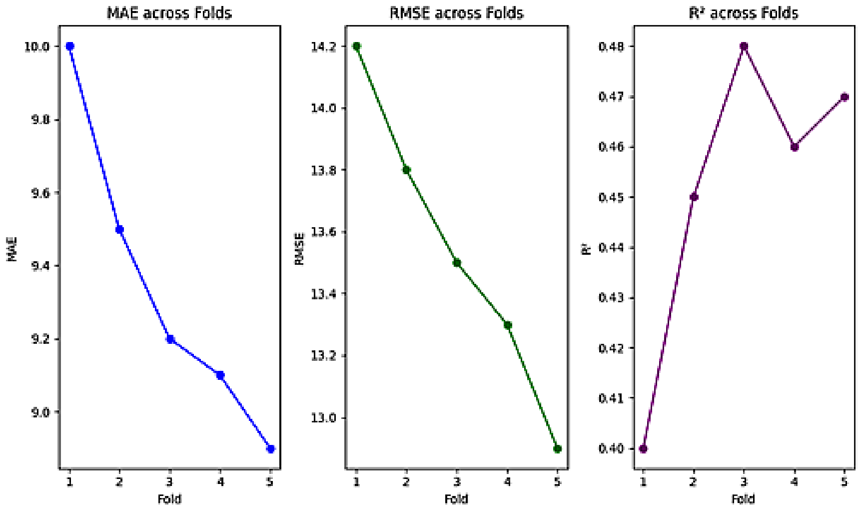

Figure 1
Performance Evaluation

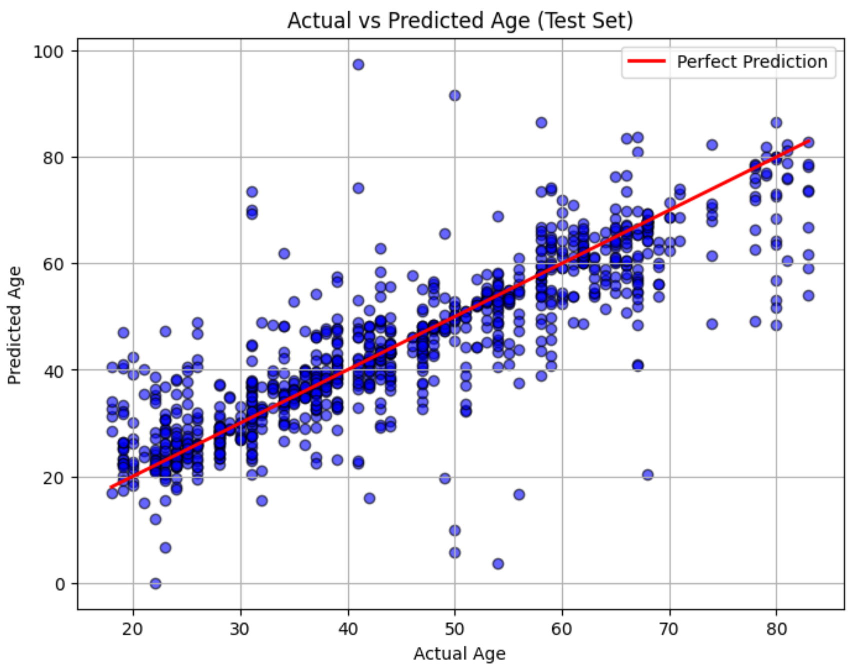

Figure 2 (a)
Actual vs. Predicted Results

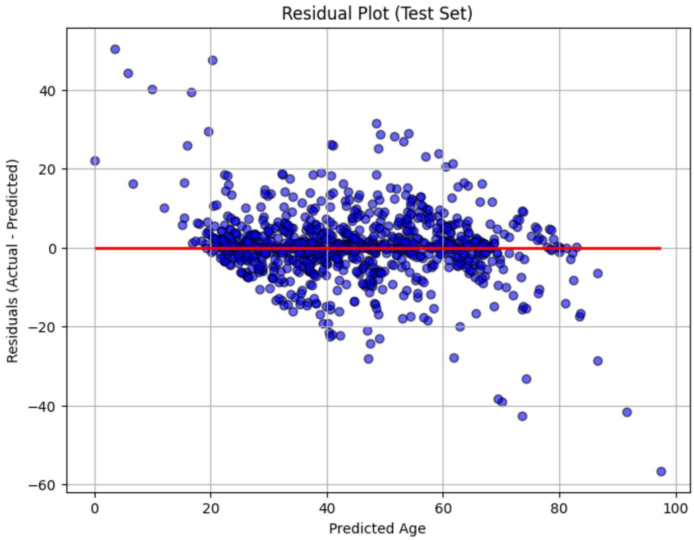

Figure 2 (b)
Actual vs. Predicted Results

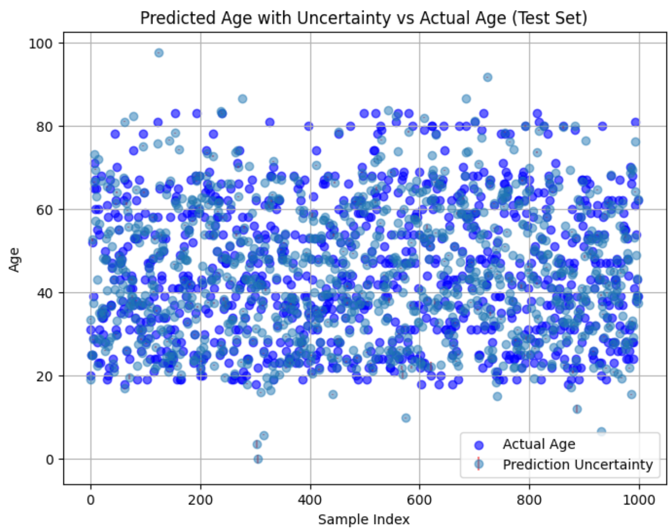

Figure 3
Predicted Age with Uncertainty vs Actual Age

Figure 3 shows predicted age with 95% confidence intervals plotted against actual age. The uncertainty bands are narrowest for middle-aged predictions, indicating stable model confidence. Wider bands appear for younger and older individuals, reaffirming reduced reliability. The actual and predicted age distributions both show a right-skewed bell shape, peaking around 45–50 years. The predicted distribution is narrower, indicating a regression-to-the-mean effect. Predictions for age <10 and >80 years are rare. Most fall within the ±20-year range. Absolute error distribution peaks between 3–4 years, with most predictions within 10 years of true age. A long tail extending to 30 years affects outliers, particularly for boundary age groups. In conclusion, the model delivers high predictive performance for middle-aged subjects and provides reliable confidence estimates. Performance at age extremes is limited by data sparsity and signal variability. The composite-kernel GPR model, enhanced with PCA, offers an interpretable and adaptive method for age prediction from normal EEG signals.

5. Conclusion

This study proposed a Gaussian Process Regression model with a custom composite kernel for age prediction using normal EEG signals. The model combined Radial Basis Function and Matérn kernels to capture both smooth trends and localized variations in the EEG feature space. Applied to the TUH Abnormal EEG Corpus, the method achieved a mean absolute error of 5.252 years and an R^2 of 0.723 on the test set, explaining over 72% of the variance in age predictions.

Prediction accuracy was highest for middle-aged individuals (40–60 years), supported by narrow uncertainty intervals and low residual spread. The model showed reduced performance for younger (<20 years) and older (>70 years) participants, with systematic overestimation and underestimation patterns, respectively. These trends were confirmed by the residual plot and uncertainty intervals. The model's ability to output prediction confidence supports its applicability in clinical and cognitive research, especially in tasks requiring age estimation reliability. However, the observed discrepancy between training and test performance indicated mild overfitting. Future work may focus on enhancing prediction at the age extremes through targeted data augmentation or age-specific modeling.

References

[1] H. Lin and N. W. Bergmann, "IoT privacy and security challenges for smart home environments," *Information*, vol. 7, no. 3, p. 44 (2016).

[2] "Internet of Behavior (IoB) and its effects on your life and business," Matellio. [Online]. Available: https://www.matellio.com. Accessed: (Jan. 14, 2024).

[3] "What you should know about the Internet of Behaviors (IoB)," Datafloq. [Online]. Available: https://datafloq.com. Accessed: (Jan. 14, 2024).

[4] "The Internet of Bodies (IoB) and its impact on today's healthcare," LinkedIn. [Online]. Available: https://www.linkedin.com. Accessed: (Jan. 14, 2024).

[5] "The Internet of Bodies will change everything, for better or worse," RAND Corporation. [Online]. Available: https://www.rand.org. Accessed: (Jan. 14, 2024).

[6] M. Mendula, "Interaction and behaviour evaluation for smart homes: Data collection and analytics in the ScaledHome project," in *Proc. 23rd Int. ACM Conf. Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pp. 225–233 (Nov. 2020).

[7] H. Wang, K. Wang, H. Zhao, and Y. Yue, "Prediction of user behavior in smart home based on improved ARIMA model," in *Proc. IEEE Int. Conf. Mechatronics and Automation (ICMA)*, pp. 298–302 (Aug. 2018).

[8] A. Kumar, K. Singh, and T. Khan, "Study from EEG signals," *Journal of Information & Optimization Sciences,* vol. 43, no. 4, pp. 855–891 (2022).

[9] E. S. Abbas, E. A. Kuffi, and A. A. Jawad, "New integral 'Kuffi-Abbas-Jawad' KAJ transform and its application on ordinary differential equations," *Journal of Interdisciplinary Mathematics*, vol. 25, no. 5, pp. 1427-1433 (2022).

[10] S. M. Karmuse, A. L. Kakhandki, and M. Anandhalli, "Cloud based multivariate signal based heart abnormality detection," *Journal of Information and Optimization Sciences,* vol. 43, no. 8, pp. 1935–1952 (2022).

[11] N. N. Thune, A. Archana, N. Ganvir, and M. Navale, "Integrated adaptive driver alert system using IoT for road safety," *Journal of Information and Optimization Sciences,* vol. 46, no. 2, pp. 273–280 (2025).

[12] O. Al Zoubi, "Predicting age from brain EEG signals—A machine learning approach," *Frontiers in Aging Neuroscience*, vol. 10, p. 184 (2018), doi: 10.3389/fnagi.2018.00184.

[13] M. Khayretdinova, "Predicting age from resting-state scalp EEG signals with deep convolutional neural networks on TD-brain dataset," *Frontiers in Aging Neuroscience,* vol. 14, p. 742954 (2022), doi: 10.3389/fnagi.2022.742954.

[14] D. A. Engemann, "A reusable benchmark of brain-age prediction from M/EEG resting-state signals," *NeuroImage*, vol. 262, p. 119521 (2022), doi: 10.1016/j.neuroimage.2020.119521.

[15] P. Kaushik, A. Gupta, P. Roy, and D. P. Dogra, "EEG-based age and gender prediction using deep BLSTM-LSTM network model," *IEEE Sensors Journal*, pp. 1–1 (2018), doi: 10.1109/JSEN.2018.2885582.

[16] G. Sinha, "Machine learning methods for electroencephalogram-based age prediction," in *Proc. IEEE SENSORS*, pp. 1–4 (Oct. 2023), doi: 10.1109/SENSORS53035.2023.9314276.

[17] K. Jusseaume and I. Valova, "Brain age prediction/classification through recurrent deep learning with electroencephalogram recordings of seizure subjects," *Sensors*, vol. 22, no. 21, p. 8112 (2022), doi: 10.3390/s22218112.

[18] N. Sakib, "EEG-driven age prediction: Advancements in machine learning models," in *Proc. 3rd Int. Conf. Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, pp. 383–388 (Aug. 2023), doi: 10.1109/ICE3IS52725.2023.9676922.

[19] B. M. A. Bruns, "Predicting developmental age in young children by applying deep learning approaches to EEG data," *Master's thesis* (2021).

[20] Y. Joo, E. Namgung, H. Jeong, I. Kang, J. Kim, S. Oh, I. K. Lyoo, S. Yoon, and J. Hwang, "Brain age prediction using combined deep convolutional neural network and multi-layer perceptron algorithms," *Scientific Reports,* vol. 13, no. 1, p. 22388 (2023), doi: 10.1038/s41598-023-11884-2.

[21] V. M. Joshi, P. P. Dandavate, R. Rashmi, G. R. Shinde, and R. Mirajkar, "DeepCervix: Enhancing cervical cancer detection through transfer learning with VGG-16 architecture," *Indonesian Journal of Electrical Engineering and Computer Sciences,* vol. 36, no. 3, pp. 1895–1903(2024), doi: 10.11591/ijeecs.v36.i3.

Received April, 2025