# Multimodal Analysis of Emotion-Related Behavioral Patterns During Affective Labeling Using Hand Tracking and Eye Tracking

Gabriele Trovato[1][0009-0008-6751-5144], Vito Cammarata[1][0009-0003-2459-6486], Adriano Schimmenti[2][0000-0001-5181-2648], and Nicole Dalia Cilia[1][0000-0002-9631-5802]

[1] Department of Engineering and Architecture, University of Enna "Kore", Italy
{gabriele.trovato, vito.cammarata, nicoledalia.cilia}@unikore.it
[2] Department of Human and Social Sciences, University of Enna "Kore", Italy
adriano.schimmenti@unikore.it

**Abstract.** Understanding how human behavior simultaneously reflects transient emotional states and stable individual differences in emotion recognition remains a central challenge in affective computing. This study investigates emotion-related behavioral patterns during an affect-labeling task by integrating handwriting (HT) and eye-tracking (ET) signals. Participants viewed emotionally evocative images and labeled the perceived emotion through handwriting, while gaze was recorded via Tobii Pro Glasses 3 and graphomotor data via a Wacom One tablet. Four classifiers with hyperparameters optimized via evolutionary algorithms were trained for emotion classification, with SHAP-based explainability highlighting feature contributions. Results show that HT features dominated emotion classification, reflecting embodied affective expression through motor behavior, whereas ET metrics served as predictors of Theory of Mind scores in regression analyses, revealing stable physiological markers of mental-state attribution independent of task-specific labeling. Together, these findings suggest that multimodal signals differentially encode transient emotional responses and stable individual traits.

**Keywords:** Autism Spectrum Disorder · Skeletal Motion Analysis · Sequence Modeling · Feature Fusion · Deep Learning

## 1 Introduction

Today, Affective Computing stands as one of the most challenging frontiers in computational emotion prediction, seeking to endow intelligent systems with the ability to recognize, interpret, and simulate human emotions. Understanding emotional states is not only crucial for enhancing user experience but also plays a fundamental role in fields such as clinical psychology, healthcare, and education. However, traditional literature has historically focused on explicit expressive modalities, such as facial expressions or vocal tone. While widely utilized, these channels suffer from the issue of "Social Masking": individuals can, consciously or unconsciously, manipulate their mimicry or voice to conceal their true inner state, rendering systems based on such data vulnerable and less robust [1].

To overcome these limitations, research is shifting toward multimodal approaches that integrate physiological and behavioral signals less prone to voluntary control. In this context, while the use of Eye-Tracking (ET) to measure visual attention and arousal is well-documented, a significant gap remains in the combined exploration of ET and HT analysis (HT). Indeed, HT is not merely a motor act but a complex psychophysical process that reflects the writer's inner state through variations in pressure, speed, and micro-tremors.

The present work aims to bridge this gap through an innovative approach that combines HT analysis and ET for emotion classification, while simultaneously introducing a validation based on implicit emotion recognition. To evaluate the latter, the Reading the Mind in the Eyes Test (RMET) was employed [2] [3]. The RMET is a standardized measure of Theory of Mind (ToM) (i.e., the capacity to attribute mental states). Participants view 36 images showing only the eye region of faces and select the term that best describes the expressed mental state from four alternatives that reflect complex affective states. The RMET indexes implicit emotion recognition in ToM, as performance relies on the automatic decoding of subtle affective cues from the eyes, entailing an intuitive perceptual-affective processing that precedes explicit emotional reasoning [4]. The underlying hypothesis is that ocular patterns during writing are not only indicators of a momentary emotional state but are also modulated by the subject's individual capacity for ToM.

The methodology of this study is structured into three sequential phases designed to address the challenges of optimization and interpretability in advanced Machine Learning (ML) frameworks. First, the study performs a multimodal classification of emotional states. To maximize predictive performance, Evolutionary Algorithms are employed for the automated search of optimal model hyperparameters, ensuring that each classifier, ranging from Neural Networks to Ensemble methods, is finely tuned to the specificities of the dataset. Second, the framework incorporates model explainability through the SHAP (SHapley Additive exPlanations) library. This step is crucial for identifying the most significant features within the motor and ocular channels, offering a transparent view of the algorithm's decision-making process and quantifying the contribution of each behavioral marker to the final prediction. Finally, the research extends beyond categorical classification to investigate the relationship between physiological signals and psychological traits. Specifically, linear and non-linear regressors are utilized to perform a correlation analysis with the participants' Reading the Mind in the Eyes Test (RMET) scores. While HT-related features are expected to dominate the transient emotion classification task, this regression phase aims to reveal subtle pre-motor and ocular signals that correlate significantly with the subjects' ToM levels.

## 2 Related works

Research in emotion recognition is increasingly prioritizing physiological and behavioral signals over explicit expressive modalities to circumvent the issue of social masking, the tendency of individuals to consciously or unconsciously manipulate their visible emotional expression [1]. In this context, neurophysiological approaches based on EEG, often combined with eye-tracking to measure valence and arousal, represent a well-established paradigm that has demonstrated the advantage of multimodal fusion over single-modality systems [5] [6] [7]. However, these methods rely on invasive or cumbersome acquisition setups, limiting their applicability in naturalistic settings. A less explored but promising direction concerns the use of handwriting as a non-invasive behavioral marker of affective states. HT is not merely a motor act but a complex psychophysical process that reflects the writer's inner state through variations in pressure, speed, and microtremors. Lopez et al. [8] proposed a multi-level approach to detect emotional alterations through pressure and stroke variations, providing an objective tool for psychological support beyond traditional stress-focused studies. Similarly, Azmi et al. [9] leveraged routine writing activities from 129 participants, using Random Forest models to successfully predict emotional states from stroke-level features.

Despite these results, a gap remains in understanding the synchronized relationship between motor execution and ocular planning during affective tasks. While HT captures stable motor patterns, ET offers a unique window into the cognitive load and attentional shifts underlying emotional expression. This study addresses this gap by proposing a multimodal framework that combines HT kinematics with ET dynamics, using SHAP analysis to explain the underlying predictive mechanisms and regression models to investigate how these features correlate with Theory of Mind levels as measured by the RMET [2] [3].

## 3 Methodology

The methodology was structured to address two distinct research questions: (RQ1). Which sensory modality contributes most significantly to the recognition of emotional states? (RQ2). To what extent are involuntary behavioral patterns correlated with Theory of Mind (ToM) levels?. To address these research questions, we first detail the dataset construction and preprocessing pipeline in 4. Subsequently, the results and discussion are organized as follows: Section 5 investigates the contribution of different sensory modalities to emotion recognition (RQ1), while Section 6 explores the correlation between behavioral patterns and ToM (RQ2).

The first experiment evaluates emotion recognition through a classification-based ablation analysis. Three configurations are compared to quantify the contribution of each modality: (i) ET only, to isolate autonomic responses and visual attention; (ii) HT-only, to assess how emotional states alter fine motor dynamics; and (iii) Multimodal fusion, to verify if a holistic approach outperforms single-channel performance and mitigates modality-specific ambiguities.

The second experiment investigates the link between involuntary behavioral patterns and ToM. Modeled as a regression analysis, it explores the association between ET features and RMET scores. Based on the theoretical premise that gaze-based spatial exploration anticipates motor acts, this analysis focuses exclusively on the ET channel to isolate the specific ocular signals that best explain the variance in participants' ToM levels.

## 4 Dataset

For the experimental validation, a specifically acquired multimodal dataset was utilized, involving a total of 65 participants aged between 20 and 27 years, ensuring gender balance. To maintain subject independence, we performed a subject wise split of the data. Given that each participant evaluated eight different images, all trials from a single subject were assigned exclusively to either the training or the testing set, thereby preventing any overlap of individual specific features between the two sets. The subjects were exposed to visual stimuli designed to elicit four specific emotions: Joy, Sadness, Disgust, and Relaxation. Each emotion was mapped onto two slides, resulting in a total of eight carefully selected slides. The acquisition protocol simultaneously integrated two data streams: the first comprises kinematic, spatial, and pressure signals acquired via a Wacom One 13 graphic tablet; the second includes ET metrics recorded using Tobii Pro Glasses 3. Additionally, the dataset incorporates psychometric data, obtained through RMET profiling, and demographic information collected via a questionnaire.

Following data acquisition, three distinct data groups were identified: HT, ET, and psychometric/demographic data. The HT data required significant preprocessing, as the acquisition software provided raw time series. These were processed by extracting descriptive statistics (min, max, mean, and std). A key methodological choice concerns the "pen phase" considered for these statistics: kinematic features were calculated exclusively during the contact phase (Stroke), whereas pressure and pen-orientation features were calculated across the entire duration of the time series (Hover and Stroke combined). In contrast, the ET data were preprocessed by the proprietary Tobii Pro Lab software and provided as pre-engineered features. Prior to model training, the column containing slide IDs was removed to prevent the models from predicting emotions based on the sequential order of the stimuli.

Upon completion of the preprocessing stage, the resulting dataset consists of eight rows per subject, for a total of 520 observations and 75 columns. Table 1 provides a detailed description of the features included in the dataset, categorized by type.

Table 1: Description of features.

| Modality | Features Summary |
|---|---|
| **Demographics** | ID, Hand, Gender, Age, Educ, Sleep, Coffee, Mood, RMET. |
| **ET** | Fixation (Dur, Count, Pupil), Saccades (Count, Time to 1st), Kinematics (Vel, Ampl, 1st Saccade metrics). |
| **HT** | Stats (Mean, SD, Min, Max) for: Pressure, X-Y Coords, Rotation, Azimuth, Tilt. |

## 5 First Experiment: Modality-based Ablation

The primary objective of this first experiment was to isolate the predictive contribution of each sensory modality. Independent models were trained and optimized using feature sets related to HT dynamics and ET movements, subsequently progressing toward a multimodal fusion approach.

To identify the optimal hyperparameters for each model, genetic algorithms were implemented via the PyGAD library, enabling an autonomous and comprehensive exploration of the configuration space. To ensure experimental reproducibility, Table 2 details the search spaces defined for the evolutionary optimization. The core parameters for the genetic algorithms, common to all models, included random wheel selection, single-point crossover, a 20% mutation rate, and a fitness function based on classification accuracy.

Table 2: Hyperparameter search spaces for the evolutionary optimization process.

| Model | Evolutionary Search Space |
|---|---|
| **MLP** | Layers $\in \{1, 2, 3\}$; Neurons $\in \{8, 16, 32, 64, 128\}$; Activation: {relu, sigmoid, tanh, elu, swish, selu, softplus, gelu}; LR $\in [10^{-4}, 10^{-2}]$; Dropout $\in [0.1, 0.5]$ |
| **LR** | $C \in [0.01, 10.0]$; Max Iter $\in [50, 500]$ |
| **RF** | N. Estimators $\in [10, 200]$; Max Depth $\in [2, 20]$ |
| **NB** | Var. Smoothing $\in [10^{-9}, 1.0]$ |
| **Voting Clf.** | Weights (RF, LR, NB) $\in [0.1, 5.0]$ |
| **Bagging (KNN)** | N. Est. $\in [10, 100]$; $k \in [1, 20]$ |
| **Bagging (SVM)** | N. Est. $\in [5, 50]$; $C \in [0.1, 5.0]$ |
| **Bagging (NB)** | N. Est. $\in [10, 100]$; Var. Smoothing $\in [10^{-9}, 1.0]$ |
| **Bagging (LR)** | N. Est. $\in [10, 100]$; $C \in [0.01, 10.0]$ |
| **Boosting (LR)** | N. Est. $\in [10, 100]$; LR $\in [0.01, 2.0]$; $C \in [0.01, 10.0]$ |

To establish a robust baseline, the experimentation initially included probabilistic and linear models. Specifically, Logistic Regression (LR) [10] was employed as a baseline to verify the linearity of the feature space, while the Naive Bayes classifier (NB) [11] was selected for its computational efficiency and its assumption of feature independence, which is often effective for statistical data.

Subsequently, we used ensemble learning techniques, with particular emphasis on bootstrap aggregating and bagging strategies. This meta-algorithm was systematically applied to various base estimators, including Support Vector Machines (SVM) [12], K-Nearest Neighbors (KNN), [13], Naive Bayes (NB), and Logistic Regression (LG), to reduce model variance and mitigate the impact of noise within the dataset. Within this framework, Random Forest (RF) [14] was also utilized as a natural evolution of bagging applied to decision trees, proving particularly effective in capturing the non-linear relationships inherent in HT dynamics. Regarding Boosting strategies, the AdaBoost algorithm was implemented using LG as a weak learner. Furthermore, a Voting Classifier was configured to combine predictions from RF, LR, and NB, leveraging the heterogeneity of their error patterns to achieve a more robust collective decision. Finally, we trained a Multi-Layer Perceptron (MLP). For this neural architecture, the genetic algorithm was tasked with evolving not only the training hyperparameters, such as learning rate and dropout, but also the network topology itself, dynamically defining the number of hidden layers and neurons, thus allowing for a data-driven architectural design. All models were evaluated using Stratified K-Fold Cross-Validation ($K = 5$) to ensure the generalizability of the results across different subjects.

The classification performance across the three sensory configurations (HT only, ET only, and Multimodal Fusion) is summarized in Table 3. The results demonstrate that the integration of both modalities consistently enhances stability and, in several instances, predictive accuracy, validating the holistic approach proposed in this study.

Specifically, the Voting Classifier achieved the highest overall performance in the Multimodal configuration (64.74%), whereas the MLP proved to be the most effective model for the unimodal HT task (63.00%). As anticipated, the ET configuration presented the lowest baseline accuracy, with the Boosting LG model reaching a maximum of 46.15%. These results provide a quantitative foundation for the subsequent ablation analysis and explainability investigations.

For model explainability, detailed in the following sections 5.1, 5.2, 5.3 the SHAP (SHapley Additive exPlanations) framework was chosen to provide a global comparison of features. The 20 most influential features for the highest-performing model were analyzed.

### 5.1 Hand Tracking

The original dataset was filtered by removing all ET metrics, retaining only the HT features and psychometric/demographic data. These latter variables were treated as physiological covariates, as factors such as sleep duration and number of daily coffees may significantly influence motor performance. For all three modality configurations, the MLP was selected for the explainability analysis, as it achieved the highest unimodal performance and its architecture, evolved

Table 3: Accuracy results on the three modalities

| Model | HT + ET | HT | ET |
|---|---|---|---|
| Voting Classifier | **0.6474** | 0.5769 | 0.4230 |
| MLP | 0.6000 | **0.6300** | 0.4500 |
| Random Forest | 0.5769 | 0.5321 | 0.4423 |
| Bagging Log. Reg. | 0.5705 | 0.5833 | 0.3782 |
| Logistic Regression | 0.5705 | 0.5769 | 0.3782 |
| Boosting Log. Reg. | 0.5705 | 0.4872 | **0.4615** |
| Bagging SVM | 0.5449 | 0.5000 | 0.4038 |
| Bagging Naive Bayes | 0.4423 | 0.4744 | 0.2884 |
| Naive Bayes | 0.4103 | 0.4615 | 0.4038 |
| Bagging KNN | 0.3782 | 0.3718 | 0.2948 |

via genetic algorithms, is uniquely capable of capturing non-linear relationships between physiological signals and emotional states.

The SHAP analysis (Fig. 1) highlights that the decision-making process is primarily driven by kinematic dynamics and force modulation rather than static geometric properties of the stroke. The standard deviation of coordinates along the Y-axis emerged as the strongest predictor, indicating that emotional states alter the amplitude and vertical fluctuation of the stroke. The variance and mean of pressure further confirm that muscular tension acts as a reliable marker of arousal, allowing the model to discriminate emotional states beyond the potential bias introduced by word length.

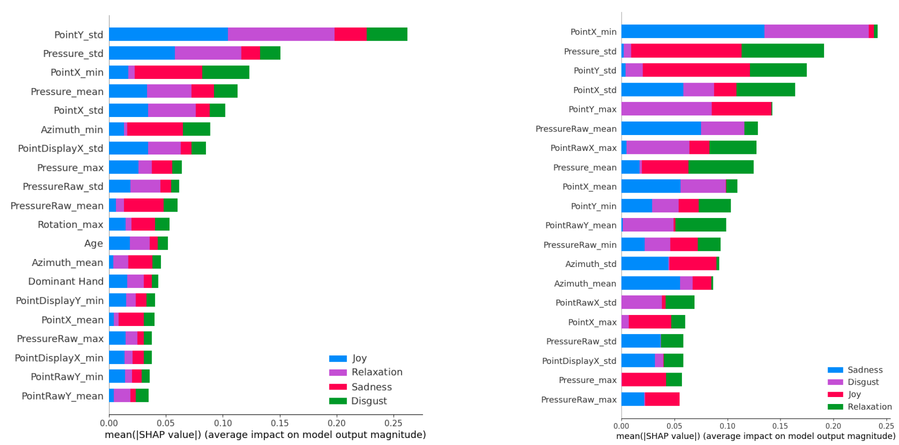


Fig. 1: Feature importance for best models (MLP e Bagging LR).

### 5.2 Eye Tracking

In this configuration the feature vector was constructed by isolating ET metrics while retaining psychometric and demographic metadata, given the well-documented physiological impact of factors such as sleep hours and caffeine intake on ocular reactivity. This modality presented the greatest classification challenges, with the Boosting (LR) model achieving the highest accuracy; however, the MLP was retained for XAI consistency across configurations.

The SHAP analysis (Fig. 2) reveals an unexpected scenario compared to classical Affective Computing literature. While previous studies frequently identify pupil diameter as the primary indicator of arousal, this metric shows only marginal impact here. Instead, the decision-making process is dominated by fixation metrics, suggesting that the model discriminates emotions based on gaze dwell times: emotions requiring higher cognitive processing generate longer or more frequent fixations compared to states of disinterest or avoidance. Psychometric and demographic variables (e.g., number of coffees, age) appear in the middle of the ranking, confirming their role in normalizing individual differences in ocular physiology.

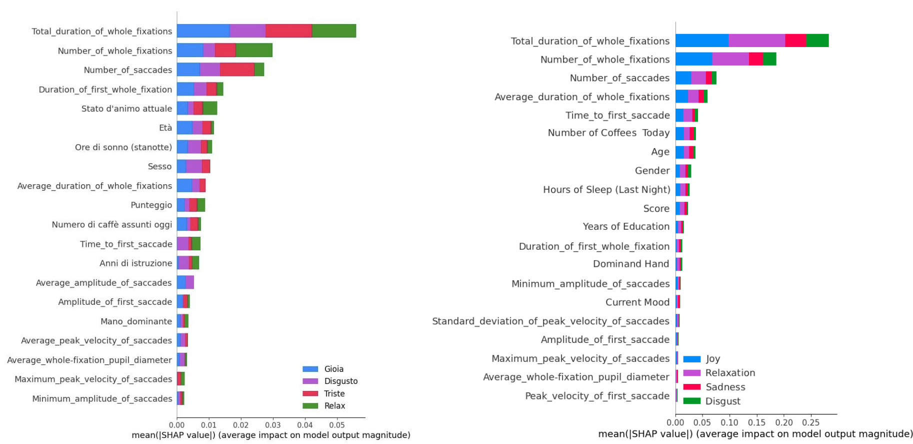


Fig. 2: Feature importance of best models (MLP and Boosting LR).

### 5.3 Multimodal

The third configuration implements a holistic approach via Feature-Level Fusion, concatenating the normalized HT and ET vectors into a single multidimensional input space. Although the Voting Classifier achieved the highest accuracy (Section 5), the MLP was retained for XAI analysis, as its Early Fusion architecture enables direct inspection of how the two modalities are weighted and integrated within the hidden layers.

The SHAP analysis (Fig. 3) reveals a precise informational hierarchy: HT dynamics remain the dominant predictors, while ET contributes a crucial supporting role. Specifically, *Total_ duration_ of_ whole_ fixations* emerges as the fifth most influential feature overall, suggesting that the model integrates motor activation with visual cognitive load to resolve ambiguous cases where the motor signal alone is insufficient.

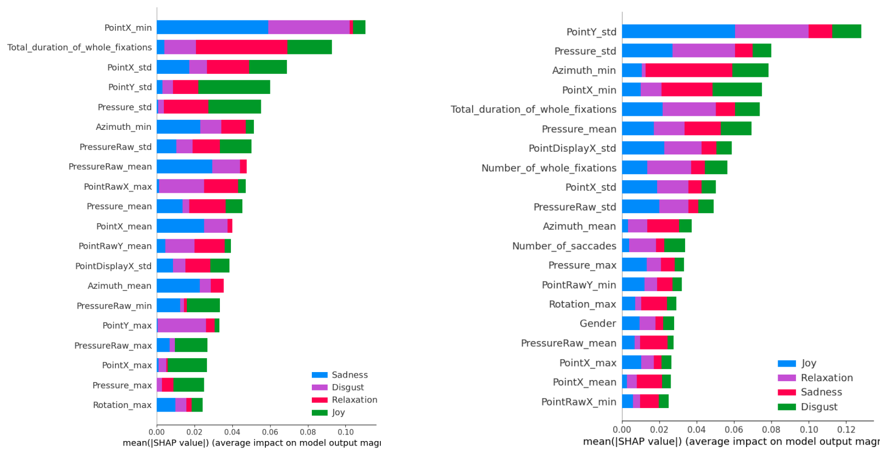


Fig. 3: Feature Importance for the best models (Voting Classifier and MLP).

## 6 Second Experiment: Statistical Analysis of Theory of Mind

While the first experiment focused on the classification of emotional states, the second investigation transitions to a regression-based framework to explore the relationship between behavioral patterns and ToM, as measured by RMET scores ($M = 24.03, SD = 3.31$). The goal is to determine if involuntary physiological responses can act as proxies for stable psychological traits. This analysis is structured upon three distinct methodological pillars:

- Task Independence and Psychological Consistency: To avoid psychometric confounding, the analysis was first conducted on the dataset of each individual stimuli. Mixing data from different emotional tasks (e.g., Joy vs. Disgust) to directly predict a stable trait like ToM would be theoretically unsound, as momentary emotional responses could act as noise. Therefore, the analysis was initially performed on a per-task basis.
- Aggregation for Trait Stability (Mean Task): To capture the subject's stable "behavioral signature" and mitigate the variance of individual trials, a new synthetic task was generated. This was achieved by calculating the mean

of the ET features across all eight tasks for each subject. This aggregated approach focuses on the subject's consistent ocular patterns, which are more likely to reflect permanent traits such as ToM.

- Isolation of Ocular Metrics: Only ET features were utilized for this experiment. This choice was made to decouple the analysis from the specific verbal content being written. While HT is inextricably linked to the motor production of a specific word, ocular movements, specifically pre-motor planning and fixations, offer a more direct window into the cognitive and affective processing of ToM.

Table 4: Model performance comparison ($R^2$) across tasks.

| Task | Lasso | OLS (Stats) | Random Forest | Ridge |
|---|---|---|---|---|
| 1 | 0.1684 | 0.4647 | 0.8311 | 0.3025 |
| 2 | 0.1517 | 0.4007 | 0.8441 | 0.2863 |
| 3 | 0.1921 | 0.5150 | 0.8700 | 0.3525 |
| 4 | 0.1544 | 0.4308 | 0.8354 | 0.2695 |
| 5 | 0.1571 | 0.5062 | 0.8428 | 0.3489 |
| 6 | 0.1730 | 0.4129 | 0.8439 | 0.3229 |
| 7 | 0.1608 | 0.3825 | 0.8343 | 0.3029 |
| 8 | 0.3965 | 0.6513 | 0.8782 | 0.5138 |
| ALL | 0.1742 | 0.5880 | 0.8519 | 0.3332 |

### 6.1 Statistical Framework

Given the limited sample size ($N = 65$) resulting from subject-level aggregation, which precludes the use of standard machine learning for predictive generalization, we transitioned to a statistical regression framework to investigate the trait-level relationship with ToM. Rather than aiming for out-of-sample prediction, we evaluated multiple regression architectures to capture the complex dependencies between ocular behavior and psychometric scores. To ensure the reliability of the associations found, we assessed our findings through $p$-value analysis and explained variance ($R^2$). This hybrid approach combines the flexibility of machine learning with the rigorous inferential standards of classical statistics, ensuring that the identified correlates of ToM are both robust and statistically significant. The analysis in the Table 4 reveals that ET metrics serve as significant indicators of individual ToM scores. Specifically, the aggregated model (Task 9) demonstrated high explanatory power ($R^2 = 0.588$), significantly outperforming individual task models. This suggests that averaging ocular behavior across multiple stimuli effectively filters out transient environmental noise, revealing the underlying stable psychological traits.

Table 5: Statistically significant features ($p < 0.05$) from OLS analysis.

| **Task** | **Feature** | **P-Value** | **Coeff. ($\beta$)** |
|---|---|---|---|
| 1 | Total amplitude of saccades | 0.0160 | -5.7225 |
| | Gender | 0.0190 | 0.0351 |
| | Amplitude of first saccade | 0.0323 | 0.7606 |
| | Peak velocity of first saccade | 0.0389 | -1.5283 |
| | Age | 0.0485 | 0.0356 |
| 2 | Gender | 0.0309 | 0.0301 |
| 3 | SD of peak velocity | 0.0065 | -4.1519 |
| | Max peak velocity | 0.0113 | 3.9882 |
| | Gender | 0.0497 | 0.0255 |
| 4 | Gender | 0.0055 | 0.0451 |
| | Direction of first saccade | 0.0404 | 1.3093 |
| 5 | Number of whole fixations | 0.0352 | -0.0642 |
| | Age | 0.0443 | 0.0360 |
| | Avg duration whole fixations | 0.0475 | -0.0684 |
| 6 | Age | 0.0449 | 0.0411 |
| 7 | Age | 0.0333 | 0.0381 |
| 8 | Time to first saccade | 0.0167 | -0.0317 |
| | Avg velocity first saccade | 0.0336 | 2.7293 |
| ALL | Gender | 0.0043 | 0.0372 |
| Mean | Amplitude of first saccade | 0.0054 | 5.0985 |
| | Peak velocity first saccade | 0.0118 | -9.3404 |
| | Avg pupil diameter | 0.0167 | -8.5593 |
| | Min peak velocity | 0.0337 | 7.0738 |

### 6.2 Global Predictors and Physiological Markers

While non-linear models provided higher overall explanatory power (see 4), we report the OLS coefficients to offer a direct and interpretable assessment of the relationship between individual features and ToM scores. The features that reached statistical significance are summarized in Table 5. Beyond demographic factors such as female ender ($p = 0.004$), which remained the most consistent predictor across configurations, the most critical physiological markers of ToM were linked to the initial orienting response. Specifically, the Amplitude of the first saccade ($p = 0.005$) and its Peak Velocity ($p = 0.011$) emerged as primary indicators. Females' advantage on emotion-recognition tasks, likely related to evolutionary adaptiveness is well documented in meta-analytic studies [15]. Also, higher performance in emotion recognition is associated with more selective social attention (greater fixation on eyes and fewer large exploratory saccades): this can plausibly be interpreted as an attentional trade-off, that is, prioritising accuracy and focused processing of socially relevant cues over rapid, exploratory eye movements [16]. Accordingly, the inclusion of the average pupil diameter as a significant predictor in the aggregated model ($p = 0.016$) reinforces the hypothesis that autonomic nervous system activity captures stable features of emotional response. Interestingly, this metric was marginal in the transient emotion classification task (Experiment 1), but becomes a "Golden Feature" when assessing long-term ToM sensitivity.

### 6.3 Task-Specific Dynamics

From the Table 5, we can note that the variance in $R^2$ across individual tasks (ranging from 0.38 to 0.65) highlights that implicit emotion recognition is not a static signal but a context-dependent reaction:
Positive Stimuli (Task 8): This task, related to the emotion of joy, achieved the highest individual performance ($R^2 = 0.651$). The significance of Time to first saccade here suggests that positive emotional stimuli "trigger" emotional salience more rapidly, leading to a faster attentional shift.
Cognitive Processing of Negative stimuli (Task 5): Unique to this task, related to sadness, was the significance of fixation metrics (Number and Average duration of whole fixations, $p < 0.05$). In this context, higher ToM scores correlate with reduced fixation, suggesting emotional avoidance and distancing from negative emotion.
Saccadic Variability (Task 3): The significance of the Standard deviation of peak velocity ($p = 0.006$) in this task, related to disgust, reflects a less rhythmic, more "engaged" and narrow visual exploration pattern in individuals high in ToM when confronted with the aversive stimulus.

Taken together, the RMET findings support a view of emotion recognition as dynamic, and inherently context- and content-based: individuals with higher ToM ability appear to engage in balanced implicit processing of emotional stimuli, characterized by approach tendencies toward positive affect and physiological avoidance of negative affect. This pattern points to a principled link between ToM and emotion regulation capacity, a core dimension of mental health.

## 7 Discussion

The present study investigated multimodal behavioral signals during an emotion elicitation and labeling task, combining HT dynamics and ET data. While the classification performance across four emotional categories was satisfactory, the dominance of HT-related features suggests that affective information, in this specific protocol, is primarily externalized through graphomotor behavior. Features such as writing speed, pressure, and stroke dynamics proved highly discriminative; however, they are inherently tied to the motor act of labeling. This indicates that HT captures embodied affective expression, the way an emotional state modulates a physical action, rather than a "pure" decoding of latent affect. In contrast, ET features represent a more direct measure of perceptual and cognitive processing, as they precede both motor execution and linguistic production. Even the modest classification results obtained using ET alone indicate the presence of early emotion-related perceptual mechanisms, supporting the hypothesis that affective processing begins well before explicit verbalization or motor labeling take place. Since ocular movements offer a unique window into these earliest stages of response, they are less susceptible to the task-specific biases that characterize handwriting. To decouple these transient emotional states from stable psychological characteristics, we transitioned to a statistical modeling of ToM. This shift is fundamental: while HT is inextricably linked to the motor production of a specific word (potentially introducing lexical confounds), ToM represents a stable, latent construct independent of the task's verbal content. By focusing the second experiment on ET metrics, we targeted the anticipatory cognitive processing that precedes motor execution. The results from the regression analysis ($R^2 = 0.587$ for the aggregated model) provide a clear distinction between state-dependent and trait-dependent markers. While ocular features contributed less to instantaneous emotion classification, they emerged as "Golden Features" for explaining the variance in ToM scores. The significance of the initial orienting response (e.g., first saccade amplitude and velocity) and autonomic nervous system activity (pupil diameter) suggests that individuals with higher ToM possess a more efficient "attentional trade-off." They prioritize the focused processing of socially relevant cues over broad exploratory movements, a pattern that becomes evident only when filtering out environmental noise through subject-level aggregation. These findings support a view of emotion recognition as a dynamic process: individuals with high ToM ability engage in balanced implicit processing, characterized by approach tendencies toward positive affect and physiological avoidance of negative stimuli. Thus, the multimodal approach does not merely classify emotions but serves as a marker of trait-level emotional regulation capacity, a core dimension of mental health.

## 8 Conclusion and future works

This work explored behavioral markers of affective processing by combining handwriting dynamics and eye-tracking data. The results indicate that while emotional information is predominantly expressed through handwriting, this channel remains closely tied to the motor and lexical constraints of the task. In contrast, the analysis of ToM reveals that ocular behavior captures intrinsic, trait-level characteristics. By demonstrating that involuntary eye movements explain nearly 60% of the variance in ToM scores, this study underscores the value of eye-tracking as a window into the anticipatory cognitive processes that precede explicit action.

Future work will focus on refining experimental protocols to further isolate perceptual mechanisms from motor execution, possibly through non-verbal response paradigms. Furthermore, we intend to explore the generalizability of these physiological markers in clinical settings where ToM deficits represent a primary diagnostic feature. Such efforts will contribute to the development of more robust and interpretable models, capable of distinguishing between transient emotional expressions and stable psychological traits of the subject.

## References


1. D. B. Grahlow M, Rupp CI, The impact of face masks on emotion recognition performance and perception of threat, PLoS One 17 ((2):e0262840) (2022) PMC8836371.
2. S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, I. Plumb, The Reading the Mind in the Eyes test: Revised version, Journal of Child Psychology and Psychiatry (2001).
3. S. Baron-Cohen, T. Jolliffe, C. Mortimore, M. Robertson, Another advanced test of theory of mind: Evidence from very high functioning adults with autism or asperger syndrome, Journal of Child Psychology and Psychiatry 38 (7) (1997) 813–822.
4. B. F. M. Oakley, R. Brewer, G. Bird, C. Catmur, Theory of mind is not theory of emotion: A cautionary note on the reading the mind in the eyes test, Journal of Abnormal Psychology 125 (6) (2016) 818–823.
5. W.-L. Zheng, B.-N. Dong, B.-L. Lu, Multimodal emotion recognition using eeg and eye tracking data, in: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014, pp. 5040–5043. doi:10.1109/EMBC.2014.6944757.
6. Z. Fan, Brain-computer interface for emotion recognition based on electroencephalography, Emotion 85 (2024).
7. O. Bazgir, Z. Mohammadi, S. A. H. Habibi, Emotion recognition with machine learning using eeg signals, in: 25th national and 3rd international iranian conference on biomedical engineering (ICBME), IEEE, 2018, pp. 1–5.
8. J.-M. López-Gil, J. Virgili-Gomá, R. Gil, T. Guilera, I. Batalla, J. Soler-González, R. García, Method for improving eeg based emotion recognition by combining it with synchronized biometric and eye tracking technologies in a non-invasive and low cost way, Frontiers in computational neuroscience 10 (2016) 85.
9. M. Azmi, S. L. Fathima, M. Wasid, Unveiling emotions through handwriting: a data analysis approach, in: 2023 12th International Conference on Advanced Computing (ICoAC), IEEE, 2023, pp. 1–6.
10. D. W. J. Hosmer, S. Lemeshow, R. X. Sturdivant, Applied logistic regression, John Wiley Sons. (2013).
11. L. Breiman, An empirical study of the naive bayes classifier, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence (2001) 41–46.
12. C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.
13. T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13 (1) (1967) 21–27.
14. L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.
15. A. E. Thompson, D. Voyer, Sex differences in the ability to recognise non-verbal displays of emotion: a meta-analysis, Cognition and emotion (28(7)) (2014) 1164–1195.
16. V. Bachurina, M. Arsalidou, Multiple levels of mental attentional demand modulate peak saccade velocity and blink rate, Heliyon (8(1)) (2022) e08826.