

RESEARCH ARTICLE

Saturation in modified correlation metrics: a reanalysis of advanced Matthews coefficient result for EEG stress classification

Mahfuzah Mustafa^{1*}, Zarith Liyana Zahari², Lailatul Niza Muhammad¹, Faradila Naim¹, Rafiuddin Abdubrani³

¹Faculty of Electrical and Electronics Engineering Technology, Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Pahang, Malaysia

²Electronics Technology Section, British Malaysian Institute, Universiti Kuala Lumpur, 53100 Kuala Lumpur, Malaysia

³Project Management, Petra Resources Sdn. Bhd., 47820 Selangor, Malaysia

Abstract - Mental stress is increasingly assessed through electroencephalography (EEG), which offers objective, temporally resolved measurement. The Advanced Matthews Correlation Coefficient (MCCA) has been proposed for this application, replacing the false-positive and false-negative terms of the conventional Matthews correlation coefficient with ratios of EEG arousal and valence indices raised to a fixed power, arousal classification rising from about 10% to above 98% has been reported. This study pursues two objectives: to determine analytically how MCCA behaves as its power exponent increases, and to establish when the previously reported accuracies may be generalised. The stress features examined are power spectral density values over the delta, theta, alpha and beta sub-bands of ten frontal and temporal channels, from which the arousal ratio $(\alpha + \beta) / (\delta + \theta)$ and its left-right difference are formed. Random Forest, support vector machine, and k-nearest neighbours classifiers were evaluated under 10-fold cross-validation. MCCA is shown to converge to one of three fixed values, +1, -1, or 0, according to whether the substituted ratios exceed unity, with the true-positive and true-negative counts eliminated in the limit. The exponent needed to reach this regime scales with confusion-matrix size, so no fixed value is universally appropriate, a selection rule is derived and verified numerically. Because the limiting value is fixed by signal-derived arousal and valence ratios rather than by classifier outcomes, MCCA is better interpreted as an index of the recorded EEG signal than as a measure of classification quality. An enhanced multimodal fusion formulation is specified for future validation.

Article History

Received : 18 August 2026
 Revised : 5 October 2026
 Accepted : 8 October 2026
 Published : 22 October 2026

Keywords

Electroencephalography
Stress classification
Matthews Correlation Coefficient
Multimodal fusion
Power spectral density

1. Introduction

Mental stress has emerged as a critical public health concern globally, with physiological manifestations detectable through multiple biosignals [1], [2]. Prolonged exposure to psychosocial stressors is associated with measurable dysregulation of the hypothalamic pituitary adrenal axis and with sustained autonomic imbalance, and the consequent burden on healthcare systems and workplace productivity is now widely acknowledged [3]. The burden is particularly pronounced in academic populations, where sustained cognitive load, evaluative pressure and irregular sleep patterns combine to produce chronic rather than episodic stress exposure. Despite this recognition, stress remains systematically under characterised in practice, largely because the dominant assessment instruments are introspective. Self-report inventories such as the Perceived Stress Scale, the Depression Anxiety Stress Scales, and the International Stress Management Association questionnaire require respondents to summarise an interval of days or weeks in a single score. Such instruments are susceptible to recall bias, social desirability effects, and limited temporal resolution, and by construction they cannot resolve the moment-to-moment fluctuations that characterise acute stress responses. A measurement modality that is both objective and temporally continuous is therefore required if stress monitoring is to move beyond retrospective self-assessment.

Electroencephalography offers precisely this complement. Because cortical potentials are sampled at the millisecond scale, EEG provides a non-invasive and temporally precise window into the neural correlates of stress, making it suitable for real-time monitoring applications [4], [5]. EEG-based stress analysis commonly extracts features from the delta (0.5 – 4 Hz), theta (4 – 7 Hz), alpha (8 – 12 Hz) and beta (13 – 30 Hz) frequency bands, each of which carries information related to distinct cognitive and affective states [6], [7]. Suppression of frontal alpha power, elevation of beta activity, and asymmetry between left and right prefrontal channels have each been reported as candidate markers of heightened arousal and negative valence [7], [8]. The maturation of low-cost dry-electrode consumer headsets has further extended the practical reach of EEG-based stress monitoring beyond the controlled laboratory environment [5]. Progress in this domain, however, is now constrained less by sensing hardware than by the evaluation and fusion methodology applied to the extracted features. Two problems recur across the literature. First, stress classification datasets are typically imbalanced across arousal and valence categories, and the accuracy metrics conventionally reported are

known to be optimistic under such imbalance [9], [10]. Second, when EEG features are combined with information originating from an external stimulus modality, the fusion weights are frequently fixed a priori rather than derived from the data [11], [12]. Both problems bear directly on whether reported classification performance can be regarded as trustworthy. The first is addressed analytically in this work; the second is formalised as a specification for subsequent validation. The present study addresses the first of these problems directly, and the specific motivation for re-examining MCCA can be stated plainly. The improvement previously reported for arousal classification is approximately 87 percentage points, and it was obtained by substituting new terms into an established coefficient rather than by changing the classifier, the features, or the recording protocol. A gain of that magnitude arising from a change of metric alone is more readily explained by a property of the substitution than by an increase in discriminative capacity. Establishing which of the two accounts is correct is the purpose of this work.

2. Materials and Methods

Figure 1 summarises the complete analysis pipeline, from signal acquisition through to the analytical characterisation that constitutes the principal contribution of this work. Each stage is described in the subsection indicated.

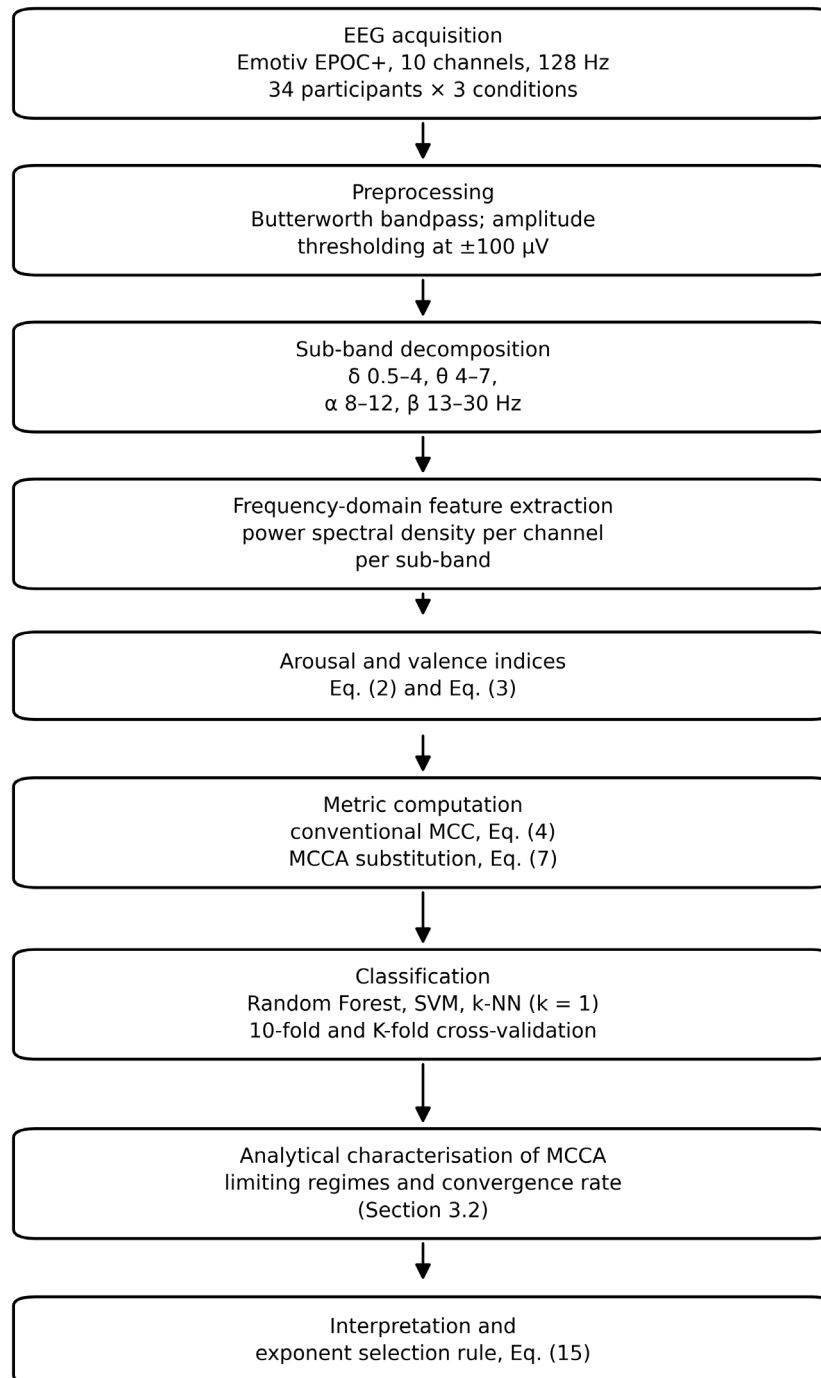


Figure 1. Top-down process flow of analysis from EEG acquisition to exponent selection rule

2.1. Data Source, Participants, and Experimental Protocol: Provenance of the Analysed Data

The experimental data analysed in this study were derived from the doctoral thesis entitled “An Enhanced Stress Indices in Signal Processing Based on Advanced Matthew Correlation Coefficient (MCCA) and Multimodal Function Using EEG Signal” by Zarith Liyana Zahari, Faculty of Electrical and Electronics Engineering Technology, Universiti Malaysia Pahang, 2023 [26]. All quantities reported in Section 3 are the archived analytical outputs of that investigation. No new signal acquisition, preprocessing, or feature extraction was performed for the present work, and the raw EEG recordings are no longer accessible. The methodological description below documents the procedure as originally executed and recorded and is presented at the level of detail supported by the surviving records. Where a parameter could not be recovered, this is stated explicitly rather than inferred, and the consequences are assessed in Section 3.6.

2.1.2. Participants

The dataset comprises EEG recordings from 34 undergraduate participants, evenly divided by sex (17 male and 17 female), aged between 18 and 30 years. Participants were recruited from the undergraduate student population of the host institution. Recruitment materials referred explicitly to stress, and the possibility of self-selection by individuals who already perceived themselves as stressed therefore cannot be excluded; this bears directly on the distribution reported in Section 3.4. Written informed consent was obtained before participation, and ethical approval for the original data collection is recorded in Section 6.2. Mean age, standard deviation, and the inclusion and exclusion criteria applied were not recorded in the surviving documentation. The sample size of 34 is modest by the standards of contemporary EEG classification studies and constrains the generality of the findings; this limitation is addressed directly in Section 3.6 rather than deferred to a concluding remark.

2.1.3. Acquisition hardware and channel selection

Signals were acquired using an Emotiv EPOC+ wireless headset. Although the device provides 14 saline-based electrode sites, 10 channels were retained for analysis: AF3, F7, F3, FC5, and T7 on the left hemisphere, and AF4, F8, F4, FC6, and T8 on the right. The four excluded sites sample the parietal and occipital cortex, regions less directly implicated in the affective and executive processes underlying the stress response than the frontal and temporal regions retained. The selected montage is symmetric across hemispheres, which is a prerequisite for the valence index defined in Section 2.3, since that index is computed as a left-right differential. Electrode positions conform to the international 10-20 system. Signals were sampled at 128 Hz, supporting a Nyquist limit of 64 Hz, which is sufficient for the four sub-bands analysed but precludes examination of high-gamma activity. Whether this rate represents the device output following internal down sampling could not be confirmed. The device layout and the measurement area analysed are shown in Figure 2.

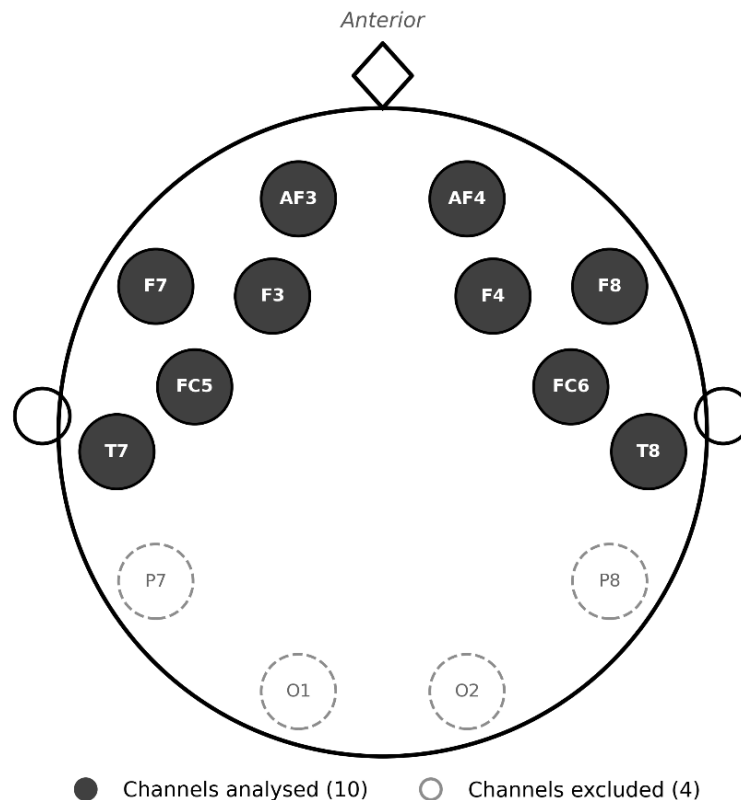


Figure 2. Emotiv EPOC+ electrode layout in the international 10-20 system, showing the ten frontal and temporal channels analysed and the four parietal and occipital channels excluded

2.1.4. Stress assessment instrument and experimental protocol

Perceived stress was assessed using the International Stress Management Association questionnaire, a 25-item self-report instrument scored from 0 to 25. Each participant completed three conditions in a fixed order: a 2-minute resting state baseline recorded with eyes open, followed by four minutes of exposure to Music Stimulus 1 and 4 minutes of exposure to Music Stimulus 2, with a 2-minute rest interval separating consecutive conditions. Both stimuli were fixed rather than participant selected, Music Stimulus 1 was a pop-punk excerpt and Music Stimulus 2 a lullaby. Because the presentation order was not counterbalanced across participants, order effects cannot be separated from stimulus effects, and the resting-state condition necessarily preceded both music conditions for every participant. This constraint is noted in Section 3.6.

2.2. Signal Processing and Feature Extraction

2.2.1. Preprocessing

Recorded signals were bandpass filtered using a Butterworth filter and decomposed into four frequency sub-bands: delta (0.5 – 4 Hz), theta (4 – 7 Hz), alpha (8 – 12 Hz), and beta (13 – 30 Hz). Artefact handling was performed by amplitude thresholding, with samples falling outside the range of $-100 \mu\text{V}$ to $+100 \mu\text{V}$ rejected. The filter order, whether zero-phase filtering was applied, the epoch length and overlap, and whether any further artefact rejection procedure was applied could not be recovered from the surviving records. The filter order is not specified in the surviving records. Because the order determines the transition-band steepness and therefore the degree of spectral leakage between adjacent sub-bands, its absence places a bound on how exactly the band power features can be reproduced; this is recorded in Section 3.6.1.

2.2.2. Feature extraction

The original investigation extracted features across frequency, time-frequency, and wavelet representations. The results reported in Section 3, and consequently the analysis presented here, are confined to the frequency domain; the time-frequency and wavelet feature sets are not analysed in this work and are not described further. Within the frequency domain, the signal power in each sub-band was computed as

$$\text{Power} = \int_{f_1}^{f_2} S_x(f) df \quad (1)$$

where $S_x(f)$ denotes the power spectral density estimate, and f_1 and f_2 are the sub-band boundaries. The specific power spectral density estimation method, for example, the averaged periodogram of Welch [27], could not be recovered from the surviving records.

2.2.3. Feature set dimensionality

21 features were extracted per channel per frequency sub-band, giving a total of 85,680 feature data points as the product of 34 participants, 3 experimental conditions, 10 channels, 21 features, and 4 sub-bands. The surviving feature documentation enumerates twenty of these: minimum, maximum, mean, root mean square, standard deviation, standard deviation error, median, mean deviation, variance, skewness, kurtosis, range of entropy values, magnitude-squared coherence, cross-power magnitude coherence, mean square error, mean square error estimate, mean absolute variation, one extended feature, and two probability measurement functions. The identity of the 21 features could not be recovered from the archived records. The reported total of 85,680 is retained here because it is the figure produced by the original analysis, had only twenty features been extracted, the product would be 81,600. The discrepancy is confined to the descriptive feature inventory and does not affect any quantity reported in Section 3, since the analysis presented here operates on the summary confusion-matrix outputs rather than on the feature matrix. It is recorded as a limitation in Section 3.6.1. No dimensionality reduction was applied prior to classification.

2.3. Arousal and Valence Indices

Rather than treating stress as a unitary construct, the analysis adopts the two-dimensional circumplex representation [28] in which affective state is characterised jointly by arousal, the intensity of activation, and valence, its positive or negative direction [8]. Both dimensions were estimated from ratios of EEG band power. Arousal was computed as the ratio of fast to slow band activity summed across channels, as given in Eq. (2):

$$\text{Arousal} = \frac{\alpha + \beta}{\delta + \theta} \quad (2)$$

The rationale is that cortical activation is accompanied by a shift of spectral power towards higher frequencies, with alpha desynchronisation and beta enhancement occurring together during states of heightened alertness or stress [7]. Valence was computed as the difference in this ratio between the right and left hemispheres, as given in Eq. (3):

$$\text{Valence} = \frac{\alpha_R + \beta_R}{\delta_R + \theta_R} - \frac{\alpha_L + \beta_L}{\delta_L + \theta_L} \quad (3)$$

where the subscripts R and L denote the right-hemisphere channels (AF4, F8, F4, FC6, T8) and left-hemisphere channels (AF3, F7, F3, FC5, T7), respectively. This differential form derives from the frontal alpha asymmetry model [33], in which relatively greater left-frontal activation is associated with approach-oriented and positively valenced states, and relatively greater right-frontal activation with withdrawal-oriented and negatively valenced states [8].

Two properties of these indices are material to the analysis that follows and have not previously been made explicit. First, both are ratios of strictly positive power quantities and are therefore unbounded above; the arousal index has no

theoretical maximum. Second, the denominator may become small in epochs dominated by fast activity, in which case the ratio grows without bound. Neither property is problematic when the indices are used descriptively. Both become consequential when the indices are substituted into a coefficient originally defined over bounded, non-negative integer counts, as described in Section 2.4.

2.4. MCCA Formulation and Power-Parameter Extension

2.4.1. The conventional Matthews correlation coefficient

The Matthews correlation coefficient summarises a binary confusion matrix as a single correlation-like quantity [9], [29]:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

Three properties of this definition are essential to what follows. First, the four arguments are non-negative integer counts. Second, given that domain, the coefficient is bounded on the interval from minus one to plus one, attaining plus one for perfect classification, minus one for perfect inverse classification, and zero for performance equivalent to chance. Third, the bound is a consequence of the domain: it follows from the Cauchy–Schwarz inequality applied to the count representation and is not preserved automatically if the arguments are replaced by quantities drawn from a different domain.

2.4.2. The MCCA substitution

The MCCA formulation replaces the false-positive and false-negative counts with arousal-valence-derived quantities, as defined in Eqs. (5) and (6):

$$FP_a = \frac{\text{Arousal}_{FP}}{\text{Valence}_{FP}} \quad (5)$$

$$FN_a = \frac{\text{Arousal}_{FN}}{\text{Valence}_{FN}} \quad (6)$$

Substituting these into Eq. (4) and introducing a power exponent n gives the generalised statistic:

$$MCCA_n = \frac{TP \times TN - FP_a^{(n)} \times FN_a^{(n)}}{\sqrt{(TP + FP_a^{(n)})(TP + FN_a^{(n)})(TN + FP_a^{(n)})(TN + FN_a^{(n)})}} \quad (7)$$

The original study evaluated the exponent at 2, 4, and 6, selecting 6 by analogy with six-sigma reasoning [18].

2.4.3. Domain consequences of the substitution

The substitution alters the domain of the coefficient in three respects, none of which was addressed in the original formulation. First, as established in Section 2.3, the arousal index is a ratio of strictly positive band powers and is unbounded above, whereas the valence index is a left–right difference and may therefore be positive, negative, or zero. The quotients defined by Eqs. (5) and (6) are consequently real valued rather than non-negative, are unbounded in magnitude, and are undefined where the valence index vanishes. The treatment of such epochs in the original implementation could not be recovered from the surviving records.

Second, because the substituted terms may be negative, the parity of the exponent is not a free choice. For odd exponents, the substituted terms retain their sign, and a sufficiently negative term can render a factor beneath the radical negative, making the statistic complex. For even exponents, the terms are guaranteed non-negative, and the radicand remains well defined. The exclusive use of even exponents in the original study is therefore not incidental; it is a necessary condition for the statistic to remain real valued. This constraint has not previously been stated. Third, once the count domain is abandoned, the bound on the interval from minus one to plus one no longer follows from Eq. (4) and must be re-established, if it holds at all.

2.4.4. The analytical question

Because the substituted terms are unbounded and are raised to a power, the behaviour of Eq. (7) as the exponent increases is not self-evident. Whether the exponent sharpens the discrimination of the statistic, as the six-sigma motivation implies, or instead drives it towards a degenerate limit, is a question that can be settled analytically without reference to any dataset. Section 3.2 provides that analysis. It is the principal contribution of the present work, and it bears directly on how the accuracy differentials reported in Section 3.1 should be interpreted.

2.5. Enhanced Multimodal Fusion

The multimodal formulation employed in the original study combined the EEG and music-stimulus streams using a fixed weight, as given in Eq. (8):

$$P_{\text{multimodal}}^x = \alpha p_{\text{eeg}}^x + (1 - \alpha) p_{\text{music}}^x, \quad \alpha \in \{0, 1, 2\} \quad (8)$$

Two difficulties arise. The admissible set is discrete and admits only three configurations, of which two are degenerate: a weight of zero discards the EEG stream entirely, and a weight of one discards the music stream. Furthermore, a weight of two yields a coefficient of minus one on the music term, which does not correspond to a convex combination and lacks

an interpretation as a weighting. In effect, the original formulation permitted no genuine fusion. To remove these difficulties, a fusion over a normalised simplex is proposed in Eq. (9):

$$P_{\text{enhanced}}^x = \alpha^* p_{\text{eeg}}^x + \beta^* p_{\text{music}}^x + (1 - \alpha^* - \beta^*) g_{\text{mcca}} \quad (9)$$

subject to α^* greater than or equal to zero, β^* greater than or equal to zero, and the sum of α^* and β^* not exceeding one, so that all three coefficients are non-negative and sum to unity. The third term introduces the MCCA statistic as a fusion stream rather than as an evaluation metric applied after fusion.

Calibration of Eq. (9) requires a two-dimensional grid search over the admissible region, with each weight traversing the unit interval in increments of 1/10 and configurations that violate the simplex constraint discarded, yielding 66 admissible combinations. Selection should be performed on a validation partition held out from both training and test data, using subject-wise partitioning for the reasons set out in Section 2.6. This calibration was not performed in the present work. The required data are unavailable, and Eq. (9) is therefore offered as a specification for subsequent validation rather than as a validated result. No performance claim is made for it.

2.6. Classification and Validation

2.6.1. Classifiers

Three classifiers were evaluated: Random Forest, a support vector machine, and k-nearest neighbours. The selection spans an ensemble method, a margin-based method, and an instance-based method, which supports assessment of whether observed performance reflects the feature representation or the inductive bias of a particular learner. The k-nearest neighbours classifier was configured with k equal to one and Euclidean distance. Hyperparameters for the Random Forest and support vector machine, and any feature normalisation applied before classification, could not be recovered from the surviving records. The hyperparameter settings for the Random Forest and the support vector machine, and the confusion matrices from which the reported accuracies and coefficient values were computed, were not preserved in the archived records and cannot be reconstructed from the summary outputs. Their absence is material rather than incidental: the true-positive and true-negative counts enter the analysis of Section 3.2 directly, and without them the regime occupied by the data cannot be determined. This is recorded in Section 3.6.1.

2.6.2. Validation design

Performance was assessed using 10-fold and K-fold cross-validation. The value of K employed in the separately reported K-fold condition could not be recovered. The rationale for selecting 10 folds was not documented. 10-fold partitioning is the conventional default in this literature and represents a compromise between the pessimistic bias associated with few folds and the variance and computational cost associated with leave-one-out partitioning. With 34 participants, 10 folds place three or four participants in each held-out partition under participant-level partitioning, which is close to the smallest evaluation unit for which a fold-level accuracy remains interpretable; fewer folds would leave too few training participants, and more folds would produce held-out sets of one or two participants whose accuracy estimates would be dominated by individual differences. Independent analyses have shown that cross-validation on small samples yields error bars far wider than commonly assumed [19] and that validation schemes differ markedly in bias when sample sizes are limited [20]; both findings apply directly to the present design. This rationale is reconstructed from the design rather than documented in the original records, and the choice should be read as conventional rather than as optimised for this dataset.

The partitioning unit is decisive for interpretation. Barry et al. [30] demonstrated that where folds are constructed by randomly partitioning epochs, segments originating from the same participant and from temporally adjacent recording intervals are distributed across training and test partitions. The classifier may then identify participant-specific spectral signatures rather than stress-related activity, and the resulting accuracy estimate reflects within-subject recognition rather than generalisation to unseen individuals. Whether partitioning was performed at the epoch level or the subject level could not be established from the surviving records. Given a dataset derived from 34 participants, epoch-level partitioning would place data from every participant in every fold. Under that condition, the reported accuracies estimate within-subject performance and should not be read as estimates of cross-subject generalisation. Because the partitioning unit is unconfirmed, the more conservative reading is adopted throughout Section 3.

2.6.3. Proposed six-level stress categorisation

The ISMA instrument yields three ordinal stress classes: low, medium, and high. A refinement to six ordinal levels is proposed here, from L1 (no stress) to L6 (extreme stress), obtained by subdividing each of the three ISMA classes into two via fractional scoring. The refinement is intended to provide finer descriptive granularity than the three-level scheme adopted in earlier work [31], which assigns a substantial majority of any general cohort to a single intermediate class. Analyses were performed in MATLAB (MathWorks, Natick, Massachusetts, United States). The subdivision is presented as a methodological proposal rather than as a validated instrument. The score boundaries separating the two halves of each ISMA class could not be recovered from the archived records, and no independent criterion, such as a concurrent physiological or clinical measure, was available against which the six-level assignment could be validated. The scheme should therefore be regarded as a hypothesis requiring future validation, and the distribution reported in Section 3.4 as descriptive of this cohort rather than as evidence that six levels are separable.

3. Results and Discussion

The results are organised by objective. Section 3.1 presents the reported MCC and MCCA values that motivate the analysis, and Section 3.2 addresses Objective 1 and RQ1 by deriving and numerically verifying the limiting behaviour of MCCA. Sections 3.3 and 3.4 address Objective 2 and RQ2 by establishing the precision of the reported classifier performance and the conditions under which it may be generalised. Section 3.5 draws the two strands together.

3.1. MCC Versus MCCA Performance

Table 1 compares conventional MCC and MCCA performance in the frequency-domain analysis. The visualisation of these values is shown in Figure 3. Under conventional MCC, arousal classification performs at 10.1%, 11.7%, and 12.0% for the resting state, Music 1, and Music 2 conditions, respectively, while valence classification performs at 93.7% to 97.2% across the same conditions. Under MCCA, arousal rises to between 98.3% and 99.4%, and valence to between 99.1% and 99.4%. Three observations follow, of which the third has not previously been made.

First, the differential is confined almost entirely to arousal. Valence classification was already performing above 93% under conventional MCC and gains between two and six percentage points under MCCA. Arousal, by contrast, moves from a level indistinguishable from chance to near ceiling. Any account of the mechanism must therefore explain why the substitution affects one index and not the other. Second, the final column of Table 1, labelled Difference, reports the arithmetic difference between 2 percentages and is therefore expressed in percentage points. It was labelled Improvement in the original analysis, and that description is not sustainable: the two quantities being differenced are not two measurements of the same construct obtained by different methods, but two different statistics, one defined over confusion matrix counts and the other over arousal-valence ratios, as established in Section 2.4.2. A difference between them does not carry the interpretation that the word improvement would imply, and the column has been relabelled accordingly.

Third, and most informative, the MCCA values exhibit markedly less variation across experimental conditions than the MCC values they replace. The three MCCA arousal figures span 1.1 percentage points, and the three MCCA valence figures span 0.3 percentage points, whereas the corresponding MCC figures span 1.9 and 3.5 points, respectively. This compression is difficult to reconcile with a metric that remains sensitive to classifier performance, since the resting state and music stimulus conditions are intended to elicit materially different affective states and would be expected to yield materially different classification difficulty. It is, however, precisely what Section 3.2.2 predicts for a statistic operating near a saturation limit: once the limiting value is approached, differences in the underlying confusion matrix are compressed towards a common value. The clustering of all six MCCA figures within a band of approximately one percentage point is therefore consistent with saturation rather than with uniformly excellent classification, and the figures should not be read as evidence of the latter without the additional reporting specified in Section 3.2.6.

Table 1. Conventional MCC and MCCA values across frequency-domain conditions

Domain	Condition	MCC arousal value (%)	MCCA arousal value (%)	MCC valence value (%)	MCCA valence value (%)	Difference (pp)
Frequency	Rest	10.1	99.4	97.2	99.3	+89.3
Frequency	Music 1	11.7	98.5	97.2	99.1	+86.8
Frequency	Music 2	12.0	98.3	93.7	99.4	+86.3

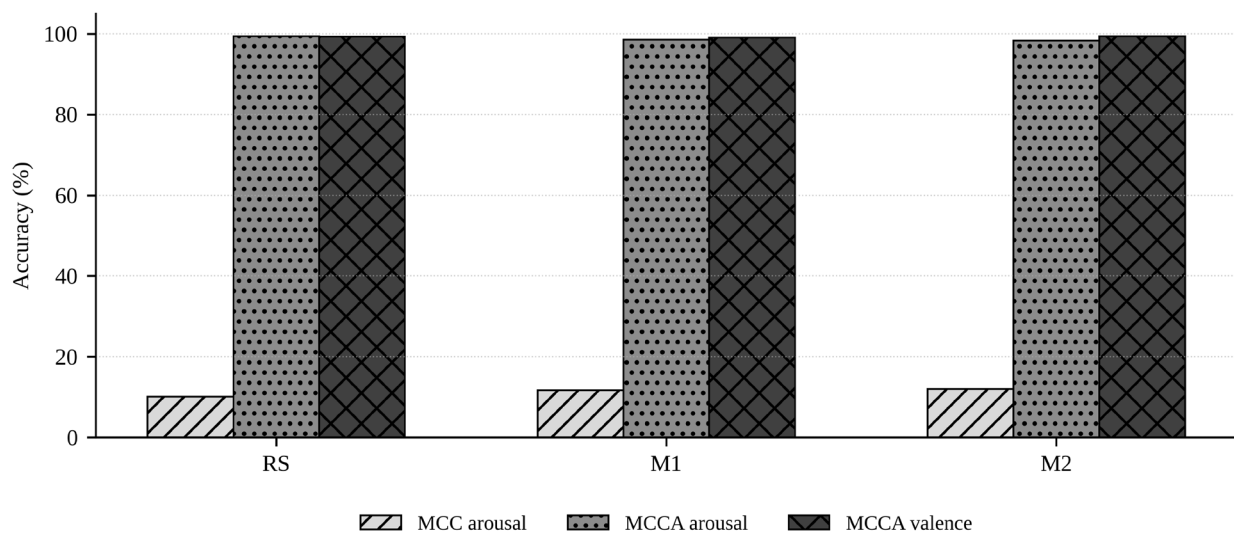


Figure 3. Conventional MCC and MCCA values for the frequency-domain PSD analysis

3.2. Analytical Behaviour of the Generalised MCCA Statistic

3.2.1. Reduction to two variables

The derivation that follows rests on five assumptions, stated here so that each step can be checked independently. First, $TP > 0$ and $TN > 0$; where either count is zero, the denominator of Eq. (7) vanishes and the statistic is undefined, as it is for conventional MCC. Second, the exponent n is even, which Section 2.4.3 shows to be necessary rather than conventional, since odd exponents may render the radicand negative. Third, the substituted quantities p and q are finite, which requires the valence index to be non-zero; epochs at which it approaches zero are excluded from the domain of the statistic, and their treatment in the original implementation could not be recovered. Fourth, TP and TN are held fixed as n varies, so that the limit describes the behaviour of the metric on a given confusion matrix rather than the behaviour of a classifier. Fifth, the generic case $|p|$ is not equal to 1 and $|q|$ is not equal to 1 is assumed throughout Sections 3.2.2 and 3.2.4; the boundary cases are treated separately at the end of Section 3.2.2 and yield a constant that depends on TP and TN rather than a limit of +1, -1, or 0.

Let a denote TP and b denote TN , both strictly positive, and let p and q denote the quantities defined by Eqs. (5) and (6). Write P for the absolute value of p and Q for the absolute value of q . For even exponents, as required by Section 2.4.3, the exponentiated terms equal P raised to n and Q raised to n respectively, both non-negative. Substituting u for P to the power n and v for Q to the power n , Eq. (7) reduces to

$$MCCA_n = \frac{ab - uv}{\sqrt{(a+u)(a+v)(b+u)(b+v)}} \quad (10)$$

The behaviour of the statistic in n is therefore governed entirely by how u and v evolve, and since these are powers of P and Q , the controlling quantities are the positions of P and Q relative to unity.

3.2.2. Limiting behaviour

Three regimes arise as the exponent increases without bound. Regime I, in which both P and Q are less than one. Both u and v tend to zero. The numerator of Eq. (10) tends to the product ab , and the denominator tends to the same product. The statistic therefore tends to plus one. Regime II, in which both P and Q exceed one. Both u and v grow without bound. The numerator is dominated by the negative of the product uv , and the denominator by the same product. The statistic therefore tends to minus one. Regime III, in which P and Q straddle unity. Taking P greater than one and Q less than one without loss of generality, u grows without bound while v tends to zero. The denominator is dominated by u multiplied by the square root of ab . Three sub-cases arise according to the product PQ . Where that product exceeds one, the numerator behaves as the negative of PQ raised to n , and the quotient tends to the negative of Q raised to n divided by the square root of ab , which vanishes. Where the product is less than one, the numerator tends to ab and the quotient tends to the square root of ab divided by P raised to n , which also vanishes. Where the product equals one, the numerator tends to ab minus one and the quotient again vanishes. In all sub-cases the statistic tends to zero. The boundary cases in which P or Q equals unity exactly are non-generic. Where P equals one and Q is less than one, for instance, u remains equal to one for every exponent and the statistic converges to the square root of ab divided by the square root of the product of a plus one and b plus one, a constant strictly between zero and one that depends on the confusion matrix counts.

3.2.3. Interpretation

The consequence of Section 3.2.2 is that, as the exponent increases, the statistic ceases to be a function of classification performance. In Regime I the limit is plus one irrespective of the values of TP and TN ; in Regime II the limit is minus one on the same terms; in Regime III the limit is zero. The true positive and true negative counts, which are the only arguments of Eq. (10) carrying information about classifier behaviour, are eliminated in the limit. What the statistic converges to instead is a three-valued indicator of whether the arousal to- valence ratios in the false-positive and false-negative partitions exceed unity in magnitude. That is a property of the recorded signal and of the arousal and valence index definitions, not of the classifier applied to it. A high value of the statistic at large exponent is therefore not evidence of accurate classification, and the quantity should not be interpreted as a classification metric in the sense established for MCC by Chicco and Jurman [9].

3.2.4. Rate of convergence

The limits established above describe the asymptotic behaviour of the statistic but not the exponent at which that behaviour becomes apparent. Since the exponents of practical interest are small, the rate of approach is the operative question. The analysis simplifies considerably in the symmetric case in which P and Q are equal, for which u equals v and the radicand becomes a perfect square. Equation (10) then reduces to the closed form

$$MCCA_n = \frac{ab - u^2}{(a+u)(b+u)}, \quad u = P^n \quad (11)$$

Two consequences follow directly. First, the statistic changes sign when u equals the square root of ab , that is, when

$$n_0 = \frac{\ln(\sqrt{TP \cdot TN})}{\ln(P)} \quad (12)$$

The exponent at which the statistic crosses zero is therefore determined jointly by the arousal-valence ratio and by the geometric mean of the true-positive and true-negative counts. It is not a property of the metric alone.

Second, the deviation from the limiting value behaves asymmetrically between the two regimes. Expanding Eq. (11) for small u gives, in Regime I,

$$1 - \text{MCCA}_n \approx P^n \left(\frac{1}{a} + \frac{1}{b} \right) \quad (13)$$

whereas expanding for large u gives, in Regime II,

$$\text{MCCA}_n + 1 \approx \frac{a + b}{P^n} \quad (14)$$

In Regime I the confusion matrix counts appear in the denominator, so convergence is accelerated as those counts grow. In Regime II they appear in the numerator, so convergence is retarded by the same growth. Larger datasets therefore reach the asymptotic limit more quickly in one regime and more slowly in the other.

The practical contrast is pronounced. Evaluating Eq. (11) with counts of fifty and a ratio of 0.7, representative of Regime I, the statistic reaches 0.995 at an exponent of six and changes only in the third decimal place at higher exponents. Evaluating the same expression with a ratio of 1.5, representative of Regime II, the statistic at an exponent of six is approximately 0.63, retaining not merely the wrong magnitude but the wrong sign relative to its limit, and does not cross zero until an exponent of approximately 9.6, consistent with Eq. (12).

3.2.5. Implications for the selection of the power exponent

Three conclusions follow, and they revise rather than confirm the position advanced in the original study. First, no single exponent is universally appropriate. The value required for the statistic to attain its asymptotic behaviour scales logarithmically with the size of the confusion matrix, as Eqs. (12) and (14) make explicit. A value adequate for one dataset will be inadequate for a larger one. The six-sigma motivation [26], [18] provides no basis for this scaling, since six-sigma reasoning concerns tolerance limits in process control and carries no implication for the exponent of a correlation statistic. The choice is therefore best regarded as arbitrary with respect to the mathematics, whatever its heuristic appeal.

Second, a principled selection rule is available in its place. For the statistic to lie within a tolerance $\varepsilon > 0$ of its limits in Regime II, Eq. (14) requires

$$n > \frac{\ln\left(\frac{TP + TN}{\varepsilon}\right)}{\ln(P)}, \quad P > 1 \quad (15)$$

with the corresponding condition in Regime I following from Eq. (13). Reporting the exponent alongside the confusion matrix totals and the observed range of the substituted ratios would allow a reader to verify that the condition is satisfied. No published application of MCCA currently reports these quantities.

Third, the resolution of the first objective of this study is conditional rather than absolute. Where the data occupy Regime I, the original omission of higher exponents entails no loss, since the statistic is already saturated at an exponent of six and additional exponents cannot alter it materially. Where the data occupy Regime II or III, the omission is consequential: the evaluations in Section 3.2.4 show that extending the sweep would change the statistic substantially, and in Regime II would reverse its sign. Determining which regime applies requires the empirical distribution of the substituted ratios in the original dataset. Those values were not preserved in the surviving analytical records, and the question cannot be settled here. It is stated as an open question in Section 4 rather than resolved.

3.2.6. Recommendations for the use of MCCA

Four recommendations follow for subsequent work. The conventional MCC should be reported alongside MCCA on the same confusion matrix, so that the contribution of classifier performance can be distinguished from that of the arousal valence substitution. The observed range of the substituted ratios should be reported, since the regime, and hence the interpretation of the statistic, cannot be determined without it. The exponent should be selected using Eq. (15) rather than by convention. And where MCCA is retained, it is more defensibly described as a signal-derived affective index than as a measure of classification quality, for the reason given in Section 3.2.3.

3.2.7. Numerical verification of the limiting behaviour

The derivation in Sections 3.2.2 and 3.2.4 was verified numerically using synthetic confusion matrices, following the practice adopted for metric characterisation by Chicco and Jurman [9], [23]. No EEG data enter this verification. The true positive and true negative counts are assigned fixed values, and the substituted ratios P and Q are assigned representative values in each of the three regimes, so that the quantity evaluated is the function defined by Eq. (10) rather than any empirical result. The purpose is to confirm that the algebra is correct and to make the rate of convergence visible. Table 2 reports MCCA evaluated with $TP = TN = 50$ across even exponents from 2 to 20, and Figure 4 extends the same evaluation to an exponent of 40. The three regimes converge to +1, -1, and 0, respectively, confirming Section 3.2.2. The closed form of Eq. (11) reproduces the general expression of Eq. (10) exactly in the symmetric case, and the sign change in Regime II occurs at an exponent of 9.65, matching the value predicted by Eq. (12). The practical implication is visible at the exponent of six used in the original study, marked in Figure 4. In Regime I, the statistic has already reached 0.995 and lies within half a percentage point of its limit, so further increases are immaterial. In Regime II it stands at 0.629, retaining the opposite sign to its limiting value of -1, which it does not approach until an exponent beyond 20. An exponent adequate in one regime is therefore not adequate in another, which is what Eq. (15) formalises. The verification

also shows that the loss of dependence on TP and TN is not a remote asymptotic curiosity: by an exponent of 14 the Regime II value is -0.708 regardless of how well the classifier performed.

3.3. Classifier Performance

Table 3 summarises the reported classifier results for the selected frequency-domain MCCA configuration. Random Forest attains the highest 10-fold accuracy at 96.8%, followed by the support vector machine at 95.2% and k-nearest neighbours at 80.6%. Under the separately reported K-fold condition, the corresponding figures are 85.0%, 85.0%, and 72.5%. The ten-fold accuracies are compared in Figure 5. Values correspond to 30/31, 59/62, and 25/31 correctly classified instances, respectively, inferred from the reported percentages. 95% Wilson score intervals [32] are 83.8 – 99.4% (Random Forest), 86.7 – 98.3% (SVM), and 63.7 – 90.8% (k-NN). The evaluation partition comprises tens of instances; the figure of 85,680 in Section 2.2.3 denotes extracted features, not independent observations.

Table 2. Numerical evaluation of Eq. (10) with $TP = TN = 50$ for the three limiting regimes

Exponent n	Regime I ($P = 0.7, Q = 0.7$)	Regime II ($P = 1.5, Q = 1.5$)	Regime III ($P = 1.5, Q = 0.7$)
2	0.981	0.914	0.947
4	0.990	0.816	0.903
6	0.995	0.629	0.812
8	0.998	0.322	0.660
10	0.999	-0.071	0.464
12	0.999	-0.444	0.278
14	1.000	-0.708	0.146
16	1.000	-0.859	0.071
18	1.000	-0.935	0.033
20	1.000	-0.970	0.015
Limit	+1	-1	0

Table 3. Classifier performance for the frequency-domain MCCA configuration

Domain	Method	Classifier	10-fold (%)	K-fold (%)
Frequency	MCCA	Random Forest	96.8	85.0
Frequency	MCCA	Support vector machine (SVM)	95.2	85.0
Frequency	MCCA	k-nearest neighbors (k-NN)	80.6	72.5

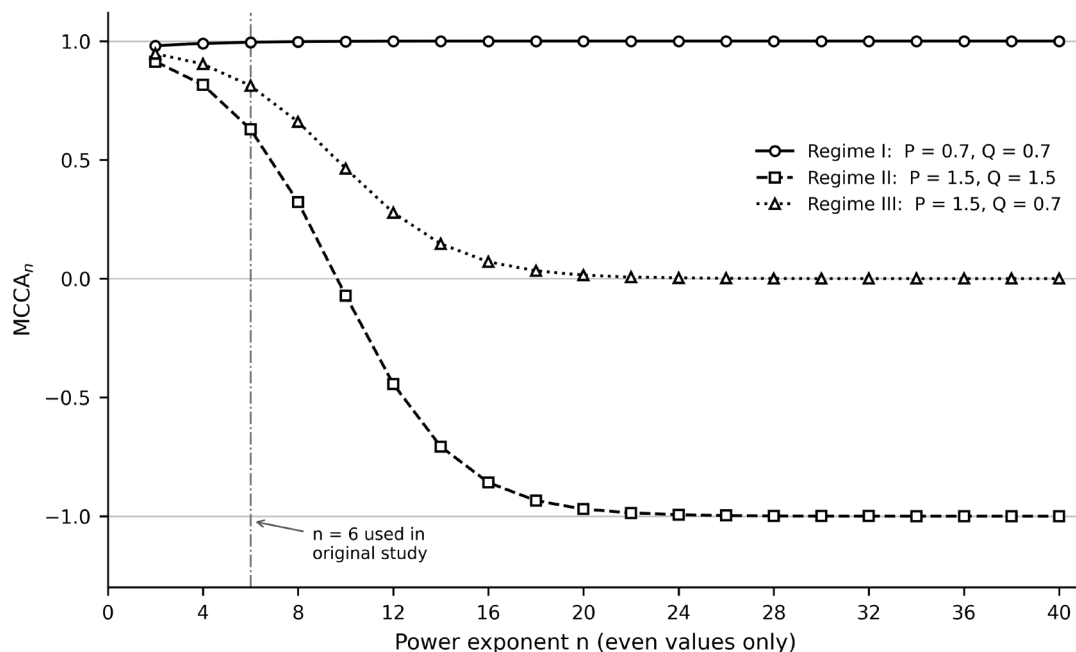


Figure 4. Numerical evaluation of Eq. (10) against the power exponent for the three limiting regimes, with $TP = TN = 50$

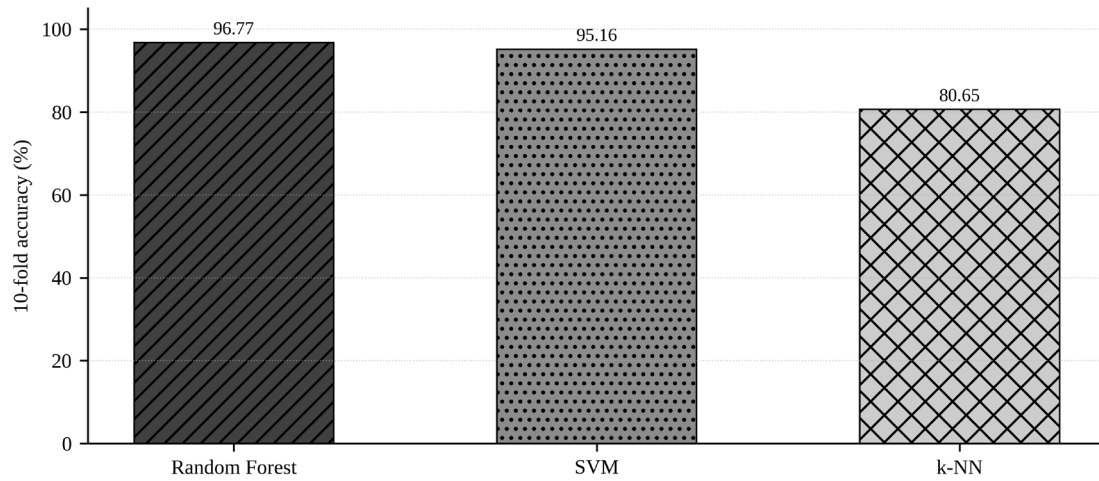


Figure 5. Ten-fold classifier accuracy for the frequency-domain MCCA configuration

3.3.1. Effective evaluation sample size

The reported percentages are informative about the size of the evaluation partition. The value 96.8% is consistent with 30 out of 31, 95.2% to 59 out of 62, and 80.6% to 25 out of 31. The exact partition size could not be confirmed against the surviving records, but the correspondence of all three reported values to fractions with these denominators is unlikely to be coincidental. If the inference holds, it carries a substantial implication. A test partition of approximately 31 instances is not consistent with classification performed over 85,680 individual feature points; it is consistent instead with classification performed over aggregated per participant or per condition representations, of which there are 34 and 102 respectively. The effective sample size available for performance estimation is therefore of the order of tens, not tens of thousands. The figure of 85,680 describes the volume of extracted features, not the number of independent observations available for validation, and the two should not be conflated in reporting.

3.3.2. Uncertainty in the reported accuracies

On the inferred denominators, 95% Wilson score intervals for the 10-fold accuracies are approximately 83.8% to 99.4% for Random Forest, 86.7% to 98.3% for the support vector machine, and 63.7% to 90.8% for k-nearest neighbours. Two conclusions follow. The intervals for Random Forest and the support vector machine overlap across almost their entire range, and the difference of 1.6 percentage points between them corresponds to approximately one test instance. The two classifiers are therefore not distinguishable on these data, and no ranking between them is asserted. The k-nearest neighbours result is lower, but its interval remains wide and overlaps the lower bound of the Random Forest interval; the separation is suggestive rather than established.

3.3.3. Evidence bearing on the partitioning unit

The k-nearest neighbours classifier was configured with k equal to one. Under epoch-level partitioning with overlapping or temporally adjacent segments, a one nearest neighbour classifier is the configuration most susceptible to inflation by leakage, since a near-duplicate of any test segment would frequently appear in the training partition and would be selected as its nearest neighbour. That this classifier returned the lowest accuracy of the three, at 80.6%, is therefore mild evidence against gross leakage of that kind. The observation does not exclude leakage at the participant level, nor does it establish the partitioning unit, which remains unconfirmed. It is offered as the only evidence internal to the reported results that bears on the question.

3.3.4. The discrepancy between validation schemes

The gap between the 10-fold and K-fold results is substantial: 11.8 percentage points for Random Forest and 10.2 for the support vector machine. A discrepancy of this magnitude between two cross-validation schemes applied to identical features and identical classifiers indicates that the schemes are not partitioning the data equivalently. The most common source of such a discrepancy is a difference in partitioning unit, with the higher figure arising from epoch-level partitioning and the lower from partitioning that respects participant boundaries [30]. That both Random Forest and the support vector machine return exactly 85.0% under the K-fold condition, despite differing under 10-fold, is consistent with a coarser partition in which both classifiers misclassify the same small number of instances. Where the partitioning unit differs between the two schemes, the lower figure is the more conservative estimate of generalisation to unseen participants, and it is the figure that should be cited when the method is characterised for deployment.

3.4. Six-Level Stress Classification

Table 4 summarises the distribution of the 34 participants across the six proposed levels: 23% at L3 denoting mild stress, 44% at L4 denoting moderate stress, 21% at L5 denoting severe stress, and 12% at L6 denoting extreme stress. No participant was assigned to L1 or L2. The distribution is internally consistent with the subdivision structure described in

Section 2.6.3. Aggregating the paired levels recovers the three underlying ISMA classes: no participant falls within either subdivision of the low class, 67% fall within the two subdivisions of the medium class, and 33% within those of the high class. The refinement therefore separates a cohort that the three-level instrument would have described as two-thirds medium and one-third high into four populated categories, which is the descriptive gain the proposal was intended to deliver.

The absence of observations at the two lowest levels nonetheless warrants direct comment. Recruitment materials referred explicitly to stress, as recorded in Section 2.1.2, so preferential participation by individuals who already perceived themselves as stressed is a plausible and partially documented contributor. Miscalibration of the instrument thresholds against an undergraduate cohort, and genuinely elevated baseline stress in that population, remain possible additional contributors and cannot be separated from the sampling effect on these data. Whichever explanation predominates, the operational consequence is the same. A classifier trained on this dataset receives no examples of L1 or L2 and cannot learn to identify them. The scheme is therefore four-level in practice, and any deployment claim should be scoped to the L3 to L6 range even though the categorisation itself defines six.

The distribution across the four populated levels is also uneven, with the moderate level accounting for 44% of participants and the extreme level for 12%. Under this imbalance, the accuracy figures reported in Section 3.3 are inflated relative to a balanced design, since a classifier predicting the majority level for every instance would achieve 44% without learning anything. This is precisely the condition under which correlation-based metrics are preferred to accuracy [9], [10], and it strengthens rather than weakens the rationale for adopting a coefficient-based evaluation, provided the coefficient retains its dependence on the confusion matrix. Regarding sex differences, three of the four participants at the extreme level are female, and one is male. With four observations in the cell, no inferential test has meaningful power, and the marginal totals are identical by design at 17 participants of each sex. No conclusion regarding sex differences in stress susceptibility can be drawn from this sample, and none is offered.

The relationship between the extracted EEG features and the assigned stress level is indirect in this design and warrants explicit statement. The stress level is determined by the ISMA questionnaire response, not by the EEG signal, the band-power features enter only as the classifier input against which that self-reported label is predicted. The distribution in Table 4 therefore characterises the cohort as described by self-report. The contribution of individual features to the assignment cannot be quantified from the archived records, which contain neither feature importance rankings nor per-feature or per-band performance. What can be stated is that the frequency-domain features supported prediction of those labels at the accuracies given in Section 3.3, and that the arousal and valence ratios formed from the same delta, theta, alpha, and beta band powers determine the MCCA values in Section 3.1 through the mechanism established in Section 3.2. Because those ratios enter the coefficient directly, the band definitions and the channel montage used to form the left-right difference propagate into the reported MCCA figure in a way they do not propagate into conventional MCC.

3.5. Discussion

3.5.1. An account of the reported performance differential

Section 3.1 identified a feature of Table 1 requiring explanation: the substitution of arousal valence terms improves arousal classification by approximately 87 percentage points while improving valence classification by between two and six. Any mechanism proposed for the MCCA gain must accommodate this asymmetry. The saturation behaviour established in Section 3.2 supplies such a mechanism, and does so without appeal to any improvement in classification. If the statistic converges towards plus one in Regime I irrespective of the confusion matrix, then the apparent gain observed for any index is bounded by the distance of that index from the saturation limit before substitution. Valence was already performing between 93.7% and 97.2% under conventional MCC and had at most six percentage points of headroom; arousal was performing near chance and had nearly ninety. The observed asymmetry is therefore not merely compatible with saturation but is quantitatively what saturation predicts. An account based on genuine improvement in discriminative capacity would require an explanation of why the substitution should assist arousal classification so specifically, and no such explanation is available from the arousal and valence definitions themselves, which enter the two computations symmetrically. This interpretation is reinforced by the compression of variance across experimental conditions noted in Section 3.1 and by the near identity of the six MCCA figures. Taken together, the three observations point consistently towards a statistic operating close to a fixed point rather than tracking classification quality.

Table 4. Six-level stress distribution across 34 participants

Level	Description	Subjects (%)	Male	Female
L1	No stress	0	0	0
L2	Modest	0	0	0
L3	Mild	23	4	4
L4	Moderate	44	9	6
L5	Severe	21	3	4
L6	Extreme	12	1	3

3.5.2. Implications for metric modification in biomedical classification

The finding generalises beyond the particular statistic examined here. The Matthews correlation coefficient is recommended for imbalanced classification precisely because its value is constrained by all four cells of the confusion matrix [9], [10]. That constraint is what makes it difficult to obtain a high value without genuinely good classification, and it is a consequence of the coefficient's algebraic structure over count data, not an incidental property. Substituting terms drawn from a different domain preserves the outward form of the coefficient while removing the guarantee that motivated its adoption. The general lesson is that a modified metric requires the same analytical characterisation that established the properties of the original. Bounds, monotonicity, and behaviour under limiting conditions cannot be assumed to transfer. Where a modification produces a large improvement in reported performance, the possibility that the improvement is an artefact of the modification should be excluded analytically before it is attributed to the underlying method. The present case suggests that this check is not routine, and the absence of such analysis in the metric modification literature is itself a gap worth noting.

3.5.3. What the original analysis does support

The foregoing does not invalidate the wider investigation from which these data derive. The feature extraction pipeline described in Section 2.2 follows established practice [4], [6], [33], and the conventional MCC results in Table 1 are computed over the standard domain and retain their usual interpretation: they indicate that valence was classified well and arousal poorly under the feature set employed. That finding is informative in its own right, and the poor arousal performance it records is the observation that motivated the MCCA development in the first place. The classifier comparison in Table 3, subject to the precision constraints of Section 3.3.2, indicates that accuracies above 95% are attainable on this feature representation under the validation scheme employed. The six-level categorisation proposed in Section 2.6.3 provides a more granular description of the cohort than the three-level instrument it refines, notwithstanding the two unpopulated levels.

3.5.4. Relation to current practice

The direction indicated by this analysis is consistent with trends in the wider literature. The movement away from fixed fusion weights towards adaptive schemes [11] to [31] reflects a growing preference for parameters derived from data over parameters fixed by convention, and the six-sigma selection of an exponent is an instance of the latter. Similarly, the increasing attention to leakage in cross-validation design for EEG [30] reflects recognition that reported accuracy in this field has often exceeded what the underlying evaluation could support. The present findings suggest that metric definition warrants comparable scrutiny.

3.6. Limitations

3.6.1. Unavailability of the primary data

The principal limitation of this study is that the raw EEG recordings underlying the analysis are no longer retrievable. The surviving records comprise the summary outputs reported in Section 3 together with the methodological description in [26]. Three consequences follow. The empirical distribution of the substituted ratios cannot be recovered. As Section 3.2.5 established, the regime in which the statistic operates, and hence whether an exponent of six is adequate, depends entirely on whether these quantities exceed unity in magnitude. The objective is therefore resolved analytically but not empirically, and which regime applies to this dataset remains open. Several methodological parameters could not be confirmed against the original records. These comprise the cross-validation partitioning unit, the interpretation of the conventional MCC values in Table 1, the handling of epochs at which the valence index approaches zero, the mean and standard deviation of participant age, the inclusion and exclusion criteria, the point in the academic calendar at which recording took place, the filter order and phase characteristic, the epoch length and overlap, the power spectral density estimation method, the identity of the 21 extracted feature, the Random Forest and support vector machine hyperparameters and their confusion matrices, any feature normalisation applied, the value of K in the separately reported K-fold condition, the score boundaries mapping ISMA responses to the six proposed levels; the MATLAB release, and the exact size of the evaluation partition. The description in Section 2 is consequently insufficient for exact replication. This is stated plainly rather than obscured. The enhanced fusion formulation specified in Section 2.5 could not be calibrated, and no performance claim is made for it.

3.6.2. Precision of the performance estimates

The accuracy figures in Section 3.3 rest on an evaluation partition inferred to contain approximately 31 instances. At that size, the 95% confidence interval on an observed accuracy of 96.8% extends from 83.8% to 99.4%, a width exceeding fifteen percentage points. Differences between classifiers of one or two percentage points cannot be resolved, and the ordering in Table 3 should be read as descriptive of this particular partition rather than as evidence of a general performance ranking. The constraint originates in the number of participants rather than in the analysis: with 34 individuals, no partitioning scheme yields an evaluation set large enough to resolve small performance differences. Studies seeking to establish classifier rankings in this domain require participant numbers an order of magnitude greater.

3.6.3. Distinction between demonstrated and inferred results

Two classes of result are reported here and should not be conflated. The limiting behaviour of Eq. (10), its closed form in the symmetric case, the sign-change condition, and the exponent selection rule are demonstrated: they follow from the algebra and are verified numerically in Section 3.2.7 without reference to any dataset. By contrast, the size of the evaluation partition, and therefore the confidence intervals and the classifier-comparison argument that depend on it, are inferred from the reported percentages and could not be confirmed against the archived records. The inference is consistent with the reported values but is not established by them, and any conclusion resting on it is correspondingly provisional.

3.6.4. Design and sampling constraints

The cohort comprises 34 undergraduate participants recruited at a single institution, and the findings should not be generalised to other age groups, occupational populations, or clinical samples. Recruitment materials referred explicitly to stress, so the sample is unlikely to be representative of the undergraduate population as a whole. Two of the six proposed stress levels are unpopulated, so the scheme is four-level in practice and the classifier cannot learn the absent categories. Conditions were presented in a fixed order without counterbalancing, so order effects cannot be separated from stimulus effects. Data were collected in a single session, so test-retest reliability of the stress categorisation is unknown. Where the partitioning unit for cross validation was the epoch rather than the participant, the reported accuracies estimate within subject rather than cross subject performance, and no claim of generalisation to unseen participants is made anywhere in this work.

3.6.5. Reproducibility

The limitations above are largely consequences of one circumstance: analytical outputs were preserved while the data and processing parameters that generated them were not. The analysis presented here was possible only because the summary statistics were reported with sufficient numerical precision to permit the inferences in Sections 3.2 and 3.3. Deposition of raw signals, preprocessing parameters, and intermediate feature matrices in a persistent repository at the time of the original work would have allowed the open questions identified in Section 3.2.5 to be settled directly. This is recorded not as a criticism of the original investigation, which met the standards prevailing at the time, but as a concrete illustration of the cost that deferred data deposition imposes on later reanalysis.

4. Conclusions

This study re-examined the Advanced Matthews Correlation Coefficient as applied to EEG-based mental stress classification, using the archived analytical outputs of an earlier doctoral investigation. Two objectives were pursued. The first was to characterise the behaviour of the generalised statistic as its power exponent increases. The analysis establishes that the statistic converges to one of three fixed values, plus one, minus one, or zero, determined solely by whether the arousal-to-valence ratios in the false-positive and false-negative partitions exceed unity in magnitude. In each case the true-positive and true-negative counts are eliminated in the limit, so the statistic ceases to depend on classifier performance. The rate at which this occurs is governed by the size of the confusion matrix as well as by the ratios themselves, with the sign changing when the exponentiated ratio equals the geometric mean of the true-positive and true-negative counts. A selection rule for the exponent follows from this relation and is given as Eq. (15). The six-sigma reasoning originally invoked to justify a fixed exponent of six has no bearing on the required value, which scales with dataset size and cannot be fixed a priori. The second objective was to establish the conditions under which the previously reported accuracies may be generalised. The reported percentages are consistent with an evaluation partition of approximately thirty-one instances, which suggests that the frequently cited figure of 85,680 denotes extracted features rather than independent observations. On this partition, the 95% confidence interval surrounding an accuracy of 96.8% spans more than fifteen percentage points, and the difference between the two leading classifiers corresponds to a single test instance. No ranking between them is supported. Separately, the compression of the reported MCCA values into a band of approximately one percentage point across three distinct experimental conditions, together with the concentration of the apparent gain in the index that had furthest to rise, is consistent with metric saturation rather than with improved discrimination. A refinement of the three-class ISMA instrument into six ordinal stress levels is proposed and applied descriptively. A multimodal fusion formulation admitting continuous weighting over a normalised simplex is specified in Section 2.5 as a methodological proposal; it has not been calibrated, and no performance claim is made for it.

The analysis also clarifies how the EEG features and the frequency band ratios enter the two statistics differently, which answers a question the original formulation left implicit. Conventional MCC is computed entirely from classification outcomes and is therefore independent of the band-power features except through the classifier that consumes them. MCCA admits the band ratios directly: the arousal index formed from $(\alpha + \beta) / (\delta + \theta)$ and the valence index formed from the left-right difference of that ratio occupy two of the four terms of the coefficient. As the exponent increases, those ratios come to determine the value of the statistic entirely, and the classifier outcome is eliminated, so it is the magnitude of the EEG band ratios relative to unity, rather than the separability of the stress classes, that governs MCCA at the exponent used. The choice of band boundaries and of the channels entering the left-right difference therefore propagates into the reported figure, and both must be stated whenever MCCA is reported. The effect of the validation design is comparable in magnitude and opposite in character. The 10-fold and K-fold conditions differ by 11.8 percentage points for Random Forest on identical features and an identical classifier, which is larger than any difference between the three classifiers under either scheme. Selection of the cross-validation design therefore affects the

reported accuracy more than selection of the learner does, and the lower of the two figures is the appropriate one to cite whenever generalisation beyond the recorded participants is at issue. Three directions follow for future work. The regime occupied by EEG-derived arousal-valence ratios should be established empirically, since it determines whether the exponent selection rule proposed here is binding in practice. The fusion formulation requires calibration on a dataset with participant-level partitioning and a substantially larger cohort. And the practice of modifying established classification metrics warrants systematic analytical scrutiny of the kind applied here, since the outward form of a coefficient does not guarantee the properties that motivated its adoption.

Acknowledgements

The authors would like to acknowledge the Faculty of Electrical and Electronics Engineering Technology, Universiti Malaysia Pahang Al-Sultan Abdullah, for the facilities, academic support, and research environment provided throughout this study.

Funding

This research received no external funding.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRedit Authorship Contribution Statement

M. Mustafa (Conceptualisation; Methodology; Formal analysis; Writing - original draft; Writing - review and editing; Supervision)

Z. L. Zahari (Investigation; Data curation; Methodology; Writing - review and editing)

L. N. Muhammad (Formal analysis; Validation; Writing - review and editing)

F. Naim (Methodology; Validation; Writing - review and editing)

R. Abdubrani (Software; Visualisation; Writing - review and editing)

Availability of the Data and Materials

The data used to support the findings of this study are included within the article.

Ethical Declaration

The study protocols, procedures, and consent forms were approved by the IIUM Research Ethics Committee (IREC) with the ID number IREC 2021-039.

Generative Artificial Intelligence Declarations

During preparation of this manuscript, generative AI tools were used as follows. ChatGPT (OpenAI, GPT-5.6) was used for language editing and structural reorganisation. No experimental data were generated by any AI tool.

REFERENCES

- [1] S. Acharya, A. Khosravi, D. Creighton, R. Alizadehsani, and U. R. Acharya, "Neurostressology: A systematic review of EEG-based automated mental stress perspectives," *Inf. Fusion*, vol. 124, p. 103368, Dec. 2025, doi: 10.1016/j.inffus.2025.103368.
- [2] A. Apicella, P. Arpaia, M. Frosolone, G. Masi, and E. Pasero, "Wearable electroencephalography and multi-modal mental state classification: A systematic literature review," *Comput. Biol. Med.*, vol. 150, p. 106134, Nov. 2022, doi: 10.1016/j.combiomed.2022.106134.
- [3] World Health Organization, *Mental Health at Work: Policy Brief*, Geneva, Switzerland: World Health Organization, 2022.
- [4] A. M. Algarni, M. A. Aslam, M. U. Khan, A. R. Baig, and S. A. M. Al-Sarem, "Enhanced classification of video-evoked stress response using power spectral density features," *Appl. Sci.*, vol. 14, no. 20, p. 9527, Oct. 2024, doi: 10.3390/app14209527.
- [5] F. Pinto, M. Cunha, and J. P. S. Cunha, "Stress monitoring using low-cost electroencephalogram devices: A systematic literature review," Mar. 2024, *arXiv:2403.05577*, doi: 10.48550/arXiv.2403.05577.
- [6] M. A. Al-Fahoum and A. A. Al-Fraihat, "Trends in EEG signal feature extraction applications," *Front. Artif. Intell.*, vol. 5, p. 1072801, Jan. 2023, doi: 10.3389/frai.2022.1072801.
- [7] L. Zapata-Salazar, C. A. Ferrer-Riesgo, and A. A. Torres-García, "Evaluation of EEG signals by spectral peak methods and statistical correlation for mental state discrimination," *Sensors*, vol. 24, no. 11, p. 3316, May 2024, doi: 10.3390/s24113316.
- [8] A. Khan, M. A. Shah, and A. M. Khattak, "EEG-based emotion recognition: A systematic review of traditional and deep learning methods," *Premier Sci.*, vol. 2025, p. 1257, 2025, doi: 10.52700/pjs.v25.1257.
- [9] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, p. 6, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

- [10] S. Agarwal, A. Sharma, R. Kumar, and P. Singh, "Machine learning evaluation of imbalanced health data: A comparative analysis of balanced accuracy, MCC, and F1 score," in *Proc. Int. Conf. Mach. Learn. Med. Imaging*, vol. 15239, 2024, pp. 167–181, doi: 10.1007/978-3-031-72215-8_12.
- [11] Y. Li, X. Chen, Z. Zhang, and L. Wang, "Real-time music emotion recognition based on multimodal fusion," *Alexandria Eng. J.*, vol. 113, pp. 448–458, Dec. 2024, doi: 10.1016/j.aej.2024.11.082.
- [12] S. Li, R. Wang, and J. Zhang, "A novel signal channel attention network for multi-modal emotion recognition," *Front. Neurobot.*, vol. 18, p. 1442080, Aug. 2024, doi: 10.3389/fnbot.2024.1442080.
- [13] Z. L. Zahari, "An enhanced stress indices in signal processing based on advanced Matthew correlation coefficient (MCCA) and multimodal function using EEG signal," Ph.D. dissertation, Faculty of Electr. Electron. Eng. Technol., Univ. Malaysia Pahang, Pekan, Malaysia, 2023.
- [14] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. 15, no. 2, pp. 70–73, Jun. 1967, doi: 10.1109/TAU.1967.1161901.
- [15] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980, doi: 10.1037/h0077714.
- [16] R. J. Davidson, "Anterior cerebral asymmetry and the nature of emotion," *Brain Cogn.*, vol. 20, no. 1, pp. 125–151, Sep. 1992, doi: 10.1016/0278-2626(92)90065-T.
- [17] P. Stoica and P. Babu, "Pearson–Matthews correlation coefficients for binary and multinary classification," *Signal Process.*, vol. 222, p. 109511, Sep. 2024, doi: 10.1016/j.sigpro.2024.109511.
- [18] K. Osman, M. A. Mosleh, and A. H. Ali, "Applying the six-sigma methodology to determine the limits of quality control tests for a specific linear accelerator," *J. Appl. Clin. Med. Phys.*, vol. 25, no. 10, p. e14492, Oct. 2024, doi: 10.1002/acm2.14492.
- [19] G. Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," *NeuroImage*, vol. 180, pp. 68–77, Oct. 2018, doi: 10.1016/j.neuroimage.2017.06.061.
- [20] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLoS ONE*, vol. 14, no. 11, p. e0224365, Nov. 2019, doi: 10.1371/journal.pone.0224365.
- [21] R. J. Barry et al., "k-Fold cross-validation can significantly over-estimate true classification accuracy in common EEG-based passive BCI experimental designs," *Sensors*, vol. 23, no. 13, p. 6077, Jul. 2023, doi: 10.3390/s23136077.
- [22] S. Kumar, R. Sharma, and A. Verma, "Optimized EEG-based stress detection: A novel approach," *Biomed. Pharmacol. J.*, vol. 17, no. 4, pp. 2307–2318, Dec. 2024, doi: 10.13005/bpj/2989.
- [23] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Min.*, vol. 14, p. 13, Feb. 2021, doi: 10.1186/s13040-021-00244-z.
- [24] L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Stat. Sci.*, vol. 16, no. 2, pp. 101–133, May 2001, doi: 10.1214/ss/1009213286.
- [25] X. Li, W. Zhang, and Y. Liu, "An automated extraction of spectral-temporal and spatial-temporal features of EEG for emotion detection," *Brain Inform.*, vol. 12, p. 265, Jan. 2025, doi: 10.1186/s40708-025-00265-y.
- [26] M. A. Khan, A. R. Baig, and S. A. M. Al-Sarem, "A novel optimized hybrid deep learning framework for mental stress detection using electroencephalography," *Brain Sci.*, vol. 15, no. 8, p. 835, Aug. 2025, doi: 10.3390/brainsci15080835.
- [27] H. Zhang, J. Zhu, and Q. Zhao, "A hybrid EEG-based stress state classification model using multi-domain transfer entropy and PCANet," *Brain Sci.*, vol. 14, no. 6, p. 595, Jun. 2024, doi: 10.3390/brainsci14060595.
- [28] M. T. Bilal, S. A. Khan, and M. U. Rehman, "Optimal variational mode decomposition based automatic stress classification system using EEG signals," *Appl. Acoust.*, vol. 228, p. 110297, Dec. 2024, doi: 10.1016/j.apacoust.2024.110297.
- [29] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
- [30] W. Chen, X. Li, H. Liu, and Y. Zhao, "Emotion recognition based on multimodal physiological electrical signals," *Front. Neurosci.*, vol. 19, p. 1512799, Jan. 2025, doi: 10.3389/fnins.2025.1512799.
- [31] Y. Wang, H. Li, X. Chen, and Z. Zhou, "Learner's cognitive state recognition based on multimodal physiological signal fusion," *Appl. Intell.*, vol. 55, p. 5958, Dec. 2024, doi: 10.1007/s10489-024-05958-1.
- [32] A. Ahmad, T. Khan, and S. Hussain, "Enhancing EEG-based emotion detection with hybrid models: Insights from DEAP dataset applications," *Sensors*, vol. 25, no. 6, p. 1827, Mar. 2025, doi: 10.3390/s25061827.
- [33] T. Li, Y. Chen, Z. Wang, and X. Sun, "Leveraging deep learning for robust EEG analysis in mental health monitoring," *Front. Neuroinform.*, vol. 18, p. 1494970, Jan. 2025, doi: 10.3389/fninf.2024.1494970.