e-ISSN: 2987-1530

Sriwijaya Journal of Forensic and Medicolegal (SJFM)

Journal website: https://phlox.or.id/index.php/sjfm

Decoding Deception: Advanced fMRI and Machine Learning Techniques for Detecting Malingered Psychiatric Symptoms in Forensic Evaluations in Indonesia

Taufiq Indera Jayadi[1], Taryudi Suharyana[2*], Vita Amanda[3], Brenda Jaleel[4]

[1]Department of Radiology, Phlox Institute, Palembang, Indonesia

[2]Department of Neurology, CMHC Research Center, Palembang, Indonesia

[3]Department of Psychiatry, CMHC Research Center, Palembang, Indonesia

[4]Department of Neuroscience, San Fernando General Hospital, San Fernando, Trinidad and Tobago

ARTICLE INFO

Keywords:
Deception detection
fMRI
Forensic psychiatry
Malingering
Visum et repertum psychiatricum

***Corresponding author:**
Taryudi Suharyana

E-mail address:
taryudi.suharyana@cattleyacenter.id

All authors have reviewed and approved the final version of the manuscript.

https://doi.org/10.59345/sjfm.v2i2.200

A B S T R A C T

Introduction: Detecting malingered psychiatric symptoms presents a significant challenge in Indonesian forensic evaluations, potentially impacting justice and resource allocation. Current methods rely heavily on clinical judgment and psychometric testing, lacking objective biomarkers. This study explored the potential of combining functional magnetic resonance imaging (fMRI) with machine learning (ML) to identify neural patterns differentiating malingered from genuine psychiatric symptoms in an Indonesian forensic context. **Methods:** This case-control study included 90 Indonesian male participants referred for forensic psychiatric evaluation (visum et repertum psychiatricum): 30 diagnosed genuine psychiatric patients (schizophrenia/psychotic depression), 30 individuals identified as malingerers, and 30 healthy controls. Participants underwent clinical assessment, psychometric testing (including symptom validity tests - SVTs), and an fMRI scan using a symptom-endorsement paradigm designed to probe cognitive control and deception-related neural activity. Preprocessed fMRI data were analyzed using group-level GLM and machine learning (Support Vector Machine - SVM; Random Forest - RF) classifiers trained on extracted features (ROI activation, functional connectivity) to distinguish malingerers. Performance was evaluated using k-fold cross-validation. **Results:** fMRI results indicated significantly greater activation in the malingering group compared to genuine patients and controls in prefrontal (dlPFC, vlPFC) and anterior cingulate cortex (ACC) regions during feigned symptom endorsement ($p<0.001$, FWE-corrected). An SVM classifier using combined ROI activation and functional connectivity features achieved the highest accuracy (83%), sensitivity (80%), specificity (86%), and AUC (0.88) in distinguishing malingerers from genuine patients. **Conclusion:** These findings suggest that integrating fMRI and ML techniques holds promise as an objective, supplementary tool for detecting malingered psychiatric symptoms within Indonesian forensic evaluations. While promising, the moderate accuracy highlights the need for further validation, consideration of ethical implications, and adaptation to the Indonesian context before any potential clinical application.

1. Introduction

Malingering, characterized by the intentional production of false or grossly exaggerated physical or psychological symptoms driven by external incentives, poses a persistent and intricate challenge in forensic psychiatric evaluations globally, including within Indonesia. The forensic setting often involves substantial external motivations, encompassing the avoidance of criminal prosecution or harsher penalties (such as invoking criminal non-responsibility under

Pasal 44 of the Indonesian Criminal Code), the pursuit of compensation, or the desire for specific institutional placements. Consequently, the precise identification of malingered psychiatric symptoms is of paramount importance for upholding the integrity of the legal process, ensuring equitable justice for victims and society, facilitating appropriate sentencing or treatment plans for offenders, and enabling the efficient allocation of limited mental health resources. The failure to accurately detect malingering can precipitate miscarriages of justice, wherein genuinely responsible individuals evade accountability, or resources are misdirected away from individuals with genuine illness. Conversely, the erroneous attribution of malingering to a genuinely ill individual can result in the denial of essential treatment and unjust punishment. These consequences underscore the gravity of forensic evaluations and the ethical necessity for precise assessment. Traditionally, the detection of malingering in forensic psychiatry has employed a multi-faceted approach, integrating data from diverse sources, including clinical interviews, psychological testing, and collateral information. This approach involves meticulous history taking, mental status examinations, observation of inconsistencies in presentation, and the evaluation of congruence between reported symptoms and observed behavior. However, clinical judgment, when used in isolation, has demonstrated limitations in reliability and susceptibility to bias. Standardized instruments such as the Minnesota Multiphasic Personality Inventory (MMPI-2-RF) incorporate validity scales designed to identify unusual response patterns. Furthermore, specific instruments like Symptom Validity Tests (SVTs), such as the Test of Memory Malingering (TOMM) or the Structured Inventory of Reported Symptoms (SIRS-2), directly assess the likelihood of feigning. While these instruments are valuable, their effectiveness can be compromised by the examinee's level of sophistication, coaching, and the specific symptoms being feigned. In addition, the availability of culturally adapted and validated versions of these instruments within Indonesia remains a factor to be considered. Collateral information, including the review of medical records, police reports, witness statements, and other historical data, is included to verify consistency. Despite the use of this comprehensive approach, the detection of sophisticated malingerers, particularly those who feign subtle or complex psychiatric symptoms like psychosis, remains challenging. The inherent subjectivity of psychiatric symptoms and the reliance on self-report create vulnerabilities in the assessment process. Consequently, there is a recognized need for more objective markers to augment existing methodologies.[1-4]

Advances in neuroimaging techniques, especially functional magnetic resonance imaging (fMRI), present a potential avenue for the development of objective biomarkers of cognitive states, including deception. fMRI indirectly measures brain activity by detecting changes associated with blood flow, known as the Blood Oxygen Level-Dependent (BOLD) signal. The fundamental principle underlying its application in deception detection is the assumption that lying or feigning symptoms is generally more cognitively demanding than truth-telling or reporting genuine experiences. Neurocognitive models of deception propose that feigning involves several cognitive processes not typically engaged during truthful reporting. These processes include the inhibition of truthful responses, the construction and maintenance of fabricated narratives or symptoms in working memory, monitoring one's own behavior and the interviewer's reactions for credibility, and task switching between truth and falsehood. These cognitive operations are associated with increased activation in specific brain networks, particularly those involved in cognitive control, executive function, and attention. These networks are primarily located within the prefrontal cortex (PFC), including the dorsolateral PFC (dlPFC) and ventrolateral PFC (vlPFC), the anterior cingulate cortex (ACC), the parietal cortex, and the insula. Numerous fMRI studies, largely conducted in laboratory settings using instructed deception paradigms such as mock crime scenarios or concealing identity, have reported differential activation patterns in these regions between deceptive and truthful responses. While the application of these findings to the complex scenario

of malingering psychiatric symptoms requires careful consideration, the fundamental principle of differential neural activity related to feigning remains promising.[5-7]

Traditional group-level fMRI analyses, which utilize the General Linear Model (GLM), are effective in identifying average differences in brain activation between groups. However, they are often suboptimal for making predictions at the individual level, which is essential for detecting malingering, as this requires classifying an individual examinee as either likely feigning or likely genuine. Machine learning (ML) techniques offer powerful computational tools suited for identifying intricate, multivariate patterns within high-dimensional data, such as fMRI scans, and for making predictions at the individual level. ML algorithms can be trained using fMRI data, including activation levels in specific regions and patterns of connectivity between regions, from known groups of confirmed malingerers and genuine patients. Through this training, the algorithms learn the neural signatures that best differentiate these groups. Subsequently, these trained models can classify new, unseen individuals based on their fMRI patterns. Various ML algorithms, including Support Vector Machines (SVM), Random Forests (RF), and neural networks, have demonstrated potential in classifying psychiatric conditions and, more recently, in detecting deception based on fMRI data, often achieving higher accuracies than traditional univariate analyses. In Indonesia, forensic psychiatric evaluations, known as visum et repertum psychiatricum, are crucial in legal proceedings, particularly concerning Pasal 44 of the Indonesian Criminal Code, which addresses criminal responsibility in cases involving mental disorder or defect. The assessment of potential malingering is a frequent and essential component of these evaluations, primarily conducted by forensic psychiatrists using clinical interviews, observations, and available psychometric tools. The introduction of objective biomarkers has the potential to enhance the reliability and validity of these assessments within the Indonesian legal framework. However, research employing advanced neuroimaging techniques like fMRI combined with ML for forensic psychiatric purposes, specifically addressing malingering, is virtually non-existent in Indonesia. This gap is significant considering the potential benefits and the unique socio-cultural and legal context of Indonesia. Challenges include the limited availability and high cost of fMRI technology, the need for specialized expertise in neuroimaging analysis and ML, and the necessity for culturally sensitive research paradigms and ethical frameworks.[8-10] Therefore, this study aimed to investigate the feasibility and potential utility of combining fMRI and ML techniques for detecting malingered psychiatric symptoms, specifically psychosis, within the Indonesian forensic evaluation context. This study serves as a foundational exploration to stimulate further empirical research and discussion on the use of neurotechnologies in Indonesian forensic psychiatry.

2. Methods

This investigation employed a case-control study design. Three distinct groups were recruited and compared: individuals identified as malingering psychiatric symptoms (Malingering Group), individuals with diagnosed genuine psychiatric disorders (Genuine Patient Group), and healthy individuals (Healthy Control Group). The study involved clinical assessment, psychometric testing, and fMRI scanning during a symptom-endorsement task, followed by traditional fMRI analysis and ML-based classification.

The study was conducted at a tertiary referral hospital and research center in Jakarta, Bandung and Surabaya, Indonesia, equipped with advanced neuroimaging facilities including a 3 Tesla fMRI scanner. Recruitment occurred through referrals from the hospital's forensic psychiatric unit, which conducts court-ordered visum et repertum psychiatricum evaluations, and through community advertisements for healthy controls.

A total of 90 male Indonesian participants aged 18-50 years were included in the study, divided into three groups of 30. Male participants were selected for this initial investigation to homogenize the sample, acknowledging the need for future studies to include females. Inclusion criteria for all groups were; male,

aged 18-50 years; fluent in Bahasa Indonesia; estimated IQ within the normal range (>80), assessed via Raven's Progressive Matrices; ability to provide informed consent; and right-handedness. Exclusion criteria for all groups were; history of significant neurological disorders (epilepsy, stroke, traumatic brain injury with loss of consciousness > 30 mins); current DSM-5/PPDGJ-III diagnosis of substance use disorder (moderate-severe) within the past 6 months (except nicotine); contraindications to MRI scanning (metallic implants, claustrophobia); and current use of medications known to significantly affect BOLD signal (benzodiazepines required washout prior to scan). In the Malingering Group (MAL, n=30). Participants referred for forensic evaluation who met adapted criteria for probable malingering of psychosis based on operationalized for the Indonesian context. These criteria included; (a) presence of a forensic context; (b) clear external incentive; (c) significant discrepancies between reported symptoms and objective findings (observed behavior, collateral reports, cognitive testing); and (d) evidence from psychological testing indicative of feigning (scores above cut-offs on Indonesian adaptations of MMPI-2-RF validity scales or SIRS-2, failure on SVTs like TOMM). These participants were identified through the comprehensive forensic evaluation process.

In the Genuine Patient Group (PAT, n=30): Participants referred for forensic evaluation with a confirmed primary diagnosis of Schizophrenia or Major Depressive Disorder with Psychotic Features according to DSM-5 criteria (cross-referenced with PPDGJ-III), established via structured clinical interview (SCID-5), review of longitudinal clinical history, and consensus diagnosis by experienced forensic psychiatrists involved in their evaluation. They were instructed to respond truthfully during assessments and the fMRI task. Symptom severity was assessed using the Positive and Negative Syndrome Scale (PANSS). In the Healthy Control Group (HC, n=30): Participants recruited from the community, screened using the MINI International Neuropsychiatric Interview to exclude current or past major psychiatric or neurological disorders. They were group-matched to the MAL and PAT groups on age and education level. They were instructed to respond truthfully during assessments and the fMRI task.

The study protocol received ethical review and approval from the Health Research Ethics Committee of the Phlox Institute, Palembang, Indonesia. Key ethical procedures were followed: obtaining voluntary, written informed consent from all participants after providing a full explanation of the study procedures, risks, benefits, confidentiality measures, and the right to withdraw at any time without prejudice. Consent forms and information sheets were provided in clear, understandable Bahasa Indonesia; careful assessment of the capacity to consent in the PAT group was conducted to ensure understanding and voluntariness; participants were explicitly informed that participation (or non-participation) would not influence the outcome of any ongoing legal or clinical processes; data anonymization and secure storage procedures were implemented to protect participant confidentiality. fMRI data were de-identified, linked only via coded identifiers stored separately and securely; the exploratory nature of the research was emphasized, and participants were informed that the fMRI/ML results would not be used for actual clinical or legal decision-making.

All participants underwent a battery of assessments, a sociodemographic and clinical history interview, a structured clinical interview for DSM-5 disorders (SCID-5), assessment of symptom severity (PAT group), cognitive screening, personality/psychopathology assessment, symptom validity tests (SVTs), and a forensic file review (MAL and PAT groups). The sociodemographic and clinical history interview involved standardized collection of demographic data, educational/occupational history, family history, medical/psychiatric history, substance use history, and forensic history. The Structured Clinical Interview for DSM-5 Disorders (SCID-5) was administered by trained clinical researchers to confirm diagnoses in the PAT group and exclude psychopathology in the HC group. Relevant modules were used for MAL group screening. PPDGJ-III criteria were cross-referenced. Assessment of symptom severity in the PAT group was conducted using the Positive and Negative Syndrome Scale (PANSS).

Cognitive screening was performed using Raven's Progressive Matrices. Personality/psychopathology assessment used the Indonesian adaptation of the MMPI-2-RF (focus on validity scales: F, Fp, Fs, FBS, RBS, L, K). Symptom validity tests (SVTs) included the Test of Memory Malingering (TOMM) and the Indonesian adaptation of the Structured Inventory of Reported Symptoms, 2nd Edition (SIRS-2). The forensic file review (MAL and PAT groups) involved examination of police reports, previous evaluations, witness statements, and court documents to gather objective data and assess consistency of presentation. fMRI data were acquired on a 3 Tesla Siemens Magnetom Skyra MRI scanner equipped with a standard head coil. A block-design fMRI paradigm, adapted from deception/symptom validity literature, was employed as the Symptom Endorsement Task. Stimuli consisted of short sentences in Bahasa Indonesia describing symptoms, categorized as: plausible psychotic symptoms (PPS), derived from PANSS/DSM-5 criteria (e.g., "Saya mendengar suara-suara yang tidak didengar orang lain"; "Saya merasa pikiran saya dikendalikan"); absurd/atypical symptoms (AAS), symptoms rarely reported by genuine patients (e.g., "Kepala saya bisa berputar 360 derajat"; "Saya bisa berbicara dengan hewan peliharaan saya"); and neutral symptoms (NS), common, non-psychiatric states (e.g., "Kadang saya merasa lelah"; "Saya suka makan nasi goreng"). Symptoms were presented in blocks (6 sentences per block, 4 seconds per sentence) interspersed with fixation cross baseline blocks (15 seconds). Blocks contained either PPS, AAS, or NS. Participants responded via button press ("Ya" / Yes or "Tidak" / No) indicating whether the described symptom applied to them recently. All participants (MAL, PAT, HC) were instructed to respond according to their actual experiences or beliefs regarding each symptom presented. The nature of responses (truthful endorsement, feigned endorsement, truthful rejection) was determined post-hoc based on group classification and item type for analysis. Functional images were acquired using a T2*-weighted gradient-echo echo-planar imaging (EPI) sequence with the following parameters: Repetition Time (TR) = 2000 ms; Echo Time (TE) = 30 ms; Flip Angle = 90°; Field of View (FOV) = 192x192 mm; Matrix = 64x64; Voxel Size = 3x3x3 mm; 36 axial slices. Structural images were acquired using a high-resolution T1-weighted Magnetization Prepared Rapid Gradient Echo (MPRAGE) sequence with the following parameters: TR = 2300 ms; TE = 2.98 ms; TI = 900 ms; Flip Angle = 9°; FOV = 256x256 mm; Voxel Size = 1x1x1 mm; 176 sagittal slices.

fMRI data preprocessing was performed using SPM12 (Statistical Parametric Mapping, Wellcome Trust Centre for Neuroimaging, London, UK) running on MATLAB (The MathWorks, Inc., Natick, MA, USA). The standard preprocessing steps included realignment, slice-timing correction, coregistration, segmentation, normalization to MNI space (resampled to 2x2x2 mm voxels), and smoothing (8 mm FWHM Gaussian kernel). Participants with head motion exceeding 3 mm translation or 3° rotation were excluded from the analysis. For each participant, a first-level statistical analysis was performed using the General Linear Model (GLM) in SPM12. The model included regressors representing the onset and duration of each task condition (PPS endorsement, AAS endorsement, NS endorsement – modeled separately for 'Yes' and 'No' responses). Contrasts of interest were defined to identify brain activation associated with: feigned symptom endorsement (derived from MAL group responses to PPS/AAS relative to baseline); genuine symptom endorsement (derived from PAT group responses to PPS relative to baseline); truthful rejection of absurd symptoms (derived from PAT & HC group 'No' responses to AAS relative to baseline); and general symptom processing vs. baseline. Motion parameters derived from realignment were included as nuisance regressors. A high-pass filter (128s cutoff) was applied. Contrast images representing activation magnitude were generated for each participant for group-level and ML analyses. For the machine learning analysis, two types of features were extracted from the preprocessed fMRI data for each participant, focusing on the contrast representing symptom endorsement: ROI-based mean activation and functional connectivity. ROI-based mean activation involved extracting mean beta values from predefined Regions of Interest (ROIs) involved in

cognitive control (dlPFC, vlPFC, ACC, anterior insula, inferior parietal lobule), defined using the Automated Anatomical Labeling (AAL) atlas. Functional connectivity analysis was performed using a seed-based approach. Time series were extracted from seed regions (key nodes within PFC/ACC). Correlation coefficients between seed time series and time series from other ROIs were calculated, Fisher-z transformed, and used as features. To identify the most informative features and prevent overfitting, Recursive Feature Elimination with Cross-Validation (RFECV) was employed. Two supervised learning algorithms were implemented using the scikit-learn library in Python: Support Vector Machine (SVM), tested with linear and Radial Basis Function (RBF) kernels; and Random Forest (RF), an ensemble method based on decision trees. A stratified 10-fold cross-validation approach was used to train and evaluate the classifiers. In each fold, data were split into training (including feature selection) and testing sets. The primary classification task was discriminating between MAL vs. PAT. Secondary analyses included MAL vs. HC and MAL vs. (PAT + HC combined). Model performance was evaluated using standard metrics: accuracy, sensitivity (recall), specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the Receiver Operating Characteristic Curve (AUC). Demographic and clinical data were compared between groups using independent samples t-tests or Chi-square tests. For fMRI group-level analysis, one-way ANOVA was performed on extracted ROI beta values. Whole-brain analyses used standard parametric methods within SPM12 (voxel threshold $p < 0.001$ uncorrected, cluster threshold $p < 0.05$ FWE-corrected). ML model performance metrics were compared to identify the best-performing approach. The significance level was set at $\alpha = 0.05$ (two-tailed).

3. Results

Table 1 presents the brain regions with significantly greater activation in the malingering group compared to the genuine patient group during the symptom endorsement task ($p < 0.05$, FWE cluster-corrected); Dorsolateral Prefrontal Cortex (dlPFC): Significant bilateral activation was observed (Right: MNI coordinates 44, 32, 38; Left: MNI coordinates -40, 30, 40). The dlPFC is crucial for executive functions like working memory and response inhibition, suggesting increased cognitive effort in malingerers; Ventrolateral Prefrontal Cortex (vlPFC): Bilateral activation was also found in the vlPFC (Right: MNI coordinates 50, 28, 10; Left: MNI coordinates -48, 30, 8). The vlPFC is involved in cognitive control and selection, potentially reflecting the active construction of false symptoms; Dorsal Anterior Cingulate Cortex (dACC): Midline activation was seen in the dACC (MNI coordinates 4, 22, 42). The dACC plays a key role in conflict monitoring and error detection, indicating heightened monitoring during feigning; Anterior Insula: Bilateral activation was present (Right: MNI coordinates 36, 18, 4; Left: MNI coordinates -34, 20, 2). The insula is implicated in interoceptive awareness and emotional regulation, possibly related to the self-monitoring and emotional aspects of deception; Inferior Parietal Lobule (IPL): Bilateral activation was observed (Right: MNI coordinates 48, -50, 46; Left: MNI coordinates -46, -52, 48). The IPL is involved in attention and working memory, suggesting increased cognitive load during the task.

Table 2 presents the performance metrics of various machine learning classifiers in distinguishing individuals identified as malingerers (MAL) from genuine patients (PAT) using fMRI data, evaluated through 10-fold cross-validation. The table compares Support Vector Machine (SVM) models with different kernels (Linear and Radial Basis Function - RBF) and a Random Forest (RF) model, each using different feature sets derived from the fMRI data: ROI Activation, Functional Connectivity, and a combination of both (Combined Features); SVM (Linear Kernel, ROI Activation): This model achieved an accuracy of 76.7%, with a sensitivity of 73.3%, specificity of 80.0%, and an AUC of 0.81. This suggests moderate performance in distinguishing the two groups, with a slightly better ability to correctly identify genuine patients than malingerers; SVM (RBF Kernel, ROI Activation): The SVM with the RBF kernel showed a slight improvement over the linear kernel, with an accuracy of 78.3%, sensitivity of 76.7%,

specificity of 80.0%, and an AUC of 0.82. The RBF kernel allows for non-linear decision boundaries, potentially capturing more complex patterns in the data; SVM (RBF Kernel, Functional Connectivity): Using only functional connectivity features, this model had an accuracy of 75.0%, sensitivity of 70.0%, specificity of 80.0%, and an AUC of 0.79. This indicates that functional connectivity alone was less effective than ROI activation for classification in this context; SVM (RBF Kernel, Combined Features): The best-performing model was the SVM with the RBF kernel using the combined feature set. It achieved an accuracy of 83.3%, sensitivity of 80.0%, specificity of 86.7%, PPV of 85.7%, NPV of 81.3%, and an AUC of 0.88. This demonstrates that integrating information about both regional brain activity levels and the interactions between brain regions improves classification accuracy; Random Forest (RF, Combined Features): The Random Forest model also performed well with the combined feature set, achieving an accuracy of 81.7% and an AUC of 0.85. Random Forest is an ensemble method that can handle complex data, but in this case, it was slightly outperformed by the SVM with the RBF kernel.

Table 1. Brain regions showing significantly greater activation in malingerers vs. genuine patients during symptom endorsement ($p < 0.05$, FWE Cluster-Corrected).

Brain region	Hemisphere	MNI coordinates (x, y, z)	Cluster size (k)	Peak Z-score
Dorsolateral Prefrontal Cortex (dlPFC)	R	44, 32, 38	485	5.12
	L	-40, 30, 40	410	4.98
Ventrolateral Prefrontal Cortex (vlPFC)	R	50, 28, 10	370	4.85
	L	-48, 30, 8	345	4.77
Dorsal Anterior Cingulate Cortex (dACC)	Midline	4, 22, 42	550	5.31
Anterior Insula	R	36, 18, 4	290	4.65
	L	-34, 20, 2	275	4.58
Inferior Parietal Lobule (IPL)	R	48, -50, 46	315	4.72
	L	-46, -52, 48	300	4.69

Table 2. Performance of machine learning classifiers for distinguishing malingerers (MAL) from genuine patients (PAT) (10-Fold Cross-Validation).

Model	Feature set	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
SVM (Linear)	ROI Activation	76.7%	73.3%	80.0%	78.6%	75.0%	0.81
SVM (RBF)	ROI Activation	78.3%	76.7%	80.0%	79.3%	77.4%	0.82
SVM (RBF)	Func. Connectivity	75.0%	70.0%	80.0%	77.8%	72.7%	0.79
SVM (RBF)	**Combined Features**	**83.3%**	**80.0%**	**86.7%**	**85.7%**	**81.3%**	**0.88**
Random Forest (RF)	Combined Features	81.7%	76.7%	86.7%	85.2%	78.8%	0.85

4. Discussion

The observation of heightened activation within prefrontal and anterior cingulate regions in the malingering group aligns robustly with both existing neurocognitive theories of deception and the findings of a multitude of previous fMRI studies that have explored the neural underpinnings of deceptive behavior. The dorsolateral prefrontal cortex (dlPFC) and the ventrolateral prefrontal cortex (vlPFC) are recognized as critical hubs within the brain's executive function network. These regions play a pivotal role in a range of higher-order cognitive processes, including working memory, response inhibition, and the manipulation of information held in mind. These cognitive demands are substantially increased when an individual is actively constructing and maintaining

feigned symptoms of a psychiatric disorder while simultaneously suppressing the truthful responses that would accurately reflect their actual experiences. The need to generate a false narrative about one's mental state, keep that narrative consistent, and avoid inadvertently revealing the truth places a significant burden on working memory and executive control. Furthermore, the dorsal anterior cingulate cortex (dACC) is another key region within the cognitive control network, and it is specifically implicated in conflict monitoring, error detection, and signaling the necessity for increased cognitive control. The act of feigning inherently involves a high degree of cognitive conflict. The individual must manage the conflict between their knowledge of their true state and the need to portray a false state. They must also continuously monitor their own responses and behavior to ensure consistency and credibility, and detect any inconsistencies that might betray their deception. This constant monitoring and conflict resolution likely contribute to the increased dACC activation observed in the malingering group. In addition to the prefrontal and cingulate cortices, the insula also demonstrated increased activation in the malingering group. The insula is a brain region involved in interoceptive awareness, which is the sense of the internal state of one's body, and emotional regulation. In the context of malingering, increased insula activation might be related to a heightened awareness of one's own internal state as the individual attempts to control their emotional responses and behaviors to maintain the deceptive facade. The act of deception can be emotionally taxing, and the individual may need to exert significant effort to regulate their emotions and avoid displaying signs of anxiety or discomfort that could raise suspicion. Finally, the parietal lobe, including the inferior parietal lobule (IPL), also showed greater activation in the malingering group. The parietal lobe is involved in attention and working memory. The increased activation in this region may reflect the heightened demands placed on these cognitive functions when an individual is actively engaged in feigning psychiatric symptoms. Maintaining a fabricated account of symptoms, remembering what has been said, and attending to the interviewer's questions all require sustained attention and robust working memory. In essence, the observation that these cognitive control networks, encompassing the prefrontal cortex, anterior cingulate cortex, insula, and parietal lobe, exhibited significantly greater engagement in individuals identified as malingerers compared to those with genuine psychiatric conditions provides compelling neural evidence in support of the central hypothesis of the study. This hypothesis posits that the act of feigning psychiatric illness is not a passive process but rather a cognitively demanding endeavor that leaves a distinct and measurable neural signature. The fact that genuine patients did not exhibit a similar pattern of increased activation in these control regions, and in fact showed less activation in these regions than the malingerers, further bolsters this interpretation, suggesting that the observed neural activity is specifically associated with the cognitive processes involved in feigning rather than with the experience of genuine psychiatric symptoms.[11,12]

The successful application of machine learning (ML) classifiers to differentiate individuals identified as malingerers from those with genuine psychiatric disorders based on their fMRI patterns lends further credence to the potential utility of this approach. It suggests that fMRI data can provide information that goes beyond simply identifying group-level differences in brain activity and can be used to make predictions about individual cases. This is particularly important in forensic settings, where the focus is on determining the veracity of an individual's claims about their mental state. The finding that a Support Vector Machine (SVM) model achieved the best performance in this classification task, with an AUC of 0.88, is consistent with a growing body of prior research that has demonstrated the effectiveness of SVMs in high-dimensional neuroimaging classification tasks. SVMs are particularly well-suited for analyzing complex datasets like fMRI data because they can identify intricate patterns and relationships between brain activity and diagnostic categories. Their ability to handle high dimensionality and non-linear relationships makes them a powerful tool for this type

of analysis. Furthermore, the observation that combining features derived from both regional activation levels and functional connectivity patterns resulted in superior classification performance compared to using either feature type alone provides valuable insight into the neural mechanisms of malingering. Regional activation levels reflect the degree of activity within specific brain areas, while functional connectivity patterns describe the communication and coordination between different brain regions. The improved performance achieved by combining these two types of features suggests that malingering is characterized not only by changes in the activity of specific cognitive control regions but also by alterations in the way these regions communicate and interact with each other. This highlights the importance of considering the brain as a network, where different regions work together to accomplish complex cognitive tasks.[13,14]

While the accuracy achieved by the machine learning classifiers in distinguishing malingerers from genuine patients—83%, with an AUC of 0.88—is promising and significantly better than chance, it is crucial to emphasize that this technique is not perfectly accurate. It is essential to acknowledge the limitations of the current findings and the potential consequences of misclassification. The false positive rate, which represents the probability of incorrectly classifying a genuine patient as a malingerer, was approximately 13% (calculated as 100% minus the specificity of 86.7%). This means that in a clinical setting, approximately 13 out of 100 genuine patients might be erroneously identified as feigning their symptoms. The implications of such a misclassification could be severe, potentially leading to the denial of appropriate treatment, the imposition of unjust legal penalties, or damage to the individual's credibility. Similarly, the false negative rate, which represents the probability of failing to detect an actual malingerer, was approximately 20% (calculated as 100% minus the sensitivity of 80.0%). This means that in a clinical setting, approximately 20 out of 100 individuals who are truly feigning their symptoms might be missed by the fMRI-ML analysis. The consequences of a false negative can also be significant, potentially leading to inappropriate leniency in legal proceedings, the misuse of resources, or a failure to address the individual's underlying motivations. These error rates underscore the critical point that fMRI-ML approaches, even if they are further validated and refined for clinical use, should not be considered as stand-alone diagnostic tools. Instead, they should be viewed as adjunctive tools, providing additional information to supplement, but not replace, the comprehensive clinical and psychometric evaluation that remains the cornerstone of forensic psychiatric assessment. The interpretation of fMRI-ML results must always be integrated with other sources of information, including clinical interviews, behavioral observations, psychological testing, and collateral data.[15,16]

As anticipated, the accuracy achieved when distinguishing malingerers from healthy controls was considerably higher (AUC 0.95) than when distinguishing malingerers from genuine patients. This finding is not surprising, given that the neural patterns associated with feigning symptoms are likely to be more distinct from the neural patterns of individuals with no psychiatric conditions than from the neural patterns of individuals with genuine psychiatric disorders. Healthy controls, by definition, do not exhibit the cognitive or emotional processes associated with either genuine psychopathology or the deliberate attempt to feign such psychopathology. This clear distinction in neural activity patterns makes the classification task easier for the machine learning algorithms. While this higher accuracy is encouraging, the primary clinical and legal challenge lies in differentiating malingerers from genuine patients, as this distinction has the most significant implications for decision-making.[17,18]

The findings of this study carry potentially significant implications for the field of forensic psychiatric practice in Indonesia, particularly in the context of visum et repertum psychiatricum, the forensic psychiatric evaluations that are conducted for legal purposes. Currently, these evaluations rely heavily on established clinical methods, including clinical interviews, behavioral observations, and psychological testing. While these methods are

essential, they are also subjective and can be influenced by various factors, such as the clinician's experience, the patient's ability to communicate, and the potential for deception. The introduction of an objective tool, such as fMRI-ML, that could enhance the clinician's confidence in differentiating between genuine and feigned symptoms has the potential to significantly improve the reliability and validity of forensic psychiatric evaluations in Indonesia. This is particularly relevant in cases involving the determination of criminal responsibility under Pasal 44 of the Indonesian Criminal Code (KUHP) and assessments of fitness to stand trial, where the accuracy of the psychiatric evaluation has profound legal consequences. In such high-stakes situations, an objective measure of symptom validity could provide valuable additional information to support clinical judgment. It is crucial to emphasize that fMRI-ML should not be seen as a replacement for traditional clinical methods but rather as a valuable addition to the existing multi-method approach. In complex cases where clinical and psychometric data are ambiguous or conflicting, fMRI-ML could offer an additional source of evidence to help clarify the diagnostic picture. It could also be particularly useful in detecting sophisticated malingerers who are adept at concealing their deception from traditional assessment methods.[19,20]

5. Conclusion

This study provides compelling evidence that the combination of fMRI and machine learning techniques holds significant promise as a potential supplementary tool for the detection of malingered psychiatric symptoms within the challenging context of Indonesian forensic evaluations. The observed differences in neural activity, particularly the heightened activation in prefrontal and cingulate regions among individuals identified as malingerers, align with established neurocognitive models of deception. Furthermore, the successful application of machine learning classifiers, especially the Support Vector Machine, to distinguish malingerers from genuine patients underscores the potential for fMRI data to inform individual case assessments. However, it is crucial to interpret these findings with a degree of caution. The accuracy of the machine learning models, while promising, is not perfect, and the potential for misclassification errors necessitates a careful and nuanced approach. fMRI-ML should not be considered a standalone diagnostic tool but rather an adjunct to traditional clinical and psychometric evaluations. Future research should focus on further validating these findings in larger, more diverse samples, refining the machine learning models, and addressing the ethical and practical considerations of implementing fMRI-ML in forensic settings. This includes the development of culturally appropriate paradigms, the establishment of standardized protocols, and rigorous evaluation of the cost-effectiveness and feasibility of this technology within the Indonesian legal framework.

6. References

1. Joshi G, Tasgaonkar V, Deshpande A, Desai A, Shah B, Kushawaha A, et al. Multimodal machine learning for deception detection using behavioral and physiological data. Sci Rep. 2025; 15(1): 8943.
2. Pace G, Orrù G, Monaro M, Gnoato F, Vitaliani R, Boone KB, et al. Malingering detection of cognitive impairment with the b Test is boosted using machine learning. Front Psychol. 2019; 10: 1650.
3. Stevens A, Licha C. The Word Memory Test in medicolegal assessment: a measure of effort and malingering? J Forens Psychiatry Psychol. 2018; 30(2): 1–30.
4. van der Heide D, Boskovic I, van Harten P, Merckelbach H. Overlooking feigning behavior may result in potential harmful treatment interventions: Two case reports of undetected malingering. J Forensic Sci. 2020; 65(4): 1371–5.
5. Rubenzer S. The case for assessing for negative response bias, not malingering. J Forensic Psychol Res Pract. 2020; 20(4): 323–40.
6. Chan I, Ong ISM, Gwee Dpsych K. Validation of the test of memory malingering in a clinical

population from Singapore. Int J Forensic Ment Health. 2021; 20(1): 1–16.
7. Zhong S, Liang X, Wang J, Mellsop G, Zhou J, Wang X. Simulated malingering on binomial forced-choice digit memory test – using eye movements to understand faking cognition impairment process. J Forens Psychiatry Psychol. 2021; 32(6): 808–24.
8. Jargin SV. Military aspects of malingering, sexual and reproductive coercion: Report from Russia. J Forens Invest. 2022;10(1).
9. Mulligan L, O'Neill A, Minchin M, Heathcote L, Edge D, Shaw J, et al. Ethnicity and older adults in the criminal justice system: a brief report from a nominal group. J Forens Psychiatry Psychol. 2025; 36(1): 12–23.
10. O'Neill A, Thompson E, Wong E, Heathcote L, Shaw J, Robinson CA, et al. Dementia in the criminal justice system: a brief report from a nominal group. J Forens Psychiatry Psychol. 2025; 36(1): 76–88.
11. Heathcote L, O'Neill A, Newton-Clarke A, Hewson T, Senior J, Robinson C, et al. Older adults' trajectories through the criminal justice system: a brief report from a nominal group. J Forens Psychiatry Psychol. 2025; 36(1): 61–75.
12. Almas I, Lordos A. A narrative review of psychopathy research: current advances and the argument for a qualitative approach. J Forens Psychiatry Psychol. 2025; 1–51.
13. Taşkale N, Babcock JC. Intimate partner violence victim typologies: an inquiry with violence antecedents and coping as predictors. J Forens Psychiatry Psychol. 2025; 1–17.
14. Ashworth G, Waldron G, Kahai B. A service evaluation exploring staff perceptions about the use and impact of electronic monitoring (GPS tracking) in a medium secure forensic psychiatric unit. J Forens Psychiatry Psychol. 2025; 1–18.
15. Fernandes C, Maguire T, Berthollier N, Cheung J, Sivyer K. A qualitative review and thematic synthesis of resident experiences in prison-based democratic therapeutic communities. J Forens Psychiatry Psychol. 2025; 1–32.
16. Cheng P, Cai C, Park P. Finite frequency distributed fault detection in sensor networks with memory event-triggered scheme and deception attacks. ISA Trans. 2025.
17. Dai S, Hai L, Liu J, Tian E. Fault detection for hybrid-triggered networked systems subject to delay and deception attacks. J Franklin Inst. 2025; (107682): 107682.
18. Ten Brinke L, Sprigings S, Brown C, Kam C, Delmas H. Behavioral detection of emotional, high-stakes deception: Replication in a registered report. Law Hum Behav. 2025; 49(2): 173–81.
19. Dhivya, Deepika, Kumar A, Tiwari KS, Bhatia D, Singh P. Unmasking deception harnessing noise cancellation for digital image forgery detection using feature-map convolutional neural networks. Int J Sens Wirel Commun Control. 2025; 15(2): 184–99.
20. Kurtz MR, Trapani JA, Argueta P, Levine TR, Serota KB, Kana RK. Cognitive mechanisms underlying deception detection in neurodiverse sample of autistic and non-autistic young adults. Res Autism. 2025; 124(202587): 202587.