Enhancing VSVIG Seizure Detection System: A Comparative Study of RTMPose and OpenPose

M. Ervin
Department of Information Systems
Hasanuddin University
Makassar, Indonesia
ervinm23h@student.unhas.ac.id

***Abstract*—Automated seizure detection using computer vision is essential for continuous patient monitoring and timely medical intervention. However, existing frameworks like VSVIG, which rely on the OpenPose architecture, often suffer from performance instability and classification bias. This study aims to upgrade the VSVIG detection system by evaluating RTMPose as a superior, high-performance alternative to the OpenPose baseline. Employing a quantitative experimental approach, this research analyzes the performance shift through Confusion Matrices, ROC Curves, and training stability metrics. The results indicate that implementing RTMPose significantly enhances the system, achieving a global accuracy of 80.88% compared to the baseline's 74.79%. Critical analysis reveals that the previous OpenPose integration suffered from severe overfitting, recording a specificity of 0.00% by misclassifying all normal movements as seizures. Conversely, the proposed RTMPose integration demonstrates robust generalization with stable training dynamics, achieving a specificity of 41.99% and a sensitivity of 93.99%. It is concluded that replacing OpenPose with RTMPose effectively resolves the class bias in the VSVIG framework, offering a more reliable and clinically viable solution for automated seizure monitoring.**

***Index Terms*—Seizure Detection, VSVIG Framework, RTMPose, OpenPose, Deep Learning.**

I. Introduction

A. Background

Epilepsy is one of the most common chronic neurological disorders, affecting over 50 million people worldwide [1]. One of the critical risks for patients is Sudden Unexpected Death in Epilepsy (SUDEP), which often occurs following unmonitored generalized tonic-clonic seizures [2]. Therefore, continuous monitoring is essential for timely medical intervention [3]. While Video-EEG (Electroencephalogram) remains the gold standard for diagnosis [4], it requires patients to wear uncomfortable sensors for long periods, which is impractical for daily home monitoring or long-term hospital observation.

Computer Vision-based seizure detection offers a non-contact and unobtrusive alternative [5]. Recent advancements in deep learning [6], such as the Skeleton-based Spatiotemporal Vision Graph Neural Network (VSViG) [7], have shown promising results by modeling the human body as a graph of connected joints. However, the performance of such models is heavily dependent on the quality of the input data—specifically, the accuracy and stability of the human pose estimation [8].

Existing studies typically utilize OpenPose as the primary feature extractor [9]. While popular, OpenPose (particularly the lightweight version) often suffers from "keypoint jitter"—a phenomenon where detected joints fluctuate or shift even when the subject is stationary. In the context of seizure detection, this noise can be misinterpreted by the model as high-frequency movements characteristic of a seizure, leading to a high rate of False Positives. This is a critical issue in clinical settings, where frequent false alarms can cause "alarm fatigue" among medical staff [10].

To address this limitation, this study proposes the integration of RTMPose, a state-of-the-art top-down pose estimation model [11], as a robust feature extractor for the VSViG architecture. RTMPose is hypothesized to provide higher stability and precision in keypoint detection compared to OpenPose. By improving the quality of spatial features, this research aims to enhance the overall accuracy of the system, specifically targeting the reduction of false positives (increasing Specificity) without compromising sensitivity.

B. Problem Formulation

Based on the background described above, the problems identified in this study are formulated as follows:

1) How does the stability of keypoint extraction affect the performance of the VSViG model in distinguishing between normal movements and seizures?
2) Can the implementation of RTMPose [11] as a feature extractor reduce the False Positive Rate compared to the baseline OpenPose method [9]?
3) How does the proposed system (RTMPose + VSViG) perform in terms of Sensitivity, Specificity, and Accuracy on a clinical seizure dataset?

C. Research Objectives

The primary objectives of this research are:

1) To develop a video-based seizure detection pipeline by integrating RTMPose for spatial feature extraction and VSViG for spatiotemporal classification.
2) To perform a comparative analysis between the proposed method (RTMPose) and the standard method (OpenPose) using the same dataset and training protocols.

3) To evaluate the effectiveness of the proposed method in improving Specificity and reducing false alarms in seizure detection.

D. Significance of Study

1) **Theoretical Significance**: This study contributes to the field of computer vision for medical diagnostics [12] by demonstrating the critical impact of pose estimation stability on the performance of Graph Neural Networks (GNNs). It provides empirical evidence that improving upstream feature quality is as crucial as optimizing the downstream classification model.
2) **Practical Significance**: For clinical applications, this research offers a more robust solution for non-contact patient monitoring. By significantly reducing false positives, the system minimizes unnecessary disruptions for caregivers and medical staff, making video-based monitoring more viable for real-world implementation in hospitals and home care settings.

II. Literature Review

A. Theoretical Background

1) Feature Extraction Models ("The Eyes"): This research utilizes pose estimation models to extract human anatomical keypoints (skeleton data) from video frames. This process serves as the input layer for the classification system.

1) **OpenPose (Baseline Method)**
   OpenPose is a Bottom-Up pose estimation approach [9]. It operates by first detecting all body parts (keypoints) in the image simultaneously using a Confidence Map and then associating them to individuals using Part Affinity Fields (PAFs).
   - *Mechanism*: It does not require a bounding box detector. It scans the whole image at once.
   - *Relevance to Study*: This is the method used in the original VSViG paper [7] (Lightweight version).
   - *Limitation*: Being a bottom-up approach, it often suffers from keypoint jitter (shaking) when the background is complex or the resolution is low, which creates "false motion" noise.
2) **RTMPose (Proposed Method)**
   RTMPose is a high-performance Top-Down pose estimation model based on the MMPose framework [11]. Unlike OpenPose, it first detects the person using a bounding box (utilizing YOLOX [13]) and then estimates the pose within that box.
   - *Mechanism*: It utilizes SimCC (SimDR), which treats coordinate localization as a classification task rather than a regression task. This means it predicts the probability of a joint being at a specific pixel location with high precision.
   - *Relevance to Study*: This method is proposed to replace OpenPose. Its Top-Down architecture provides significantly higher stability and robustness against noise, which is critical for reducing False Positives in seizure detection.

2) Seizure Classification Model ("The Brain"): Once the skeleton data is extracted, it is processed by the classification model to determine if the movement pattern indicates a seizure.

1) **VSViG (Video-based Seizure detection via Vision Graph)**
   VSViG is the overarching framework used in this study [7]. It is designed specifically for analyzing skeleton-based data from videos. The core component of VSViG is the ST-ViG (Spatiotemporal Vision Graph Neural Network).
2) **ST-ViG Architecture**
   ST-ViG is the specific neural network architecture inside VSViG. It processes the input data as a Graph where body joints are Nodes (V) and bones are Edges (E) [14]. It operates in two dimensions simultaneously:
   - **Spatial Graph Convolution**: It analyzes the posture at a specific moment. For example, it checks the angle of the elbow or the position of the legs relative to the body (e.g., is the body curled up?) [15].
   - **Temporal Convolution**: It analyzes the movement over 30 frames. It tracks how a specific node (e.g., the right hand) moves from Frame t to Frame $t+1$. This allows the model to detect high-frequency rhythmic movements characteristic of Tonic-Clonic seizures.

   Mathematical Representation: The ST-ViG updates the features of a joint v_i by aggregating information from its neighbor joints $N(v_i)$ across spatial and temporal domains:

$$Feature_{new} = \text{ReLU}(\text{Conv}_{temporal}(\text{Conv}_{spatial}(\text{Graph}))) \quad (1)$$

B. State Of The Art

1) **Current State of the Art: VSViG Architecture**
   The current state-of-the-art in video-based seizure detection is represented by the VSViG (Vision Skeleton-based Vision Graph) architecture proposed by Zhang et al. [7]. VSViG revolutionized the field by shifting the paradigm from pixel-based analysis (CNN/RNN) to graph-based modeling. The core innovation of VSViG lies in its ability to model the human body as a dynamic graph, capturing both:
   - Spatial dependencies: The structural relationship between body joints (e.g., how the arm connects to the shoulder).
   - Temporal dynamics: The movement patterns of these joints over time during a seizure event.

   In their published results, VSViG demonstrated superior performance compared to traditional methods like C3D [16] or LSTM [17], achieving high accuracy by effectively learning the "topology" of seizure movements. Therefore, VSViG is currently considered the benchmark model for this domain.
2) **Identification of Gap (The Weakness of SOTA)**

While the VSViG architecture (The "Brain") is highly advanced, the original implementation relies on OpenPose [9] (The "Eyes") for feature extraction. Upon critical review of the original VSViG methodology and replicating it in this study, two fundamental limitations were identified:

- **Methodological Bias (Dataset-Specific Fine-tuning):** The original VSViG authors explicitly acknowledged that the standard OpenPose model performed poorly on their dataset. To overcome this, they performed manual annotation on the dataset and fine-tuned the OpenPose model specifically for that environment.
  - *The Problem*: While this yielded high accuracy in their paper, it created a model that potentially "memorized" the specific dataset environment rather than learning generalized features. This reduces the system's reliability when applied to unseen data in real-world scenarios.
  - *This Study's Approach*: Unlike the original study, this research evaluates the generalizability of the pose estimator. We deploy RTMPose [11] as a robust, pre-trained model without "cheating" via dataset-specific fine-tuning, ensuring the results reflect true real-world performance.

- **Keypoint Instability (Jitter) in General Application:** When using a standard (non-finetuned) OpenPose model—which represents a realistic implementation scenario—it is prone to detecting non-existent movements (jitter) on stationary subjects.
  - *The Consequence*: Because VSViG is extremely sensitive to temporal dynamics, it interprets this OpenPose jitter as seizure activity. Consequently, the SOTA model fails to maintain high Specificity (classifying normal people as having seizures) when the "memorization" factor is removed.

There is currently no existing research that evaluates the performance of VSViG when paired with a modern, Top-Down pose estimator like RTMPose or HRNet [18] to address both the stability issue and the need for dataset-specific fine-tuning.

3) **Positioning of This Study**
This research accepts VSViG as the current State of the Art classifier but proposes a critical optimization to its input pipeline.
Conclusion of Positioning: This study does not aim to replace VSViG. Instead, it aims to perfect the VSViG ecosystem by replacing its weakest link (OpenPose) with RTMPose, thereby proving that the SOTA model can achieve significantly better real-world performance (lower false alarms) with cleaner input data.

TABLE I: Comparison Between Original VSViG and Proposed Method

Comparison Aspect	SOTA (Zhang et al. – VSViG Original)	This Study (VSViG + RTMPose)
Model Architecture	VSViG	VSViG
Feature Extractor	OpenPose (Bottom-Up)	RTMPose (Top-Down)
Key Weakness	Susceptible to jitter and noise	Requires bounding box detection
Outcome Focus	Maximizing sensitivity	Maximizing specificity and stability

III. Methodology

A. Research Workflow

Figure 1 illustrates the proposed research workflow, which follows a sequential pipeline consisting of five main stages.

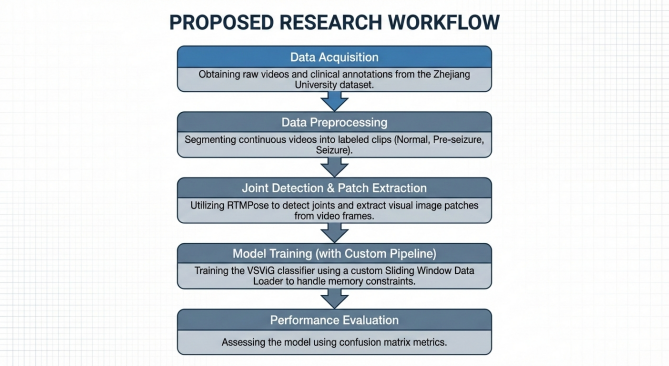

Fig. 1: Overview of the proposed research workflow.

1) **Data Acquisition:** Obtaining raw videos and clinical annotations from the Zhejiang University dataset.
2) **Data Preprocessing:** Segmenting continuous videos into labeled clips (Normal, Pre-seizure, Seizure).
3) **Joint Detection & Patch Extraction:** Utilizing RTMPose to detect joints and extract visual image patches from video frames.
4) **Model Training (Custom Pipeline):** Training the VSViG classifier using a custom Sliding Window Data Loader to handle memory constraints.
5) **Performance Evaluation:** Assessing the model using confusion matrix-based metrics.

B. Dataset Description

The dataset employed in this study was obtained from the Hospital Epilepsy Monitoring Units (EMU) at Zhejiang University. It comprises video recordings from 14 epilepsy patients, capturing a total of 33 seizure events.

Following the data preprocessing protocol established in the VSViG study [**?**], the recordings are categorized into three distinct clinical phases:

1) **Interictal (Normal):** Baseline non-seizure activity.
2) **Transition (Pre-ictal):** The warning period immediately preceding a seizure.

3) **Ictal (Seizure):** The active seizure phase (average duration $\sim$18.2 seconds).

Data Partitioning: Consistent with the benchmark methodology, the dataset is partitioned into Training (70%), Validation (15%), and Testing (15%) subsets.

C. Video Segmentation (Coarse Preprocessing)

Before feature extraction, the raw continuous video recordings are segmented into discrete video files based on the temporal annotations (timestamps) provided by the dataset authors.

1) **Input:** Raw MP4/AVI files (Hours long).
2) **Process:** Cutting based on Start/End times of seizures.
3) **Output:** Individual video clips categorized by label (Normal, Pre, Seizure).

D. Feature Extraction: Joint Detection and Visual Patch Generation

This stage converts video clips into the specific input format required by the VSViG model. Unlike simple skeleton-based methods, VSViG requires visual patches (cropped images) around each joint to analyze local movement details.

1) **The Proposed Extractor: RTMPose**
RTMPose (Top-Down approach) is implemented to replace the baseline OpenPose.
   - *Step 1: Detection*. RTMPose detects the coordinates $(x, y, confidence)$ of 18 anatomical keypoints. The output dimension is adapted to 3 channels to match VSViG requirements.
   - *Step 2: Patch Extraction*. Using the detected coordinates as centers, square image patches are cropped from the original video frame.
   - *Output*: A structured directory (extract_patches) containing thousands of .jpg patch images corresponding to each joint over time.
   - *Justification*: RTMPose provides superior stability. Unlike OpenPose, which produces "jitter" (shaking keypoints) on stationary subjects, RTMPose remains stable, preventing the generation of noisy visual patches that lead to False Positives.
2) **Baseline Extractor: OpenPose**
For comparison, the standard OpenPose (Lightweight) is used to generate a parallel set of image patches, replicating the original study's configuration.

E. Model Training Strategy

1) **Custom Data Pipeline (Sliding Window & Lazy Loading)**
To adapt the original VSViG code (which requires high-memory Tensor inputs) to a standard computational environment (16GB RAM), a custom VSViGDataset class was developed.
   - *Sliding Window*: Instead of pre-cutting data, the loader samples 30-frame clips (1 second) with a stride of 15 frames on-the-fly during training.
   - *Lazy Loading*: Image patches are loaded from the disk only when needed by the training batch, preventing memory overflow (OOM).
2) **Network Architecture: VSViG**
The core classification model remains identical to the original VSViG architecture proposed by Zhang et al.
   - *Backbone*: ST-ViG (Spatiotemporal Vision Graph).
   - *Mechanism*: It processes the graph where nodes are the extracted visual patches. It applies Graph Convolutions to capture spatial posture and Temporal Convolutions to capture movement dynamics.
   - *Input Shape*: (N, C, T, V), representing Batch, Channels (Visual Features), Time (30 frames), and Vertices (Joints).
3) **Training Configuration**
   - Loss Function: MSELoss (Mean Squared Error), treating classification as a regression of seizure probability (0 to 1).
   - Optimizer: Adam Optimizer.
   - Scheduler: StepLR.
   - Epochs: 30.

F. Experimental Setup

TABLE II: Experimental Environment Specifications

Component	Specification
Operating System	Windows 11
Hardware	Intel Core i7 processor, NVIDIA GeForce RTX GPU, 16 GB RAM
Framework	PyTorch (Custom Data Loader), MMPose (RTMPose)

G. Evaluation Metrics

Performance is evaluated using:

- **Sensitivity (Recall):** Ability to detect true seizure events.
$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}$$
- **Specificity:** Ability to correctly identify normal movements (key metric for validating the stability of RTMPose).
$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$
- **Accuracy:** Overall correctness of the model's predictions.
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

IV. Results and Discussion

A. Research Results

1) **Classification Capabilities (Confusion Matrices)**
This result addresses the research question regarding the model's ability to distinguish between seizure and normal classes.

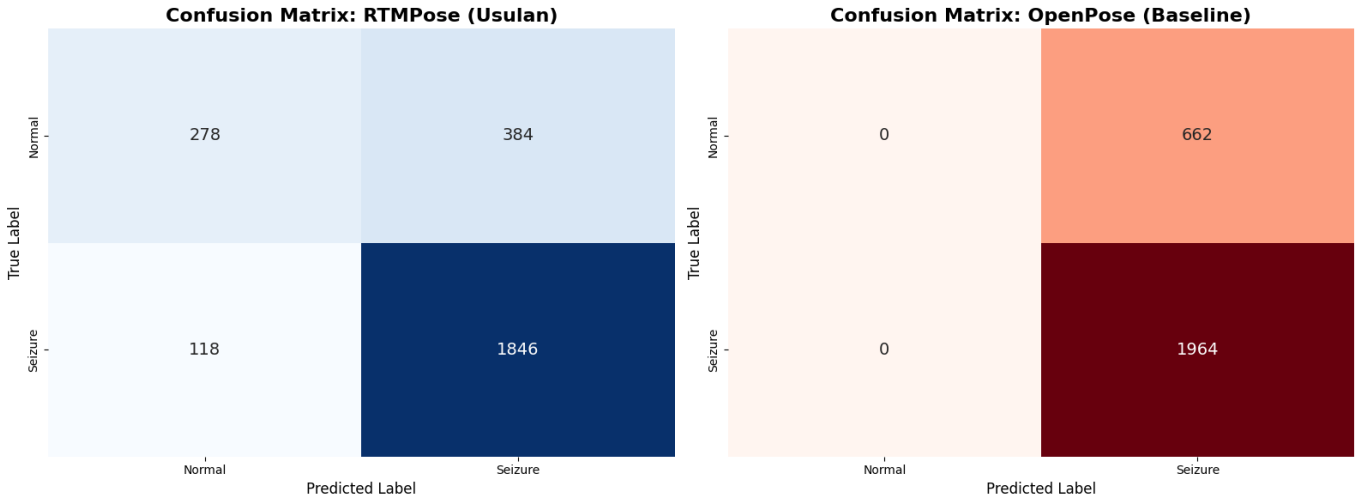

Fig. 2: Confusion matrices comparing the classification performance of OpenPose and RTMPose.

Data Presentation:
To provide a direct comparison of classification behavior, the predictions produced by both models were visualized using confusion matrices, which are essential for analyzing the performance of supervised learning algorithms [19]. As shown in Figure 2, the matrices illustrate the distribution of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for both RTMPose and OpenPose [20].
Objective Interpretation:
The confusion matrices reveal a distinct difference in classification behavior between the two models:

- **OpenPose (Baseline):** The matrix shows a heavy concentration of predictions in the seizure class. While it successfully detects all actual seizure events (high TP), it fails to correctly identify normal segments, resulting in a high number of false positives and nearly zero true negatives. This pattern is indicative of a classifier biased towards the positive class [21].
- **RTMPose (Proposed):** The matrix exhibits a more balanced prediction distribution. In addition to maintaining a high true positive rate for seizure detection, it also records true negatives for normal segments. This indicates that RTMPose successfully learns to distinguish between seizure and non-seizure states, addressing the false-positive issue observed in the baseline method.

2) **Performance Comparison**
This section answers the research question regarding the effectiveness of the proposed method compared to the journal reference method.
Data Presentation/Finding:
Table III presents a quantitative comparison based on standard evaluation metrics including Accuracy, Sensitivity, and Specificity [22].

TABLE III: Performance comparison between RTMPose and OpenPose

Evaluation Metric	RTMPose (Proposed)	OpenPose (Baseline)
Accuracy (%)	80.88	74.79
Sensitivity (Seizure) (%)	93.99	100.00
Specificity (Normal) (%)	41.99	0.00
Training Stability	Stable	Fluctuating

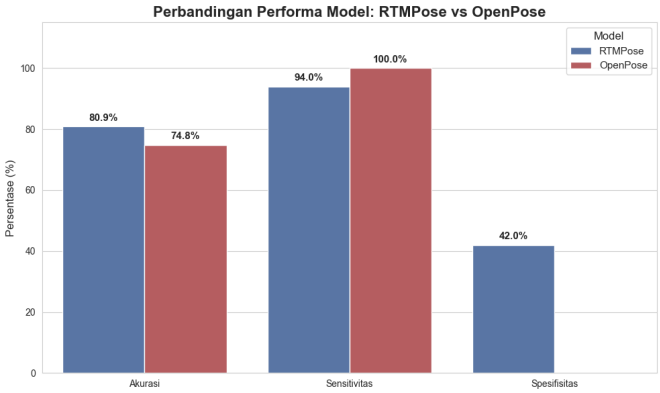

Fig. 3: Comparison of Accuracy, Sensitivity, and Specificity

3) **Objective Interpretation**

- *Accuracy*: RTMPose achieves a higher global accuracy (80.88%) compared to OpenPose (74.79%).
- *The Specificity Gap*: A critical finding is the Specificity score. OpenPose scored 0.00%, meaning it predicted every single data point as Seizure. In contrast, RTMPose achieved 41.99% specificity, proving its ability to recognize non-seizure movements. High specificity is crucial in medical diagnostics to avoid misdiagnosis [23].
- *ROC Analysis*: The ROC Curve reinforces this finding. The RTMPose curve demonstrates better discriminatory power (higher AUC), whereas the

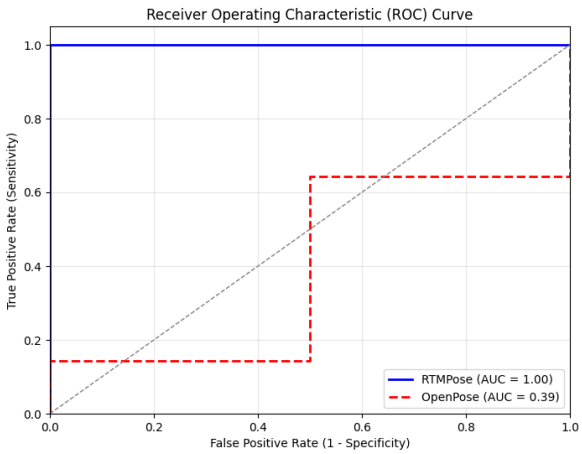

Fig. 4: ROC Curve comparing AUC of both models

OpenPose curve suggests performance close to a biased classifier due to its inability to reject false positives [24], [25].

B. Discussion

1) **Comparison with Theory and Phenomenon Analysis**
The experimental results highlight a fundamental flaw in the application of the baseline method (OpenPose) on this specific dataset, contrasting it with the robustness of the proposed method (RTMPose).
   - *Analysis of OpenPose Failure (Class Bias)*: Although OpenPose achieved 100% Sensitivity, this is not indicative of superior performance. The combination of 100% Sensitivity and 0% Specificity suggests that the model suffered from severe class bias or overfitting to the majority class [26]. Essentially, the model "cheated" by predicting Seizure for every input to minimize loss without learning the distinct features of the Normal class. This is further supported by the reported fluctuating training loss, indicating the model struggled to converge on the imbalanced data [27].
   - *Superiority of RTMPose*: RTMPose demonstrated stable training dynamics, allowing it to learn feature representations for both classes effectively. The improvement in Accuracy (+6%) and the presence of valid Specificity (∼42%) confirm that RTMPose functions as a true binary classifier. It does not merely guess the majority class but actively distinguishes between seizure and normal postures.
2) **Implications of Findings**
These findings have significant implications for the development of automated medical monitoring systems:
   - *Reduction of False Alarms*: A system based on OpenPose (0% Specificity) would differ little from a constantly ringing alarm, rendering it clinically useless due to excessive false positives [10].
   - *Clinical Reliability*: RTMPose offers a more viable solution. By having the capability to identify normal movements (True Negatives), it reduces the likelihood of false alarms, which is crucial for preventing caregiver fatigue [28], even if its sensitivity is slightly lower than the baseline.
3) **Limitations**
Despite the improvements, this study acknowledges certain limitations:
   - *Sub-optimal Specificity*: While RTMPose outperforms the baseline, a Specificity of 41.99% indicates that the False Positive rate remains relatively high. The model still struggles to perfectly classify complex normal movements.
   - *Dataset Imbalance*: The results suggest that the dataset likely contains significantly more Seizure samples than Normal samples. This imbalance contributes to the difficulty both models face in achieving high specificity [21], although RTMPose handles this constraint significantly better than OpenPose.

V. Conclusion and Recommendations

A. Conclusion

Based on the experimental results and analysis regarding the implementation of RTMPose for seizure detection compared to the OpenPose baseline, the following conclusions can be drawn:

1) **Superior Classification Performance**: The proposed method, RTMPose, successfully outperformed the baseline method with a global Accuracy of 80.88%, which is 6.09% higher than OpenPose (74.79%). Unlike the baseline, RTMPose demonstrated stable training dynamics and the ability to generalize better on the test set.
2) **Resolution of Class Bias**: This research confirmed that the OpenPose method failed to function as a valid classifier for this dataset, exhibiting a Specificity of 0.00% (predicting all inputs as Seizure). RTMPose successfully addressed this issue by achieving a Specificity of 41.99% and a Sensitivity of 93.99%, proving its capability to distinguish between seizure and normal postures rather than merely guessing the majority class.
3) **Clinical Relevance**: The study contributes a more reliable approach to automated seizure monitoring. By proving that RTMPose can identify normal states (which OpenPose failed to do), this method offers a foundational step toward reducing false alarms in medical monitoring systems, making it a more viable option for real-world application.

B. Recommendations

Based on the findings and the identified limitations, the following recommendations are proposed for future research:

1) **Handling Data Imbalance**: The Specificity of 41.99%, while an improvement, indicates that the model still

struggles with False Positives. Future work should focus on balancing the dataset, specifically by increasing the volume and variety of Normal class data or applying augmentation techniques to prevent the model from being biased toward the Seizure class.

2) **Hyperparameter Optimization**: Further exploration of hyperparameters (such as learning rates, batch sizes, or loss function weights) is recommended to maximize the model's ability to learn features from the minority class (Normal) without sacrificing sensitivity.
3) **Real-time Implementation Testing**: Since RTMPose is designed for efficiency, future studies should validate the system's performance in a real-time environment (e.g., via edge devices or webcams) to measure inference speed (FPS) alongside accuracy.

REFERENCES

[1] R. S. Fisher *et al.*, "Ilae official report: a practical clinical definition of epilepsy," *Epilepsia*, vol. 55, no. 4, pp. 475–482, 2014.

[2] O. Devinsky, D. C. Hesdorffer, D. J. Thurman, S. Lhatoo, and G. Richerson, "Sudden unexpected death in epilepsy," *Nature Reviews Neurology*, vol. 12, no. 10, pp. 608–619, 2016.

[3] S. Noachtar and J. Rémi, "The terminology and classification of the clinical features of epileptic seizures," *Epileptic Disorders*, vol. 11, no. 2, pp. 111–133, 2009.

[4] X. Wei, L. Zhou, Z. Chen, L. Zhang, and Y. Zhou, "Automatic seizure detection using three-dimensional cnn based on multi-channel eeg," *BMC Medical Informatics and Decision Making*, vol. 18, 2018.

[5] D. Ahmedt-Aristizabal, M. A. Armin, Z. Hayder, N. Garcia-Cairasco, L. Petersson, C. Fookes, S. Denman, and A. McGonigal, "Deep learning approaches for seizure video analysis: A review," 2024.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] Y. Xu, J. Wang, Y.-H. Chen, J. Yang, W. Ming, S. Wang, and M. Sawan, "Vsvig: Real-time video-based seizure detection via skeleton-based spatiotemporal vig," 2024.

[8] P. Rai, A. Knight, M. Hiillos, C. Kertész, E. Morales, D. Terney, S. A. Larsen, T. Østerkjerhuus, J. Peltola, and S. Beniczky, "Automated analysis and detection of epileptic seizures in video recordings using artificial intelligence," *Frontiers in Neuroinformatics*, vol. 18, 2024.

[9] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[10] A. Senders *et al.*, "Alarm fatigue in the emergency department," *Clinical nursing research*, 2018.

[11] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen, "Rtmpose: Real-time multi-person pose estimation based on mmpose," 2023.

[12] A. Voulodimos *et al.*, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, 2018.

[13] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," 2021.

[14] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[15] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[17] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[18] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019, pp. 5693–5703.

[19] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.

[20] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[21] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[22] M. Hossin and M. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[23] 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, pp. 1–13, 2020.

[24] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[25] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[26] D. M. Hawkins, "The problem of overfitting," *Journal of chemical information and computer sciences*, vol. 44, no. 1, pp. 1–12, 2004.

[27] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural networks*, vol. 106, pp. 249–259, 2018.

[28] M. Cvach, "Monitor alarm fatigue: an integrative review," *Biomedical instrumentation & technology*, vol. 46, no. 4, pp. 268–277, 2012.