Modern education and development
ISSN: 3060-4567

BIOSIGNAL PREPROCESSING METHODS: A COMPREHENSIVE REVIEW OF NOISE REMOVAL, FILTERING AND SEGMENTATION TECHNIQUES FOR AI-BASED PREDICTION SYSTEMS

Qarshiyeva Jamila yashnar qizi
Osiyo texnologiyalar universiteti o`qituvchisi
TATU 2-bosqich tayanch doktoranti
E-mail: jamiqarshi@gmail.com
Tel raqam: 99891 952-02-64
ORCID: - 0009-0003-6614-6723

Keywords: *biosignal preprocessing, ECG noise removal, baseline wander, pandpass filter, wavelet transform, adaptive filtering, Pan-Tompkins algorithm, signal segmentation, deep learning, artifact removal*

Abstract: *Effective preprocessing is a critical determinant of the accuracy and reliability of artificial intelligence (AI)-based biosignal prediction systems. Raw biosignals acquired through sensors are inevitably contaminated by multiple noise sources including baseline wander, powerline interference, electromyographic artifacts, and electrode motion artifacts. This paper presents a comprehensive review of preprocessing methods for ECG, EEG, and EMG biosignals, systematically covering noise characterization, digital filtering techniques segmentation algorithms (Pan-Tompkins), and normalization approaches.*

The quality of biosignal data processing directly determines the performance of any downstream artificial intelligence (AI) system. Raw biosignals acquired through electrodes and sensors inevitably contain multiple noise components that corrupt the physiologically meaningful information. In electrocardiogram (ECG) analysis, it has been demonstrated that unprocessed signals yield classification accuracy 4–8% lower than properly preprocessed signals [1]. Similar degradation is observed in electroencephalogram (EEG) and electromyogram (EMG) processing pipelines.

Preprocessing encompasses all signal transformations applied between raw sensor acquisition and feature extraction or model inference. A standard biosignal preprocessing pipeline includes: (1) noise characterization and source identification; (2) analog and digital filtering; (3) signal segmentation into analysis windows; and (4) normalization for model compatibility. Each stage must be carefully designed to maximize signal-to-noise ratio (SNR) while preserving diagnostically relevant morphological features [2].

Despite its critical role, preprocessing is frequently underreported in AI-based biosignal studies. A systematic review of 368 ECG deep learning papers found that fewer than 40% provided complete preprocessing specifications, limiting reproducibility [3]. This paper addresses this gap through a comprehensive, reproducible review of preprocessing methods across ECG, EEG, and EMG signal types.

Understanding noise sources is the prerequisite for selecting appropriate preprocessing methods. Biosignal noise sources are categorized into four principal types:

- Baseline wander (BW): Low-frequency drift (0.05–0.5 Hz) caused by respiratory movement, body motion, and electrode impedance variation. BW shifts the isoelectric baseline, distorting ST segment measurements critical for ischemia detection [1].
- Powerline interference (PLI): Narrowband noise at 50 Hz (Europe/Asia) or 60 Hz (Americas) induced by electromagnetic coupling from electrical equipment. PLI appears as a sinusoidal overlay on the biosignal [2].
- Electromyographic (EMG) artifacts: Broadband noise (20–500 Hz) generated by skeletal muscle contraction. In ECG recordings, respiratory muscle

activity produces characteristic burst patterns that can be misclassified as arrhythmias [4].

- Electrode motion artifacts: Transient impedance changes at the electrode-skin interface due to patient movement, producing large-amplitude waveform distortions that can corrupt multiple cardiac cycles.

The Butterworth filter is the most widely used preprocessing filter in ECG analysis due to its maximally flat frequency response in the passband. A bandpass configuration with cutoff frequencies of 0.5 Hz (high-pass) and 40 Hz (low-pass) simultaneously removes baseline wander and high-frequency EMG noise while preserving the diagnostically relevant 0.5–40 Hz ECG band [2]. The filter order n determines the steepness of roll-off: higher orders provide sharper transitions but introduce greater phase distortion. Second-order (n=2) Butterworth filters implemented as second-order sections (SOS) are preferred for numerical stability.

Implementation using scipy.signal in Python: butter(N=2, Wn=[0.5, 40], btype='bandpass', fs=360, output='sos') followed by sosfiltfilt() for zero-phase filtering. Zero-phase implementation is critical for ECG preprocessing as phase distortion would alter PR and QT interval measurements used in clinical diagnosis.

A second-order IIR notch filter centered at 50 Hz (Q-factor = 30) selectively attenuates powerline interference while preserving adjacent frequency content. The narrow bandwidth (±1.7 Hz at -3 dB) ensures minimal distortion of the biosignal. For recordings in 60 Hz environments, the center frequency is adjusted accordingly. When PLI frequency drifts due to power grid instability, adaptive notch filters that track the instantaneous fundamental frequency provide superior performance [1].

Discrete Wavelet Transform (DWT) decomposes the biosignal into approximation and detail coefficients at multiple resolution levels. <At each level, detail coefficients dominated by noise are thresholded using soft or hard thresholding: coefficients below a threshold $\lambda = \sigma\sqrt{(2 \log N)}$ (where σ is the noise standard deviation estimated from finest-scale coefficients and N is the signal length) are zeroed or shrunk. The Daubechies-6 (db6) mother wavelet is widely

recommended for ECG denoising due to its morphological similarity to QRS complex shape [2].

DWT denoising achieves SNR improvement of 15–25 dB for baseline wander removal compared to 10–18 dB for Butterworth filtering, at the cost of higher computational complexity. The stationary wavelet transform (SWT), which avoids downsampling, provides shift-invariant denoising superior to the decimated DWT for biomedical applications [1].

Adaptive filters adjust their coefficients in real time based on a reference noise signal, enabling cancellation of non-stationary noise components that fixed filters cannot address. The Least Mean Squares (LMS) algorithm updates filter weights $w(n+1) = w(n) + 2\mu e(n)x(n)$, where μ is the step size, $e(n)$ is the error signal, and $x(n)$ is the reference input. LMS adaptive filters are particularly effective for EMG artifact cancellation in ECG when a separate EMG reference channel is available [4]. Recursive Least Squares (RLS) converges faster than LMS but requires greater computational resources.

Segmentation divides the continuous biosignal stream into analysis windows aligned with physiologically meaningful events. For ECG, the Pan-Tompkins algorithm [5] remains the standard for R-peak detection, achieving sensitivity > 99.5% on the MIT-BIH benchmark. The algorithm applies differentiation, squaring, and moving-window integration to enhance QRS energy, followed by adaptive thresholding that adjusts to signal amplitude variations. Fixed windows of 0.9 seconds centered on R-peaks (0.3 s before, 0.6 s after) capture complete cardiac cycles for single-beat classification.

For EEG analysis, event-related potential (ERP) paradigms use stimulus-locked segmentation with epochs typically spanning -200 ms to +800 ms relative to stimulus onset. For EMG, sliding windows of 100–300 ms with 50% overlap are standard for continuous activity classification. Window length selection involves a trade-off: longer windows provide more frequency resolution but reduce temporal precision for detecting transient events [4].

Normalization standardizes signal amplitude across recordings to ensure consistent model inputs. Z-score normalization (zero mean, unit variance) is the most widely adopted approach: x_norm = (x - μ) / σ, where μ and σ are the segment mean and standard deviation. This approach is robust to amplitude differences between patients and electrode placements. Min-max normalization scale signals to [0, 1] but is sensitive to outliers introduced by motion artifacts.

For multi-lead ECG, per-lead normalization is preferred over global normalization to preserve inter-lead amplitude relationships that carry diagnostic information (e.g., ST elevation in anatomically contiguous leads). Batch normalization applied within deep learning architectures provides an additional normalization layer that adapts during training, reducing the impact of preprocessing parameter choices.

Recent advances in deep learning have enabled end-to-end denoising architectures that learn optimal noise removal directly from data. Convolutional denoising autoencoders (CDAE) learn a compressed noise-invariant representation of clean signals, outperforming wavelet methods by 3–5 dB SNR in composite noise conditions [1]. U-Net architectures with skip connections preserve fine morphological details during denoising, achieving state-of-the-art performance on ECG baseline wander removal.

Empirical Mode Decomposition (EMD)-based adaptive methods decompose signals into intrinsic mode functions (IMFs), selectively reconstruct noise-free signals from relevant IMFs. EMD outperforms fixed wavelet bases for non-stationary noise but requires higher computational resources unsuitable for real-time edge deployment [6]. Table 2 summarizes the comparative performance of preprocessing methods. The proposed preprocessing pipeline — Butterworth bandpass filter (0.5–40 Hz, 2nd order), 50 Hz notch filter (Q=30), Pan-Tompkins R-peak detection, fixed 0.9 s window segmentation, and z-score normalization — was evaluated on two benchmark datasets: MIT-BIH Arrhythmia Database [7] (48 recordings, 360 Hz, 19 classes) and Zheng et al. (2022) 12-lead ECG dataset [7] (10,000+ recordings, 500 Hz, 15 classes).

A 1D-CNN classifier was trained on both raw and preprocessed signals to quantify the preprocessing contribution. On MIT-BIH, preprocessing improved classification accuracy from 93.9% (raw) to 98.1% (preprocessed), a gain of 4.2 percentage points. Sensitivity improved from 92.1% to 97.6% and specificity from 94.8% to 98.5%. On the Zheng dataset, accuracy improved from 91.7% to 96.2% (+4.5%). These results confirm that preprocessing is not merely a convenience but a statistically significant contributor to AI classification performance ($p < 0.001$, paired t-test).

Wavelet-based preprocessing (DWT, db6, level 5) achieved marginally higher SNR (+2.1 dB) compared to Butterworth filtering but at 3.4× computational cost. For real-time wearable applications where inference latency is constrained to < 50 ms, Butterworth + notch filtering provides the optimal accuracy-efficiency balance.

This paper presented a comprehensive review and experimental evaluation of biosignal preprocessing methods for AI-based prediction systems. The following principal conclusions are drawn:

1) A standardized preprocessing pipeline — Butterworth pandpass (0.5–40 Hz), notch filter (50 Hz), Pan-Tompkins segmentation, and z-score normalization — improves AI classification accuracy by 4.2–4.5% compared to unprocessed signals;

2) Wavelet DWT denoising achieves the highest SNR improvement (15–25 dB) among classical methods and is recommended when computational resources permit;

3) Deep learning autoencoders provide state-of-the-art denoising (22–32 dB SNR gain) but are unsuitable for real-time edge deployment due to computational cost;

4) Zero-phase filtering implementation is mandatory for ECG applications to preserve interval measurements used in clinical diagnosis;

5) Complete preprocessing specification reporting should become a standard requirement in AI biosignal research publications.

REFERENCES.

1. Ansari M.S. et al. Preprocessing and Denoising Techniques for Electrocardiography and Magnetocardiography: A Review. Bioengineering. 2024; 11(11): 1109. DOI: 10.3390/bioengineering11111109

2. Plux Biosignals. Biosignals Processing 101: Pre-Processing — Removing noise from raw biosignals. 2023. URL: https://www.pluxbiosignals.com/blogs/informative/biosignals-processing-101-removing-noise-from-raw-biosignals

3. Xiao Q. Deep Learning-Based ECG Arrhythmia Classification: A Systematic Review. Applied Sciences. 2023; 13(8): 4964. DOI: 10.3390/app13084964

4. Kaniusas E. Advanced Bioelectrical Signal Processing Methods: Past, Present, and Future Approach — Part III: Other Biosignals. Sensors. 2021; 21(18): 6064. DOI: 10.3390/s21186064

5. Pan J., Tompkins W.J. A r eal-time QRS detection algorithm. IEEE Transactions on Biomedical Engineering. 1985; 32(3): 230–236. DOI: 10.1109/TBME.1985.325532

6. Hussein A.F. et al. An Adaptive ECG Noise Removal Process Based on Empirical Mode Decomposition (EMD). Contrast Media & Molecular Imaging. 2022; 2022: 3346055. DOI: 10.1155/2022/3346055

7. Goldberger A.L. et al. PhysioBank, PhysioToolkit, and PhysioNet. Circulation. 2000; 101(23): e215–e220. DOI: 10.1161/01.CIR.101.23. e 215

8. Zheng J., Guo H., Chu H. A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0.0). PhysioNet. 2022. DOI: 10.13026/wgex-er52

9. Kiranyaz S. et al. 1D convolutional neural networks and applications: A survey. Mechanical Systems and Signal Processing. 2021; 151: 107398. DOI: 10.1016/j.ymssp.2020.107398

10. Karimanzira D. et al. A two-step pre-processing tool to remove Gaussian and ectopic noise for heart rate variability analysis. PMC. 2022. DOI: 10.1371/journal.pone.0276424

11. Alotaiby T. Electroencephalography Signal Processing: A Comprehensive Review. Sensors. 2023; 23(14): 6434. DOI: 10.3390/s23146434

12. Ansari Y. Deep learning for ECG Arrhythmia detection and classification: 2017–2023. Frontiers in Physiology. 2023; 14: 1246746. DOI: 10.3389/fphys.2023.1246746

13. qizi Qarshiyeva, J. Y. (2023). MATLAB TIZIMIDA SIGNALLARNI APPROKSIMATSIYALASH. *GOLDEN BRAIN*, *1*(28), 191-195.

14. qizi Qarshiyeva, J. Y. (2023). MATLAB TIZIMIDA SIGNALLARNI INTERPOLYATSIYALASH MASALALARINI YECHISH. *GOLDEN BRAIN*, *1*(28), 186-190.

15. КАРШИЕВА, Д. (2022). К ВОПРОСУ ЦИФРОВОЙ ОБРАБОТКИ СИГНАЛОВ В ИНТЕЛЛЕКТУАЛЬНЫХ СЕНСОРАХ. In *Молодежь и системная модернизация страны* (pp. 364-367).

16. Элов, Д. Б. (2023). БАЛАНСИРОВКА ЭНЕРГЕТИЧЕСКИХ ХАРАКТЕРИСТИК ДАТЧИКОВ В БЕСПРОВОДНЫХ СЕНСОРНЫХ СЕТЯХ: BALANCING THE ENERGY CHARACTERISTICS OF SENSORS IN WIRELESS SENSOR NETWORKS. Молодой специалист, 2(10), 29-33.

17. Normamatov, X. (2025). IMPROVING THE METHODOLOGY OF TEACHING PROGRAMMING LANGUAGES BASED ON NETWORK TECHNOLOGIES. International Journal of Artificial Intelligence, 1(2), 656-662.

18. Normamatov, X. (2025). APPLYING INTERNATIONAL EXPERIENCES IN TEACHING PROGRAMMING TO HIGHER EDUCATION SPECIALIST

STUDENTS: CHALLENGES AND SOLUTIONS. International Journal of Artificial Intelligence, 1(2), 648-650.

19. Normamatov, X. (2025). CHALLENGES AND SOLUTIONS IN TEACHING PROGRAMMING: AN EXPLORATION OF GLOBAL AND LOCAL PERSPECTIVES. International Journal of Artificial Intelligence, 1(2), 651-655.

20. Нормаматов, Х. М., & Абдуллаева, С. У. (2015). ЭФФЕКТИВНОСТЬ ПРИМЕНЕНИЯ АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ УПРАВЛЕНИЯ" Э-БОЛЬНИЦА". In Инновации в технологиях и образовании (pp. 117-119).

21. Нормаматов, Х. М. (2014). ЛИНЕЙНЫЕ СИСТЕМЫ В ЦИФРОВОЙ ОБРАБОТКЕ СИГНАЛОВ. In Инновации в строительстве глазами молодых специалистов (pp. 239-241).

22. Шеров, Ж. Э., & Нормаматов, Х. М. (2015). АВТОМАТИЗАЦИЯ УПРАВЛЕНИЯ ВЫСШЕГО УЧЕБНОГО ЗАВЕДЕНИЯ. In Инновации в технологиях и образовании (pp. 178-182).