Model	Typ danych	Zastosowanie	Dokładność (%)	Architektura	Rozmiar/Precyzja (Inferred)
EmotionTFN (Emotion-Multi-Scale Temporal Fusion Network)	Multimodalne (EEG, PPG, GSR, Obraz, Dźwięk)	Monitorowanie opieki zdrowotnej, inteligentne domy, interakcja człowiek-komputer (IoT)	94.2% (zbiór MELD)	Hierarchiczna uwaga czasowa (HTA-CMRE), 1D CNN, Dilated Convolutions, Self-attention	Skalibrowane pod urządzenia IoT (Jetson Xavier NX), opóźnienie <200 ms, 8-bitowa kwantyzacja
Attention-Based CNN-BiLSTM-Transformer	Sygnały fizjologiczne (EEG, EKG, GSR, EMG)	Opieka zdrowotna, zarządzanie zdrowiem psychicznym, interakcja człowiek-komputer	88.2% (DEAP), 89.5% (AMIGOS)	CNN (przestrzenna), BiLSTM (czasowa), Transformer (zależności dalekiego zasięgu)	Wysoka precyzja obliczeniowa (FP32/Mixed), framework PyTorch/TensorFlow
MemoCMT	Multimodalne (Obraz, Dźwięk, Tekst)	Rozpoznawanie emocji w konwersacjach	82.3% (zbiór CMU-MOSEI)	Cross-Modal Transformer	Zoptymalizowany pod kątem fuzji cech multimodalnych
JOYFUL	Multimodalne (Dźwięk, Obraz, Tekst)	Social AI, analiza dialogów	81.2% (F1-macro na MELD)	Graph Contrastive Learning + Transformer	Średni rozmiar, fuzja na poziomie cech pośrednich (mid-level)
Wear-BioNet	Sygnały fizjologiczne (Tętno, EDA, Akcelerometr)	Urządzenia noszone (Wearables), śledzenie stresu	84.5% (zbiór WESAD)	Ensemble CNN-GRU (późna fuzja)	Lekki model dedykowany dla urządzeń brzegowych (Edge)
TMNet	Bimodalne (EEG, Mowa)	Analiza afektywna oparta na sygnałach mózgowych i głosie	88.7% (zbiór SEED-IV)	Cross-modal Transformer	Zoptymalizowany pod kątem korelacji EEG-Mowa
Edge-MER	Bimodalne (Twarz, Dźwięk)	Aplikacje czasu rzeczywistego na urządzeniach mobilnych	79.4% (zbiór RAVDESS)	Lekki CNN-LSTM	Bardzo niski pobór mocy, optymalizacja pod opóźnienia
DAF-MHSCA	Obraz (Twarz)	Rozpoznawanie ekspresji twarzy (FER) w rzeczywistych warunkach, medycyna, bezpieczeństwo kierowców	92.49% (RAF-DB), 67.72% (AffectNet-7), 74.33% (FER2013)	ResNet18 backbone + Adaptive Feature Calibration (AFC) + Directional Attention Fusion (DAF) + Multi-Head Spatial-Channel Attention (MHSCA)	16.5 MB / FP16/FP32
FedMultiEmo (Federated CNN)	Multimodalne (Obrazy twarzy + Sygnały fizjologiczne: tętno, EDA, temperatura skóry)	Monitorowanie emocji kierowcy w czasie rzeczywistym, systemy ADAS	87.0% (Fuzja), 77.0% (Wizualny), 74.0% (Fizjologiczny)	CNN (Wizja) + Random Forest (Fizjologia) z fuzją na poziomie decyzji (Majority Voting)	< 200 MB RAM na klienta / INT8/FP16
Raven (Tavus API)	Multimodalne (Wideo, mowa, język ciała)	Konwersacyjny interfejs wideo, obsługa klienta, sprzedaż, marketing, zdrowie psychiczne	Brak danych (SOTA)	Raven perception model (Kontekstualna percepcja w czasie rzeczywistym)	Wysoka (Cloud API/Edge)
MFER	Obraz (Twarz)	Analiza afektywna, rozpoznawanie emocji w dziczy	92.08% (RAF-DB), 67.06% (AffectNet-7), 73.45% (FER2013)	Multiscale Facial Expression Recognition (Dynamic Global & Static Local Attention)	Średni / FP32
DDAMFN	Obraz (Twarz)	Rozpoznawanie emocji w warunkach nieograniczonych	91.35% (RAF-DB), 67.03% (AffectNet-7)	Dual-Directional Attention Mixed Feature Network	Brak danych w źródle
AffectGPT	Wideo, Audio, Tekst (Multimodalne)	Generatywne rozumienie emocji, interpretowalna analiza afektywna	57.36% (S1 na MER-FG)	Multimodal Large Language Model (MLLM) z mechanizmem pre-fuzji	7B parametrów / BF16/INT4
EmotionTFN	Multimodalne	Systemy multimodalne czasu rzeczywistego	Brak danych w źródle	Hierarchical Temporal Attention (HTA)	Średni / FP16
AlexNet Variant	Obraz (AffectNet)	Mobilne/Webowe rozpoznawanie emocji w czasie rzeczywistym	55.09% (AffectNet-8)	Lightweight CNN (AlexNet variant)	Mały / INT8/FP16
Combined CNN-RNN	Wideo	Rozpoznawanie emocji w sekwencjach wideo (Valence/Arousal)	0.45 RMSE / 0.39 CCC (Arousal na Aff-Wild2)	CNN Feature Extractor + GRU (3 warstwy)	Mała liczba parametrów / FP32
AffectNet+	Obraz (wyraz twarzy)	Analiza ekspresji twarzy w systemach edukacyjnych	90.4	Głębokie sieci neuronowe (Deep Neural Networks)	Precyzja zmiennoprzecinkowa (FP32/FP16)
EVI 3 (Hume AI)	Multimodalny (mowa, tekst)	Interfejsy głosowe, robotyka, edukacja	Brak danych w źródle	Speech-to-speech (SLM), integracja z Claude 4/Gemini	Opóźnienie poniżej 300ms, model zoptymalizowany pod kątem przepustowości
EMOE	Multimodalny (obraz, akustyka, język)	Rozpoznawanie emocji (MER) i intencji (MIR)	Brak danych w źródle (SOTA)	Mixture of Modality Experts, Unimodal Distillation	Wymaga akceleracji GPU/NPU dla multimodalnej fuzji
CNN-BiLSTM-Transformer	Sygnały fizjologiczne (EEG, GSR, PPG)	Monitorowanie stanów poznawczych i emocjonalnych	88.2 (DEAP), 89.5 (AMIGOS)	CNN, BiLSTM, Transformer (z mechanizmem Attention)	Przetwarzanie sygnałów w czasie rzeczywistym na urządzeniach wearable
ArcFace (na AffectNet)	Obraz (twarz)	Rozpoznawanie 8 klas emocji	87.6	ArcFace	Wysokie zapotrzebowanie na zasoby obliczeniowe
Poria et al. (2015) Model	Tekst, dźwięk, wideo	Klasyfikacja emocji w AIEd	87.95	Głębokie sieci neuronowe (Deep Neural Networks)	Wysokie wymagania sprzętowe i zapotrzebowanie na dane
AffectGPT	Multimodalny (audio, wideo, tekst)	Złożone rozumienie emocji i generowanie wyjaśnień	Brak danych w źródle	Multimodal Large Language Model (MLLM)	Wymaga dużej pamięci VRAM (oparty na architekturze GPT)
Scaleddown EfficientNet	Obraz (twarz)	Systemy o ograniczonych zasobach	61.3	EfficientNet (skalowana w dół)	Zoptymalizowany pod kątem procesorów mobilnych/edge
ResNet50 (Video CNN)	Wideo	Rozpoznawanie emocji w robotyce społecznej	76	ResNet50	2.5M parametrów
C3D	Wideo	Aplikacje robotyczne (kompaktowy rozmiar)	76	C3D	0.33M parametrów