**Review of AI-based mental health detection models**

**Introduction:**

Artificial intelligence (AI) has advanced significantly in healthcare, particularly in mental health detection, offering accessible and accurate assessments. Unlike traditional, often time-consuming and subjective clinical evaluations, AI models use machine learning, deep learning, and natural language processing to analyse speech, text, physiological data, and behaviour, detecting conditions like depression and anxiety with improved efficiency. This paper reviews AI-based mental health detection models, their methodologies, applications, strengths, limitations, and potential for future advancements.

**Background and theoretical framework:**

Mental health disorders impact millions globally, affecting well-being and quality of life. Early, accurate detection is vital but traditional methods like clinical interviews and tools (e.g., PHQ-9, GAD-7, DSM-5) are often subjective and less accessible. AI in mental health detection draws on:

1. **Machine and Deep Learning**: Using supervised, unsupervised, and reinforcement learning, models like CNNs, RNNs, and Transformers process text, speech, and physiological data.
2. **Natural Language Processing (NLP)**: NLP analyzes linguistic markers in text and speech to detect conditions like depression and anxiety.
3. **Cognitive and Behavioral Psychology**: AI integrates psychological theories to map cognitive-behavioral patterns, enhancing assessment interpretability.
4. **Human-Computer Interaction (HCI)**: HCI principles ensure ethical, user-friendly AI tools, addressing privacy, bias, and explainability. This paper explores AI-based mental health detection models, their methodologies, and effectiveness in early diagnosis and management.

**Model 1: Ellipsis Health** [1]

**Overview**:

Ellipsis Health uses AI to detect depression and anxiety by analyzing speech content and acoustic features. Its technology assesses linguistic and vocal patterns to measure mental health severity, enabling non-intrusive screening via short voice clips.

**Model Type**:

Employs deep learning with NLP and signal processing, using transformer-based architectures to process speech data and evaluate mental health symptoms.

**Applications**:

* Mental health screening in senior populations.
* Continuous monitoring via smartphone apps.

**Study Summary** (Frontiers in Psychology, 2022):  
A study at Desert Oasis Healthcare tested the Ellipsis Health App’s feasibility for weekly voice-based screening in seniors (mean age 63). Participants recorded 5-minute voice samples over 6 weeks, alongside PHQ-8 and GAD-7 questionnaires. The app achieved an AUC of 0.82 for depression and anxiety detection, with 61% protocol completion and high performance across age groups. Transformer models outperformed LSTM, demonstrating scalability for diverse populations.

**Strengths**: Non-invasive, scalable, effective across ages.  
**Limitations**: Relies on user compliance; limited to voice data.

**Model 2:Woebot** [2]

#### **Overview:**

Woebot is an AI-powered mental health chatbot that offers **cognitive behavioral therapy (CBT)** through conversations. It's designed to support people dealing with **depression, anxiety, and stress** using daily mood tracking and supportive conversations.

#### **Model Type:**

Uses **NLP** and **rule-based conversation models**. Trained with **psychological frameworks like CBT**, it can understand user messages and respond with supportive, therapeutic replies.

#### **Applications:**

Daily check-ins and mood tracking.  
Early detection and support for depression and anxiety.  
Personalized mental health support via mobile app.

#### **Study Summary (JMIR, 2017):**

A study of 70 college students showed that those who used Woebot reported a **significant decrease in depression symptoms** after 2 weeks compared to the control group. Users found it engaging and easy to use.

#### **Strengths:**

Easily accessible through smartphones.  
Focuses directly on anxiety and depression.  
Based on clinically proven CBT techniques.

#### **Limitations:**

Not a replacement for therapists.  
Rule-based responses can feel repetitive over time.

**Model 3: MHDeep** [3]

**Overview**: MHDeep detects mental health disorders (e.g., schizoaffective disorder, depression, bipolar disorder) using wearable sensor data, providing continuous, passive monitoring without user input.

**Model Type**: Combines recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer models to analyze time-series physiological data (e.g., heart rate, movement) from wearables.

**Applications**:

* Early detection of mental health disorders.
* Remote monitoring via smartwatches and smartphones.

**Study Summary** (arXiv, 2021):  
MHDeep was tested on 74 individuals using smartwatch sensors to collect physiological data for 1.5 hours. Using grow-and-prune DNN synthesis and synthetic data, it achieved up to 100% patient-level accuracy for schizoaffective and depressive disorders, with 82.4%-90.4% instance-level accuracy. The system is computationally efficient, suitable for real-time monitoring.

**Strengths**: High accuracy, non-invasive, continuous monitoring.  
**Limitations**: Limited to physiological data; requires wearable devices.

**Model 4: Proactive Emotion Tracker** [4]

**Overview**: The Proactive Emotion Tracker monitors emotions in real-time using text, speech, and physiological data from social media, messages, and wearables, detecting mood disorders and stress.

**Model Type**: Integrates pre-trained transformer models (e.g., BERT), CNNs, RNNs, and sensor-based machine learning to analyze multimodal data for emotional assessment.

**Applications**:

* Real-time mood detection.
* Long-term mental health tracking.

**Study Summary** (arXiv, 2024):  
The tracker used a modified BERT model and wearable sensor data (EEG, smartwatches) to detect depression from social media and physiological signals. It achieved 93% accuracy in classifying depressive text. A browser extension analyzed search history, demonstrating real-time mood tracking potential.

**Strengths**: Multimodal, high accuracy, real-time insights.  
**Limitations**: Privacy concerns; relies on user-generated data.

**Model 5: Speech-Based Mental Health Models**

**Overview**: These models analyze speech patterns to detect depression, anxiety, and bipolar disorder, using linguistic and paralinguistic features for non-invasive assessments in telehealth and remote settings.

**Model Type**: Employs NLP (BERT, RoBERTa), acoustic feature extraction (CNNs, RNNs), and pre-trained speech models (Wav2Vec, Whisper) for multimodal speech analysis, with explainable AI for interpretability.

**Applications**:

* Early diagnosis of mental health conditions.
* Telemedicine and workplace well-being monitoring.

**Study Summary** (arXiv, 2024): [5]  
Using French (Callyope-GP) and Italian (Androids) datasets, the study tested HuBERT, wav2vec2, and Whisper models. HuBERT-L achieved a 0.92 F1 score for depression detection in spontaneous speech, with 5-20 second audio segments performing best. Spontaneous speech outperformed elicited speech for mental health markers.

**Strengths**: Non-invasive, scalable, high accuracy in spontaneous speech.  
**Limitations**: Language-specific models; variable performance with short audio.

**Model 6: Explainable AI for Mental Disorder Detection via Social Media**  [6]

**Overview**: This system detects mental disorders (e.g., depression, PTSD) from social media posts, prioritizing explainability to ensure transparent predictions for clinicians and users.

**Model Type**: Uses transformer-based NLP (BERT, RoBERTa), graph neural networks (GNNs), and explainable AI (SHAP, LIME) to analyze text, social interactions, and sentiment.

**Applications**:

* Early detection of mental health risks.
* Suicide prevention and psychiatric research.

**Study Summary** (arXiv, 2024):  
A survey reviewed XAI models for mental disorder detection using social media data (Reddit, Twitter). Transformer-based models with SHAP and LIME achieved high accuracy in classifying mental health signals, with GNNs improving detection via social network analysis. Challenges include privacy and generalizability across demographics.

**Strengths**: Transparent, interpretable, leverages social data.  
**Limitations**: Privacy risks; model generalizability issues.

**Comparison Matrix**

| **Model** | **Primary Data Source** | **AI Techniques** | **Primary Applications** | **Reported Accuracy** |
| --- | --- | --- | --- | --- |
| **Ellipsis Health** | **Speech (voice clips)** | **NLP, Deep Learning (Transformers)** | **Depression/anxiety screening, telemedicine** | **AUC: 0.82** |
| **Speech-Based Mental Health** | **Speech (linguistic, acoustic)** | **NLP, CNNs, RNNs, Pre-trained (HuBERT)** | **Early diagnosis, remote monitoring** | **F1-score: 0.92** |
| **MHDeep** | **Wearable sensors (physiological)** | **RNNs, CNNs, Transformers** | **Continuous monitoring, early detection** | **82.4%–100%** |
| **Proactive Emotion Tracker** | **Text, speech, wearables** | **BERT, CNNs, RNNs, Sensor ML** | **Real-time mood tracking, depression detection** | **93% (text)** |
| **Explainable AI (Social Media)** | **Social media text** | **BERT, GNNs, SHAP/LIME** | **Early detection, suicide prevention** | **High (unspecified)** |

**Challenges for AI is psychiatry** [7][8][9]

Artificial intelligence (AI) holds promise for enhancing psychiatric diagnosis and treatment, but significant challenges must be addressed to ensure its reliability and ethical use.

* **Data Quality and Privacy**: AI requires high-quality data, but psychiatric data, often based on unreliable self-reports, is limited by privacy concerns restricting collection and sharing.
* **Model Interpretability**: Complex AI models lack transparency, reducing trust among clinicians and patients who need clear reasoning for recommendations.
* **Ethical and Legal Accountability**: Unclear regulations and uncertainty about responsibility for AI errors raise ethical concerns, particularly regarding patient data usage.
* **Bias and Fairness**: AI trained on non-diverse datasets may produce biased outcomes, leading to unfair treatment for minority groups.
* **Clinical Integration**: Many clinicians lack AI training, and reluctance to adopt new technology hinders effective use in practice.
* **Human Connection and Trust**: AI lacks emotional understanding, critical for therapy, and patients may distrust sharing sensitive information with non-human systems.
* **Crisis Handling and Reliability**: AI may fail in emergencies (e.g., suicidal crises) and lacks proven reliability in real-world psychiatric settings.

**Conclusion**

AI can support psychiatry but faces barriers in data quality, interpretability, ethics, and trust. Improved data, clear regulations, and clinician training are essential to ensure AI complements, rather than replaces, human care.

References

[1] <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.811517/full>

[2] https://mental.jmir.org/2017/2/e19/

[3] <https://arxiv.org/abs/2102.10435>

[4] <https://arxiv.org/pdf/2401.13722>

[5] <https://arxiv.org/pdf/2409.19042>

[6] <https://arxiv.org/pdf/2406.05984>

[7] <https://academic.oup.com/ijnp/article/28/Supplement_1/i213/8009656?login=false>

[8] <https://annals-general-psychiatry.biomedcentral.com/articles/10.1186/s12991-023-00476-9>

[9] <https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/on-the-role-of-artificial-intelligence-in-psychiatry/C01B76C74A10CD116786D4600A3D679F>