



Article

Learning Acoustic Biomarkers in Depression Speech Using a Fractional Kolmogorov–Arnold Network

Junkang Yang ¹ , Yuchen Lu ² and Yuxuan Zhang ^{3,4,*}

¹ Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi, Kofu 400-8510, Japan; g25dtsa5@yamanashi.ac.jp

² Yantai Research Institute, Harbin Engineering University, Yantai 265615, China; luyuchen6@foxmail.com

³ College of Intelligent Science and Engineering, Beijing University of Agriculture, Beijing 102206, China

⁴ Department of Computer and Electrical Engineering, Mid Sweden University, 85170 Sundsvall, Sweden

* Correspondence: yuxuan.zhang@miun.se

Abstract

Speech-based depression detection provides a non-invasive and low-cost approach for supporting mental health screening. However, existing deep models often learn acoustic representations through fixed nonlinear transformations, which limits their ability to characterise gradual, nonlocal, and depression-related time–frequency variations in speech. To address this issue, this paper proposes a Fractional Kolmogorov–Arnold Network (FKAN) for learning acoustic biomarkers from mel-spectrograms. The proposed FKAN introduces fractional coordinate encoding to describe non-integer-order temporal–frequency responses, fractional-gated RBF-KAN blocks to learn adaptive edge-wise functional mappings, and a prototype-based classifier to construct a depression-related acoustic state space. In this way, the model not only improves classification performance but also provides functional and geometric evidence for interpreting acoustic biomarkers. Experiments on two public depression speech datasets demonstrate the effectiveness of the proposed method. The FKAN achieves an accuracy of 0.95 and an average F1-score of 0.94 on DAIC-WOZ, and an accuracy of 0.94 and an average F1-score of 0.94 on MODMA. Ablation studies, parameter sensitivity analysis, cross-dataset evaluation, and visualisation results further confirm that the fractional components and prototype-based state representation contribute to both robust depression detection and interpretable acoustic pattern analysis.

Keywords: depression detection; acoustic biomarkers; fractional modeling; fractional coordinate encoding; Kolmogorov–Arnold networks; speech signal processing



Academic Editor: António Lopes

Received: 15 June 2026

Revised: 6 July 2026

Accepted: 9 July 2026

Published: 11 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Depression is one of the most common mental health disorders and has become a major public health concern worldwide. It affects mood, cognition, social behaviour, and daily functioning, and can lead to long-term disability when not detected and treated in time [1]. The World Health Organisation reported that approximately 332 million people live with depression, and that depression is closely related to suicide risk and impaired quality of life [2]. In clinical practice, depression assessment still relies heavily on interviews, self-report scales, and expert judgment. Although these methods are clinically useful, they can be affected by subjective bias [3], delayed access to care, and the limited availability of trained professionals. Therefore, objective, low-cost, and easily deployable tools are increasingly needed to support early screening and continuous mental health monitoring.

Speech provides a promising way to meet this need. It is natural, non-invasive, and can be collected with common devices. More importantly, depression often changes how people speak. Depressed speech may show reduced energy, slower rhythm, longer pauses, flatter intonation, and altered voice quality [4]. These changes are not always easy for human listeners to perceive, but they can be captured through acoustic and time–frequency analysis. Recent studies have shown that speech and voice features can serve as potential digital or acoustic biomarkers for depression detection [5], although their clinical use still requires stronger evidence and better generalisation. This makes speech-based depression detection an important research direction at the intersection of affective computing, speech signal processing, mental health assessment, and interpretable machine learning.

Existing studies have explored various speech representations for depression detection, including handcrafted acoustic descriptors [6], spectrogram-based features [7], and deep representations learned by convolutional, recurrent, attention-based, and self-supervised models [8]. These methods have improved models' ability to detect depression-related speech patterns. Some works also use multimodal information [5], such as text, facial behaviour, and physiological signals, to further enhance performance. However, speech-only depression detection remains highly valuable because speech is easier to collect and less intrusive in many real-world scenarios. For this reason, improving the modelling ability and interpretability of acoustic representations is still a central problem in this field.

Despite recent progress, several limitations remain. First, many deep models learn powerful representations, but their decisions are often difficult to explain. This is a serious concern in depression detection, where a model should not only provide a prediction but also offer clues about the acoustic evidence supporting it. Second, most existing networks rely on fixed nonlinear transformations or standard temporal–frequency modelling operations. Such designs may not fully describe the slow, long-range, and non-integer-order changes that often appear in depressed speech, such as low energy, reduced prosodic variation, and delayed spectral dynamics. These acoustic changes are usually distributed across time–frequency regions rather than being isolated local events. Therefore, a fractional modelling perspective may be useful for characterising global and non-integer-order acoustic variations in depressive speech. Third, many models treat classification as a direct mapping from features to labels, without explicitly defining a state space in which depression-related acoustic biomarkers can be observed and analysed. As a result, the learned features may be effective for prediction but less useful for understanding how depression is reflected in speech.

To address these issues, this paper proposes a Fractional Kolmogorov–Arnold Network (FKAN) for acoustic biomarker learning in speech-based depression detection. The FKAN takes mel-spectrograms as input and introduces a fractional-coordinate functional encoding module to model non-integer-order temporal–frequency responses in speech, aiming to capture longer-term acoustic changes associated with depression. Based on this encoding, a fractional-gated edge-wise KAN block is designed to replace conventional node-wise nonlinear transformations with learnable edge functions, enabling the model to learn more flexible and more transparent functional mappings. In addition, an affective-depressive state prototype head maps each speech sample into a compact state space and performs classification by comparing it with learnable class prototypes. This design links model prediction with acoustic biomarker interpretation, as the prototype distances, state distributions, edge-function importance, and fractional-order responses can be further analysed to explain depression-related speech patterns.

Based on the above design, this work makes the following contributions:

- We propose a unified FKAN framework for acoustic biomarker learning in speech-based depression detection, offering a fractional functional modelling perspective on depression-related speech representations.
- We design a fractional-coordinate encoding module and a fractional-gated KAN block to capture slow and non-integer-order depression-related speech changes using adaptive edge-wise functions rather than fixed activation patterns.
- We introduce a prototype-based depression-state classifier that maps speech samples into an interpretable acoustic state space, thereby improving both detection performance and model transparency.

The remainder of this paper is organised as follows. Section 2 reviews related work on speech-based depression detection. Section 3 describes the proposed model in detail. Section 4 presents the experimental setup, datasets, implementation details, and evaluation metrics. Section 5 reports the main results, comparison with baseline methods, ablation studies, parameter sensitivity analysis, and visualisation results. Section 6 concludes the findings and potential directions for future work.

2. Related Work

Speech-based depression detection has attracted growing attention because speech can reflect changes in emotion, cognition, and psychomotor activity. Early studies mainly relied on hand-crafted acoustic features. Common features included pitch [9], energy [10], spectral descriptors [11], cepstral coefficients [12], jitter, shimmer, harmonic-to-noise ratio, and so on [13]. These features were often combined with traditional classifiers such as support vector machines [14], random forests [15], and logistic regression models. Such methods offered a clear link between acoustic features and depression-related speech symptoms, which made them relatively easy to interpret. However, their performance was strongly affected by feature design, recording conditions, language differences, and speaker variability. They also had limited ability to model complex temporal–frequency patterns in continuous speech.

Recent advances in deep learning have significantly improved automated disease detection and medical decision support across different clinical domains [16–18]. So more studies began to use neural networks to learn depression-related speech representations directly from spectrograms or frame-level acoustic sequences. Convolutional networks have been used to extract local spectral patterns [19], while recurrent networks and temporal models have been applied to capture changes over time [20]. Attention mechanisms further improved the ability of models to focus on informative speech segments, especially when long recordings contained silence, neutral content, or irrelevant acoustic events [21]. These methods reduced reliance on manual feature engineering and improved detection performance across many settings. However, they often require careful training and can still overfit when the available clinical data are limited.

Recent work has also explored transfer learning [22] and speech pre-trained models. Models trained on large-scale speech data can provide strong acoustic representations, and they are useful when depression datasets are small. Features extracted from wav2vec-style models [23], audio spectrogram transformers [24], and other foundation models have been used for depression classification and severity estimation. These approaches can improve robustness and reduce the burden of training a model from scratch. At the same time, their large model size and complex internal structure make them harder to interpret. Another important research direction is acoustic biomarker analysis [25,26]. Instead of treating depression detection as a pure classification task, these studies aim to identify speech patterns that are related to depressive states. Some works have used statistical analysis or explainable learning methods to connect model predictions with acoustic indicators. This

direction is valuable because it can help bridge the gap between automatic detection and clinical interpretation.

Although current speech-based methods have achieved promising results, several challenges remain. Depressed speech often shows slowly varying changes rather than just short, local events. Standard neural networks usually model these patterns through convolution, recurrence, or attention, but these operations may not fully describe temporal–frequency memory in speech. In addition, many models use fixed activation functions or ordinary feed-forward layers, which limit the transparency of the learned nonlinear mapping. The final classifier is also commonly implemented as a fully connected layer, making it difficult to explain how a sample is related to depression or non-depression states. These limitations suggest that a more interpretable functional representation may be useful for speech-based depression detection. In recent years, Kolmogorov–Arnold Networks (KANs) [27] have provided a new direction for building neural models with learnable functional transformations. Unlike conventional multilayer perceptrons, which use fixed activation functions at nodes, KANs place learnable one-dimensional functions on edges. This design offers a more flexible way to model nonlinear relations and gives a natural basis for visualising the learned functions. Because of these properties, KANs have attracted attention in various machine learning tasks [28,29]. However, their use in speech-based depression detection is still limited. Existing KAN-based studies are also rarely designed around acoustic biomarkers or depression-related speech dynamics.

The above studies show that speech-based depression detection has moved from hand-crafted acoustic features to deep representation learning and pre-trained speech models. This progress has improved prediction ability, but interpretability and depression-specific acoustic modelling remain open problems. A model that can learn flexible functional mappings, capture slow temporal–frequency changes, and map speech into a depression-related state space would be valuable for both performance and analysis. This motivates the proposed FKAN framework, which combines fractional encoding, edge-wise KAN representation, and prototype-based acoustic biomarker learning for speech-based depression detection.

3. Methods

3.1. Overall Architecture of the FKAN

As shown in Figure 1, the proposed FKAN is designed for speech-based depression detection using mel-spectrograms as input. Given a mel-spectrogram, the FKAN learns depression-related acoustic patterns via a compact processing pipeline. The architecture consists of a convolutional frontend, a fractional-coordinate encoding module, a learnable class token, a stack of KAN blocks, and a prototype-based classifier. The model predicts whether a speech sample belongs to the major depressive disorder (MDD) group or the healthy control (HC) group.

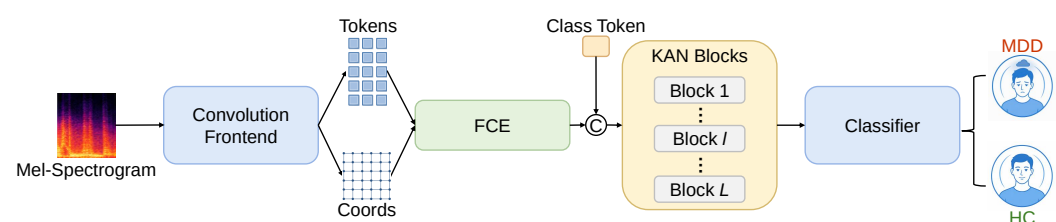


Figure 1. Architecture of the proposed FKAN.

The input mel-spectrogram is first processed by the convolution frontend to obtain a compact acoustic representation. Instead of applying KAN operations directly to the original spectrogram, this stage extracts local time–frequency patterns and converts them

into a sequence of acoustic tokens. Each token represents a local speech region and carries short-term spectral information. This design reduces computational cost and provides a stable token representation for subsequent functional modelling stages.

To preserve the time–frequency structure of the token sequence, the FKAN also builds coordinate information for the corresponding token locations. These coordinates describe the location of each token in the mel-spectrogram and provide the positional basis for fractional-coordinate encoding. In contrast to ordinary position labels, the coordinates are used to generate fractional time–frequency responses. This allows the model to describe how depression-related acoustic patterns vary over time and frequency, which is important because depressed speech often shows persistent changes, such as reduced energy variation, flatter prosody, and slower temporal dynamics.

The acoustic tokens and their coordinate information are then fed into the fractional coordinate encoding (FCE) module. FCE integrates local acoustic content with coordinate-dependent fractional responses so that each token contains both spectral information and its time–frequency functional context. Through this design, the FKAN not only models what acoustic pattern appears in a local region, but also how this pattern is organised within the whole spectrogram. The FCE output serves as depression-sensitive token representation for the subsequent KAN blocks.

After FCE, a learnable class token is concatenated with the encoded token sequence. This class token acts as a global representation carrier and interacts with the acoustic tokens during KAN-based transformation. As the sequence passes through the stacked KAN blocks, the class token gradually aggregates depression-related information from the whole utterance. This design avoids using simple global pooling as the only aggregation strategy and allows utterance-level representation to be learned through the same functional transformation process as the acoustic tokens.

The KAN blocks form the core of the FKAN’s representation learning stage. Each block uses learnable edge-wise functions to transform token features, rather than relying only on fixed activation functions. This structure allows the model to learn flexible nonlinear mappings for depression-related speech cues while keeping the transformation process more interpretable. By stacking multiple KAN blocks, the FKAN progressively refines both the acoustic tokens and the class token, enabling the model to capture local acoustic variations and higher-level depression-related patterns.

The refined class token is finally sent to the prototype-based classifier. The classifier maps it into a compact depression-related state space and compares it with learnable class prototypes. In the binary depression detection task, the two prototypes correspond to MDD and HC. The prediction is based on the distance or similarity between the speech representation and these prototypes. Compared with a standard fully connected classifier, this design provides a clearer link between the model output and the learned acoustic biomarker space.

3.2. Convolution Frontend

As shown in Figure 2, the convolution frontend converts the input mel-spectrogram into a compact sequence of acoustic tokens. Given an input spectrogram $\mathbf{X} \in \mathbb{R}^{B \times 1 \times T \times F}$, where B is the batch size, T is the number of temporal frames, and F is the number of mel-frequency bins, the frontend applies a series of lightweight convolutional blocks to extract local time–frequency patterns. Each block consists of a convolution layer, batch normalisation, and a SiLU activation. The channel dimensions are gradually increased from 1 to 192, while the time–frequency resolution is reduced by strided convolutions. A final 1×1 convolution projects the feature map to the token dimension D .

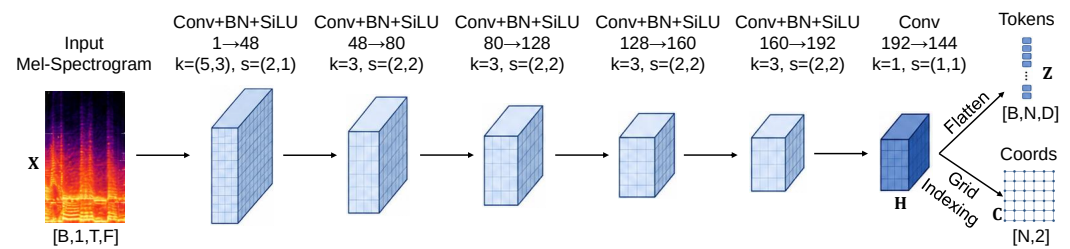


Figure 2. The architecture of the convolution frontend.

This frontend is used only as a standard acoustic tokeniser. It is not treated as the main methodological contribution of the FKAN. Its role is to provide stable local representations before the fractional functional modules and KAN blocks. In this work, the input mel-spectrogram has the shape $[B, 1, 500, 64]$. After the convolution frontend, the feature map is reduced to a compact grid with spatial size $T_g \times F_g$, where $T_g = 16$ and $F_g = 4$ in the default setting. The output feature map can be written as

$$\mathbf{H} = f_{\text{conv}}(\mathbf{X}), \quad \mathbf{H} \in \mathbb{R}^{B \times D \times T_g \times F_g}, \quad (1)$$

where $f_{\text{conv}}(\cdot)$ denotes the convolution frontend and $D = 144$ is the token dimension. The feature map is then flattened along the time–frequency grid to form a token sequence. For the grid location (i, j) , where $i \in 0, \dots, T_g - 1$ and $j \in 0, \dots, F_g - 1$, the corresponding token is obtained by

$$\mathbf{z}_{i,j} = \mathbf{H}_{:, :, i, j} \in \mathbb{R}^{B \times D}. \quad (2)$$

By using row-wise indexing, the two-dimensional grid location (i, j) is mapped to a one-dimensional token index

$$n = iF_g + j, \quad n \in 0, \dots, N - 1, \quad (3)$$

where $N = T_g F_g$ is the total number of tokens. The final token sequence is represented as

$$\mathbf{Z} = [\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_{N-1}] \in \mathbb{R}^{B \times N \times D}. \quad (4)$$

For the default input size, $N = 16 \times 4 = 64$, so the convolution frontend produces 64 acoustic tokens for each speech sample.

In addition to the token sequence, the FKAN builds a coordinate matrix to preserve the time–frequency location of each token. The coordinates of each grid point are normalised to the range $[0, 1]$ along both the time and frequency axes. For the grid location (i, j) , the normalised temporal and frequency coordinates are defined as

$$\tau_i = \frac{i}{T_g - 1}, \quad \phi_j = \frac{j}{F_g - 1}. \quad (5)$$

The coordinate vector of the corresponding token is then given by

$$\mathbf{c}_{i,j} = \begin{bmatrix} \tau_i \\ \phi_j \end{bmatrix} \quad (6)$$

Following the same row-wise indexing rule, the coordinate matrix is constructed as

$$\mathbf{C} = [\mathbf{c}_0, \mathbf{c}_1, \dots, \mathbf{c}_{N-1}]^\top \in \mathbb{R}^{N \times 2}. \quad (7)$$

Each row of $\mathbf{C}$ stores the normalised time–frequency position of one acoustic token. The coordinate matrix is not used as a simple positional label. It provides the basis for the following fractional-coordinate encoding module, in which the model learns coordinate-dependent functional responses. In this way, the frontend produces two aligned outputs: the acoustic token sequence $\mathbf{Z}$ and the coordinate matrix $\mathbf{C}$. The former describes local spectral content, while the latter preserves the locations of each local pattern in the mel-spectrogram. This design allows the following FKAN modules to model both acoustic content and its time–frequency organisation.

3.3. Fractional Coordinate Encoding

The convolution frontend produces an acoustic token sequence $\mathbf{Z} \in \mathbb{R}^{B \times N \times D}$ and a coordinate matrix $\mathbf{C} \in \mathbb{R}^{N \times 2}$. Each row of $\mathbf{C}$ stores the normalised time–frequency position of a token. A direct use of these coordinates as ordinary positional labels may be insufficient for depression detection because depression-related speech changes often appear as slow and persistent variations rather than isolated local events. For example, reduced energy fluctuation, flatter prosody, and slower temporal dynamics may extend across a relatively long region of an utterance. To describe such gradual changes, FKAN introduces a fractional coordinate encoding (FCE) module, as shown in Figure 3.

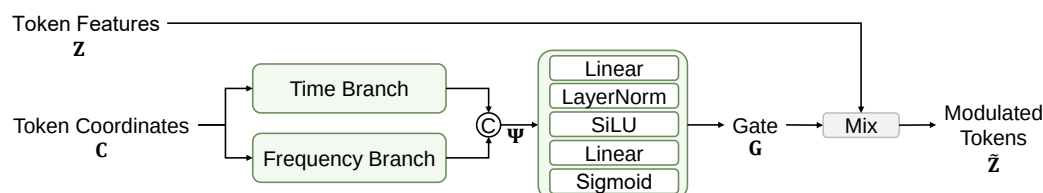


Figure 3. The architecture of the proposed fractional coordinate encoding module.

Let the coordinate of the n -th token be denoted as

$$\mathbf{c}_n = (\tau_n, \phi_n), \quad (8)$$

where $\tau_n \in [0, 1]$ and $\phi_n \in [0, 1]$ indicate the normalized temporal and frequency positions, respectively. FCE constructs coordinate-dependent fractional responses along the time and frequency axes. Different from fixed sinusoidal position encoding, the fractional orders are learnable and are constrained to a small range. This allows the model to adapt the strength of coordinate responses while avoiding an overly dominant positional effect.

For the time branch, FCE uses a set of learnable fractional orders

$$\boldsymbol{\alpha}^t = [\alpha_1^t, \alpha_2^t, \dots, \alpha_{K_c}^t], \quad (9)$$

where K_c is the number of fractional bases. Each order is constrained by

$$\alpha_k^t = \alpha_{\max} \cdot \sigma(\alpha_k^t), \quad k = 1, \dots, K_c, \quad (10)$$

where α_k^t is a learnable parameter, $\sigma(\cdot)$ is the sigmoid function, and $\alpha_{\max}$ controls the maximum fractional order. In the same way, the frequency branch uses another set of learnable orders

$$\boldsymbol{\alpha}^f = [\alpha_1^f, \alpha_2^f, \dots, \alpha_{K_c}^f], \quad k = 1, \dots, K_c. \quad (11)$$

Based on these learnable orders, the time branch generates two directional fractional coordinate responses:

$$\psi_{n,k}^{t,+} = (\tau_n + \epsilon)^{\alpha_k^t}, \quad \psi_{n,k}^{t,-} = (1 - \tau_n + \epsilon)^{\alpha_k^t}, \quad (12)$$

where $\epsilon = 10^{-8}$ is a small constant for numerical stability. The forward response $\psi_{n,k}^{t,+}$ describes the position from the beginning of the utterance, while the backward response $\psi_{n,k}^{t,-}$ describes the position from the end. This bidirectional design allows the model to encode both early and late temporal patterns. The frequency branch is constructed in a similar way:

$$\psi_{n,k}^{f,+} = (\phi_n + \epsilon)^{\alpha_k^f}, \quad \psi_{n,k}^{f,-} = (1 - \phi_n + \epsilon)^{\alpha_k^f}. \quad (13)$$

Here, $\psi_{n,k}^{f,+}$ and $\psi_{n,k}^{f,-}$ describe the coordinate responses from the low-frequency and high-frequency sides, respectively. These responses are useful because depression-related speech cues may appear differently across frequency regions, such as changes in spectral balance or voice quality.

The outputs of the time and frequency branches are concatenated to form the fractional coordinate basis:

$$\psi_n = [\psi_n^{t,+}, \psi_n^{t,-}, \psi_n^{f,+}, \psi_n^{f,-}] \in \mathbb{R}^{4K_c}. \quad (14)$$

For all tokens, the basis matrix is denoted as

$$\Psi = [\psi_1, \psi_2, \dots, \psi_N]^T \in \mathbb{R}^{N \times 4K_c}. \quad (15)$$

The fractional basis is then transformed into a token-wise gate through lightweight nonlinear mapping. As shown in Figure 3, this nonlinear layer contains a linear layer, layer norm, SiLU and Sigmoid activations. The resulting gate, $\mathbf{G} \in \mathbb{R}^{N \times D}$, is shared across samples in the batch because it is generated from token coordinates. The gate modulates the acoustic token sequence through a residual scaling operation:

$$\tilde{\mathbf{Z}} = \mathbf{Z} \odot [1 + \tanh(s)(\mathbf{G} - 0.5)], \quad (16)$$

where $\odot$ denotes element-wise multiplication and s is a learnable strength parameter. The term $\tanh(s)$ controls the modulation range. Since s is initialised with a small value, FCE only introduces a mild fractional coordinate effect at the beginning of training. The model can then learn how much coordinate-dependent modulation is needed.

Through this design, FCE does not replace acoustic token features with positional information. Instead, it softly adjusts the token features according to their fractional time–frequency responses. The output $\tilde{\mathbf{Z}} \in \mathbb{R}^{B \times N \times D}$ keeps the same shape as the input token sequence, so it can be directly concatenated with the class token and sent to the following KAN blocks. The fractional responses describe how token features are modulated at different time–frequency locations, and may reflect gradual patterns such as reduced energy variation or slower temporal change. The motivation of FCE differs from that of conventional positional encoding. Standard positional encoding provides a fixed or learnable representation of token location, but its relation to acoustic content is usually indirect. In FCE, each coordinate response follows a learnable power-law form, and the resulting gate directly modulates the acoustic token features. Therefore, the same position can have different effects across feature channels.

3.4. KAN Block

After fractional coordinate encoding, the encoded acoustic tokens are concatenated with a learnable class token and sent to a stack of KAN blocks. The purpose of these blocks is to refine the token sequence through both token-level interactions and feature-wise functional transformations. As shown in Figure 4, each KAN block contains two residual branches. The first branch uses multi-head self-attention to exchange information among tokens. The second branch uses a fractional-gated RBF-KAN layer to learn edge-wise nonlinear functions. The attention branch helps the class token collect utterance-level

information, while the KAN branch improves the functional modelling ability of each token feature.

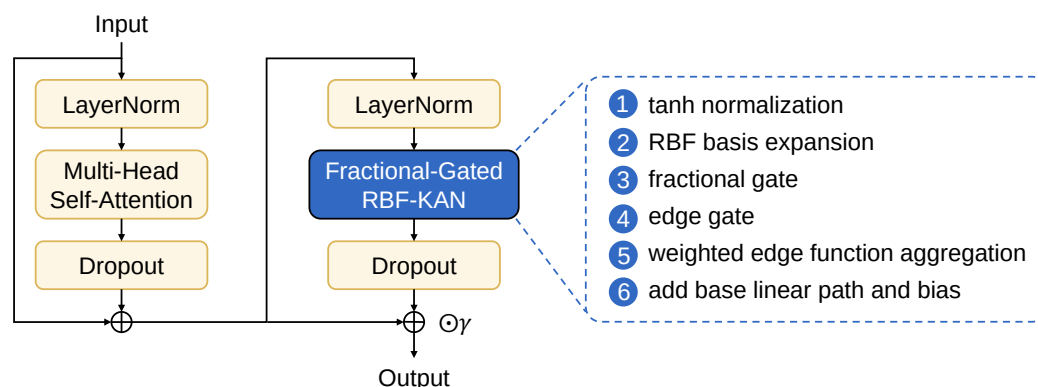


Figure 4. The architecture of the proposed KAN block.

Let the input of the l -th KAN block be

$$\mathbf{S}^{l-1} \in \mathbb{R}^{B \times (N+1) \times D}, \quad (17)$$

where N is the number of acoustic tokens, and the additional token is the class token. The self-attention branch is written as

$$\mathbf{U}^l = \mathbf{S}^{l-1} + \text{Dropout}\left(\text{MSA}\left(\text{LN}(\mathbf{S}^{l-1})\right)\right), \quad (18)$$

where $\text{LN}(\cdot)$ denotes layer normalization and $\text{MSA}(\cdot)$ denotes multi-head self-attention. This residual design keeps the original token information while allowing each token to receive context from other regions of the spectrogram.

The output of the attention branch is then processed by the fractional-gated RBF-KAN layer:

$$\mathbf{V}^l = \mathcal{K}\left(\text{LN}(\mathbf{U}^l)\right), \quad (19)$$

where $\mathcal{K}(\cdot)$ denotes the proposed fractional-gated RBF-KAN transformation. The block output is obtained by a scaled residual connection:

$$\mathbf{S}^l = \mathbf{U}^l + \text{Dropout}\left(\mathbf{V}^l \odot \gamma\right), \quad (20)$$

where $\gamma \in \mathbb{R}^D$ is a learnable channel-wise scaling vector. This parameter controls the contribution of the KAN branch and helps stabilise training.

The key component of the block is the fractional-gated RBF-KAN layer. Unlike a conventional linear layer followed by a fixed activation function, this layer learns a one-dimensional function on each edge between an input feature and an output feature. Let $\mathbf{x} \in \mathbb{R}^I$ be one token feature after layer normalisation, where I is the input dimension. The feature is first normalised by

$$\tilde{x}_i = \tanh(x_i), \quad i = 1, \dots, I. \quad (21)$$

This operation restricts the input to radial basis functions to a stable range and reduces the influence of extreme feature values.

For each input dimension, the layer expands $\bar{x}_i$ with K_r radial basis functions. Given that RBF centres $\{\mu_k\}_{k=1}^{K_r}$ and width β , the k -th basis response is

$$r_{i,k} = \exp\left(-\frac{(\bar{x}_i - \mu_k)^2}{2\beta^2}\right), \quad k = 1, \dots, K_r. \quad (22)$$

Here, $\beta = \text{softplus}(\omega) + \beta_0$, where ω is a learnable parameter and $\beta_0 = 0.4$. The RBF centres are fixed and uniformly sampled from a predefined interval $[\mu_{\min}, \mu_{\max}]$:

$$\mu_k = \mu_{\min} + \frac{k-1}{K_r-1}(\mu_{\max} - \mu_{\min}), \quad k = 1, \dots, K_r. \quad (23)$$

In this paper we set $\mu_{\min} = -2, \mu_{\max} = 2$.

To introduce fractional sensitivity into the edge transformation, the layer computes an input-wise fractional gate. For each input dimension, the fractional order is defined as

$$\rho_i = \rho_{\max} \cdot \sigma(q_i), \quad (24)$$

where q_i is a learnable parameter, $\sigma(\cdot)$ is the sigmoid function, and $\rho_{\max} = 0.5$ is the maximum fractional order. The fractional gate is then given by

$$m_i = (|\tanh(x_i)| + \epsilon)^{\rho_i}, \quad (25)$$

where $\epsilon = 10^{-8}$ is a small constant for numerical stability. This gate modulates the strength of RBF responses according to the magnitude of each normalised latent feature. It does not correspond to a single predefined clinical variable. Instead, it allows different feature dimensions to exhibit different nonlinear response strengths. Since these features are learned from time–frequency speech patterns, the gate may improve sensitivity to gradual depression-related characteristics. Edge importance shows which learned feature relations contribute more strongly to the decision, but does not correspond to a single predefined acoustic variable.

In addition to the fractional gate, the layer uses a learnable edge gate to control the importance of each input–output connection:

$$e_{i,o} = \sigma(g_{i,o}), \quad (26)$$

where $g_{i,o}$ is a learnable parameter for the edge from input dimension i to output dimension o . This gate makes the edge functions more selective and also provides a direct way to analyse edge importance.

The edge-wise function from input dimension i to output dimension o is represented by a weighted combination of RBF bases:

$$\varphi_{i,o}(x_i) = e_{i,o} \sum_{k=1}^{K_r} w_{i,o,k} m_i r_{i,k}, \quad (27)$$

where $w_{i,o,k}$ is the learnable coefficient of the k -th RBF basis. The output feature is obtained by aggregating all edge functions:

$$y_o = \sum_{i=1}^I \varphi_{i,o}(x_i) + \mathbf{a}_o^\top \mathbf{x} + b_o, \quad o = 1, \dots, O. \quad (28)$$

Here, $\mathbf{a}_o^\top \mathbf{x}$ is a base linear path and b_o is a bias term. The base linear path improves optimisation stability, while the RBF-KAN part learns flexible nonlinear edge functions.

For the whole token sequence, the same fractional-gated RBF-KAN layer is applied to each token feature.

3.5. Classifier

After the stacked KAN blocks, the FKAN uses the refined class token as utterance-level representation for depression detection. Instead of sending this representation directly to a standard fully connected classifier, the proposed classifier maps it into a compact depression-related state space and compares it with learnable class prototypes. This design makes the decision process more interpretable because each prediction can be linked to the distance between a speech sample and the class prototypes.

As Figure 5 shows, let the final class-token representation be denoted as $\mathbf{h} \in \mathbb{R}^D$, where D is the token dimension. The classifier first transforms the final class-token representation $\mathbf{h} \in \mathbb{R}^D$ with a fractional-gated RBF-KAN layer. The transformed feature is then passed through a SiLU activation, layer normalisation, and dropout to obtain a compact state vector $\mathbf{s} \in \mathbb{R}^{d_s}$, where $d_s = 96$. This vector is regarded as the depression-related acoustic state of the input speech sample. Compared with directly using $\mathbf{h}$ for classification, this step projects the utterance-level representation into a lower-dimensional state space, where its relationship to the class prototypes is more clearly measurable.

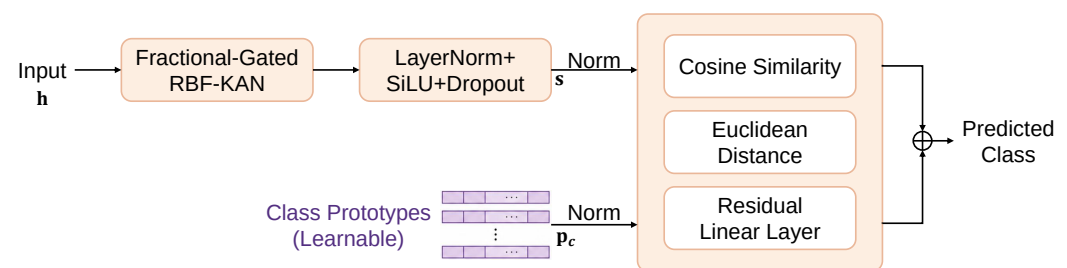


Figure 5. The architecture of the proposed classifier module.

The classifier maintains a set of learnable class prototypes:

$$\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_C]^\top \in \mathbb{R}^{C \times d_s}, \quad (29)$$

where C is the number of classes and $\mathbf{p}_c \in \mathbb{R}^{d_s}$ is the prototype of the c -th class. In binary depression detection, the two prototypes correspond to MDD and HC. To compare the sample state with the class prototypes, both the state vector and the prototypes are normalised as

$$\hat{\mathbf{s}} = \frac{\mathbf{s}}{\|\mathbf{s}\|_2}, \quad \hat{\mathbf{p}}_c = \frac{\mathbf{p}_c}{\|\mathbf{p}_c\|_2}. \quad (30)$$

The cosine similarity between the sample state and the c -th prototype is computed by

$$l_c^{\cos} = \eta, \hat{\mathbf{s}}^\top \hat{\mathbf{p}}_c, \quad (31)$$

where η is a learnable scale parameter. This term measures how close the sample direction is to each class prototype in the normalised state space. To further preserve the geometric relation between samples and prototypes, the classifier also computes the Euclidean distance:

$$d_c = \|\hat{\mathbf{s}} - \hat{\mathbf{p}}_c\|_2. \quad (32)$$

A smaller distance indicates that the sample is closer to the corresponding class prototype. Since classification logits should be larger for more likely classes, the distance term

is added with a negative sign. A small residual linear layer is also used to improve optimisation stability:

$$l_c^{\text{res}} = \mathbf{w}_c^T \mathbf{s} + b_c, \quad (33)$$

where $\mathbf{w}_c$ and b_c are the weight vector and bias of the residual classifier for class c . This branch provides a simple discriminative path during training, while the prototype comparison remains the main decision structure.

The final logit for class c is obtained by combining the three terms:

$$l_c = l_c^{\text{cos}} - \lambda_d d_c + \lambda_r l_c^{\text{res}}, \quad (34)$$

where λ_d and λ_r are weighting factors for the distance term and the residual linear term, respectively. In the implementation, $\lambda_d = 0.5$ and $\lambda_r = 0.2$. The predicted class is then given by

$$\hat{y} = \arg \max c \in 1, \dots, Cl_c. \quad (35)$$

Prototype distance indicates how close a speech sample is to the learned HC or MDD acoustic state. Samples closer to the boundary between the two prototypes may show mixed or weak depression-related acoustic patterns.

4. Experimental Setup

4.1. Data

This study uses two public speech datasets for depression detection: DAIC-WOZ [30] and MODMA [31]. These datasets differ in recording conditions, participant backgrounds, and language settings, which makes them suitable for evaluating the robustness of the proposed method.

DAIC-WOZ contains 189 clinical interview recordings collected by the University of Southern California Institute for Creative Technologies. Each participant interacted with a virtual interviewer in a controlled environment. The dataset includes 56 participants with depression and 133 healthy controls. The recordings are sampled at 16 kHz and are accompanied by transcripts, acoustic features, demographic information, and depression labels derived from PHQ-8 scores. MODMA is a multimodal dataset for analysing mental disorders. In this study, we used its audio subset, which includes recordings from 52 participants: 23 outpatients with major depressive disorder and 29 healthy controls. The patients were diagnosed by psychiatrists according to DSM-IV criteria [32] and the MINI structured interview [33]. Each participant completed the interview, reading, and picture description tasks. The recordings were stored in WAV format with a sampling rate of 44.1 kHz and a bit depth of 24 bits.

For both datasets, only the participant's speech was retained. Interviewer speech was removed from the original recordings based on available annotations. Participant segments shorter than 1 s were also discarded, since they usually contain brief acknowledgements, hesitation sounds, or noise rather than complete speech responses. The remaining participant speech segments were concatenated and divided into fixed-length 10 s samples. This process reduces the interviewer's influence and provides inputs with a consistent duration.

Each audio sample was converted into a mel-spectrogram before being fed into the model. The short-time Fourier transform was computed with 2048 FFT points. A Hann window of 1764 samples and a hop length of 882 samples were used, giving a 50% overlap between adjacent frames. The resulting linear spectrogram was mapped to 64 mel filter banks. During training, SpecAugment [34] was applied to improve model generalisation. After preprocessing, DAIC-WOZ contained 4076 healthy-control samples and

2515 depression samples, while MODMA contained 1868 healthy-control samples and 1478 depression samples.

It is worth noting that although each participant recording was segmented into multiple 10 s samples, the cross-validation split was performed at the participant level. All samples from the same participant were kept in the same fold. Therefore, no participant appeared in both the training and test sets.

4.2. Loss

The proposed FKAN was trained with a classification loss and a lightweight regularisation term. The classification loss encouraged the model to predict the correct depression label, while the regularisation term improved the sparsity of KAN edge gates and prevented the class prototypes from collapsing in the learned acoustic state space.

For a training sample with label $y \in 1, \dots, C$, the model produces a logit vector

$$\mathbf{l} = [l_1, l_2, \dots, l_C], \quad (36)$$

where C is the number of classes. In the binary depression detection task, $C = 2$, corresponding to MDD and HC. The classification loss is defined as the cross-entropy loss:

$$\mathcal{L}_{ce} = -\log \frac{\exp(l_y)}{\sum_{c=1}^C \exp(l_c)}. \quad (37)$$

This loss directly optimises the probability assigned to the ground-truth class.

To improve interpretability, the FKAN applies a regularisation loss to the edge gates in all fractional-gated RBF-KAN layers. For the m -th layer, let $e_{i,o}^m = \sigma(g_{i,o}^m)$ denote the learned gate value of the edge from input dimension i to output dimension o (mentioned in Section 3.4). The edge regularisation of this layer is defined as

$$\mathcal{L}_{edge}^m = \lambda_1 \frac{1}{I_m O_m} \sum_{i=1}^{I_m} \sum_{o=1}^{O_m} e_{i,o}^m + \lambda_2 \frac{1}{I_m O_m} \sum_{i=1}^{I_m} \sum_{o=1}^{O_m} [-e_{i,o}^m \log(e_{i,o}^m + \epsilon) - (1 - e_{i,o}^m) \log(1 - e_{i,o}^m + \epsilon)], \quad (38)$$

where I_m and O_m are the input and output dimensions of the m -th layer and $\epsilon = 10^{-8}$. The first part encourages sparse edge connections, and the second part encourages each gate to become more confident. These two constraints make the learned edge functions easier to analyse. In our implementation, $\lambda_1 = 2 \times 10^{-5}$ and $\lambda_2 = 2 \times 10^{-5}$.

We further used a prototype separation loss to avoid prototype collapse. Given the normalized class prototypes $\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, \dots, \hat{\mathbf{p}}_C$, this term is written as

$$\mathcal{L}_{proto} = \frac{1}{C^2} \sum_{i=1}^C \sum_{j=1, j \neq i}^C (\hat{\mathbf{p}}_i^\top \hat{\mathbf{p}}_j)^2. \quad (39)$$

It penalises high similarity between different class prototypes and helps maintain a clearer depression-related state space.

The final training objective is

$$\mathcal{L} = \mathcal{L}_{ce} + \sum_{m=1}^M \mathcal{L}_{edge}^m + \lambda_p \mathcal{L}_{proto}, \quad (40)$$

where M is the number of fractional-gated RBF-KAN layers used in the KAN blocks and the classifier. The prototype separation weight is set to $\lambda_p = 10^{-4}$. This objective enables the FKAN to learn discriminative depression-related representations while preserving edge functions and class prototypes that are suitable for interpretation.

4.3. Training Details

The proposed model was developed with PyTorch for Python 3.10. All experiments were conducted on a server with an AMD Ryzen Threadripper 2990WX CPU and an NVIDIA RTX TITAN GPU. The random seed was set to 42 in all experiments. The model parameters were optimised using Adam [35]. The initial learning rate was set to 1×10^{-4} and was updated during training using a step-decay strategy. At each decay step, the learning rate was multiplied by 0.9. The batch size was set to 48, and the model was trained for up to 100 epochs. The number of fractional coordinate bases and RBF bases was selected during model development using training and validation data only. The test folds were not used for hyperparameter selection.

4.4. Evaluation Metrics

To evaluate the performance of different models, four commonly used metrics are adopted, including precision, recall, F1-score, and accuracy. In the binary depression detection task, MDD is treated as the positive class, and HC is treated as the negative class. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Precision measures the proportion of correctly predicted MDD samples among all samples predicted as MDD:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (41)$$

Recall measures the proportion of correctly detected MDD samples among all real MDD samples:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (42)$$

F1-score is the harmonic mean of precision and recall:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (43)$$

Accuracy measures the overall proportion of correctly classified samples:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (44)$$

5. Results and Analysis

5.1. Results on DAIC-WOZ

Table 1 presents the subject-independent 10-fold validation performance of the FKAN on the DAIC-WOZ dataset, while Figure 6 provides the corresponding confusion matrix. The FKAN achieves an overall accuracy of 0.95 ± 0.007 , showing stable performance across different validation folds. This result indicates that the proposed model can capture depression-related acoustic patterns from DAIC-WOZ despite the natural variability in interview-based speech. For the HC class, the model achieved a precision of 0.93 ± 0.012 , recall of 0.99 ± 0.003 , and an F1-score of 0.96 ± 0.006 . The confusion matrix shows that 4051 HC samples were correctly classified, with only 25 samples incorrectly assigned to the MDD class. This indicates that the FKAN produced few false positive depression predictions under the current decision setting. For the MDD class, the FKAN reached a precision of 0.98 ± 0.006 , a recall of 0.88 ± 0.017 , and an F1-score of 0.93 ± 0.009 . Among all of the MDD samples, 2222 were correctly detected, while 293 were misclassified as HC. Although the high MDD precision indicates that predicted depressive samples were generally reliable, the lower recall shows that the current decision setting may have missed some depressive

cases. Therefore, sensitivity should be carefully balanced against the risk of false alarms in screening-oriented applications.

Table 1. The 10-fold test results of the proposed method on the DAIC-WOZ dataset. The number after “ $\pm$ ” denotes the standard deviation.

Class	Precision	Recall	F1-Score	Accuracy
HC	0.93 ± 0.012	0.99 ± 0.003	0.96 ± 0.006	0.95 ± 0.007
MDD	0.98 ± 0.006	0.88 ± 0.017	0.93 ± 0.009	

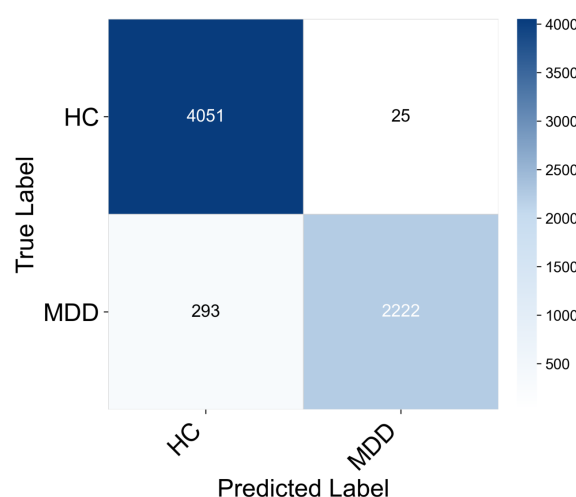


Figure 6. Confusion matrix of DAIC-WOZ dataset.

Table 2 presents a reference-level comparison between FKAN and several recent depression detection methods on the DAIC-WOZ dataset. These methods use different acoustic inputs, including MFCCs, spectrograms, waveform signals, mel-spectrograms, CQT features, and HuBERT representations. Because the compared studies may differ in preprocessing procedures, participant splits, and other experimental settings, the results should not be interpreted as a strictly controlled benchmark. Instead, the table summarises the values reported in the original studies to provide contextual reference. Under the protocol used in this work, FKAN achieves an average F1-score of 0.94 ± 0.007 using only mel-spectrograms. This value is higher than the reported scores of CNN-ANN, GNN, DMPF, LTD-Conformer, and AMAST. CNN-ANN reports an average F1-score of 0.90, while GNN reports 0.92 using MFCC, F0, and CQT features. DMPF and LTD-Conformer report average F1-scores of 0.80 and 0.87, respectively, using waveform input and combined mel-spectrogram and HuBERT features. AMAST is the closest reported method, with F1-scores of 0.94 for HC and 0.90 for MDD. FKAN obtains 0.96 and 0.93 for the two classes under the present evaluation setting. These comparisons suggest that FKAN is competitive with recent approaches, while a unified reproduction under the same preprocessing and subject-level split would be needed for a strict fairness assessment.

Table 2. Comparison with other models on the DAIC-WOZ dataset. The number after “ $\pm$ ” denotes the standard deviation.

Method	Feature(s)	Year	F1-Score (HC/MDD)	F1-Score (Average)
CNN-ANN [36]	MFCC and Spectrogram	2024	0.90/0.89	0.90
GNN [6]	MFCC, F0 and CQT	2024	-	0.92
DMPF [37]	Waveform	2025	-	0.80
LTD-Conformer [8]	Mel-Spectrogram and HuBERT Feature	2025	0.88/0.86	0.87
AMAST [7]	Mel-Spectrogram	2025	0.94/0.90	0.92
FKAN (ours)	Mel-Spectrogram	2026	0.96/0.93	0.94 ± 0.007

5.2. Results on MODMA

The performance of FKAN on the MODMA dataset is presented in Table 3, with the confusion matrix shown in Figure 7. Under the 10-fold validation setting, FKAN obtains an overall accuracy of 0.94 ± 0.014 . This result shows that the proposed model maintains strong performance on another depression dataset, even though MODMA differs from DAIC-WOZ in data collection conditions and subject characteristics. For the HC class, the model achieves a precision of 0.93 ± 0.019 , a recall of 0.97 ± 0.015 , and an F1-score of 0.95 ± 0.013 . The confusion matrix shows that 1813 HC samples are correctly classified, while 55 samples are incorrectly predicted as MDD. This indicates that FKAN can identify most healthy speech samples and maintains a low false alarm rate. For the MDD class, FKAN reaches a precision of 0.96 ± 0.018 , a recall of 0.91 ± 0.026 , and an F1-score of 0.94 ± 0.016 . Among the MDD samples, 1351 are correctly detected, and 127 are misclassified as HC. Compared with the HC class, the lower MDD recall indicates that missed depressive samples remain more common than HC-to-MDD false positives. This pattern may reflect the acoustic heterogeneity of depressive speech, as some MDD participants may exhibit speech characteristics similar to those of the HC group. In addition, the current decision boundary appears relatively conservative for MDD prediction, helping limit false alarms but potentially increasing missed detections. The confusion matrix confirms that the classification errors are not heavily biased toward one class. The number of HC samples misclassified as MDD is much smaller than the number of correctly recognised HC samples, and most MDD samples are also detected correctly. This balanced behaviour is important for practical depression screening because the model should avoid both excessive false alarms and severe missed detections.

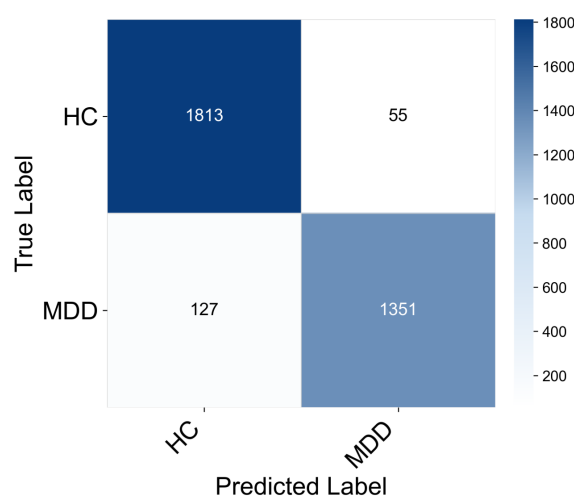
**Figure 7.** Confusion matrix of MODMA dataset.

Table 3. 10-fold test results of the proposed method on the MODMA dataset. The number after “ $\pm$ ” denotes the standard deviation.

Class	Precision	Recall	F1-Score	Accuracy
HC	0.93 ± 0.019	0.97 ± 0.015	0.95 ± 0.013	0.94 ± 0.014
MDD	0.96 ± 0.018	0.91 ± 0.026	0.94 ± 0.016	

Table 4 reports the comparison between the FKAN and recent baseline methods on the MODMA dataset. The FKAN obtained the best average F1-score of 0.94 ± 0.014 , outperforming the compared methods. Compared with CNN-ANN, the average F1-score rose from 0.89 to 0.94. The class-wise results follow the same trend, with the HC F1-score increasing from 0.91 to 0.95 and the MDD F1-score increasing from 0.88 to 0.94. This indicates that the improvement was not caused by a bias toward the majority or the upper class, but reflects a more reliable separation between healthy and depressive speech. Similar conclusions can be drawn from the comparison with GNN and DMPF. GNN reaches an average F1-score of 0.89 using MFCC, F0, and CQT features, and DMPF reports 0.77 using waveform input. In contrast, the FKAN achieved higher accuracy with only mel-spectrograms, supporting the effectiveness of its time–frequency modelling strategy. AMAST remains the strongest baseline in this group, with an average F1-score of 0.91 and class-wise F1-scores of 0.92 for HC and 0.89 for MDD. The FKAN increased the HC and MDD F1-scores to 0.95 and 0.94, respectively. The larger improvement in MDD suggests that fractional-coordinate encoding and edge-wise KAN transformations help capture subtle and variable acoustic patterns associated with depression.

Table 4. Comparison with other models on the MODMA dataset. The number after “ $\pm$ ” denotes the standard deviation.

Method	Feature(s)	Year	F1-Score (HC/MDD)	F1-Score (Average)
CNN-ANN [36]	MFCC and Spectrogram	2024	0.91/0.88	0.89
GNN [6]	MFCC, F0 and CQT	2024	-	0.89
DMPF [37]	Waveform	2025	-	0.77
AMAST [7]	Mel-Spectrogram	2025	0.92/0.89	0.91
FKAN (ours)	Mel-Spectrogram	2026	0.95/0.94	0.94 ± 0.014

5.3. Cross-Dataset Generalization

Figure 8 shows the cross-dataset generalisation results of the FKAN. When the model was trained on MODMA and tested on DAIC-WOZ, 3546 HC samples and 1861 MDD samples were correctly classified, while 530 HC samples were predicted as MDD and 654 MDD samples were predicted as HC. Even under this domain shift, the FKAN still preserved a clear separation between the HC and MDD samples. The result suggests that the model not only learns dataset-specific patterns but also captures acoustic cues that can transfer across corpora. This behaviour is related to the design of the FKAN, which encodes fractional coordinates to model gradual time–frequency changes, and the KAN blocks learn flexible edge-wise functions rather than relying on fixed nonlinear transformations.

When the FKAN was trained on DAIC-WOZ and tested on MODMA, the model correctly classified 1700 HC samples and 1182 MDD samples, misclassifying 168 HC samples and 296 MDD samples. This direction gave fewer errors than the MODMA-to-DAIC-WOZ setting, which may be because DAIC-WOZ provided a broader training distribution and helped the model learn more general depression-related acoustic patterns. In both directions, most errors arose from MDD samples being misclassified as HC, suggesting that some depressive speech patterns remained close to healthy speech when transferred across datasets. The prototype-based classifier helped reveal this behaviour because uncertain

samples were likely to lie near the boundary between the MDD and HC prototypes in the learned acoustic state space.

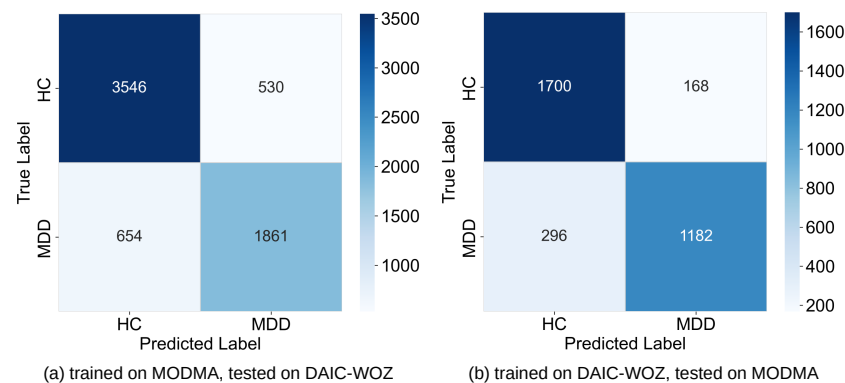


Figure 8. Confusion matrix for cross-dataset test.

5.4. Ablation Study

Table 5 reports the ablation results on DAIC-WOZ and MODMA. To clearly evaluate the contribution of each component, six variants were constructed from the full FKAN. To distinguish fractional modelling from general positional information, FCE was replaced with a learnable position embedding (PE) while keeping the remaining architecture unchanged. In the “w/o FCE” case, the fractional coordinate encoding module was removed, and the acoustic tokens from the convolution frontend were directly concatenated with the class token before being sent to the KAN blocks. In the “w/o Fraction Gate” case, the RBF-KAN structure was retained, but the fractional gate in the edge function was removed, so the edge response was only determined by the RBF basis and the learnable edge gate. In the “RBF-KAN $\rightarrow$ MLP-FFN” case, the fractional-gated RBF-KAN layer in each KAN block was replaced with a conventional feed-forward layer with fixed nonlinear activation. For a stronger comparison, the RBF-KAN layers were also replaced with standard Transformer-FFN layers while all other components remained unchanged. In the “Classifier $\rightarrow$ FC Layer case”, the prototype-based classifier was replaced with a standard fully connected classifier, and the class-token representation was directly mapped to the output logits.

Table 5. Results of the ablation study.

Dataset	Item	Precision	Recall	F1
DAIC-WOZ	w/o FCE	0.9318	0.9057	0.9186
	FCE $\rightarrow$ PE	0.9387	0.9135	0.9259
	w/o Fraction Gate	0.9442	0.9175	0.9307
	RBF-KAN $\rightarrow$ MLP-FFN	0.9169	0.8948	0.9057
	RBF-KAN $\rightarrow$ Transformer-FFN	0.9298	0.9081	0.9188
	Classifier $\rightarrow$ FC Layer	0.9256	0.9029	0.9141
	Full FKAN	0.9606	0.9386	0.9476
MODMA	w/o FCE	0.9185	0.9056	0.9120
	FCE $\rightarrow$ PE	0.9262	0.9142	0.9202
	w/o Fraction Gate	0.9324	0.9183	0.9253
	RBF-KAN $\rightarrow$ MLP-FFN	0.9048	0.8921	0.8984
	RBF-KAN $\rightarrow$ Transformer-FFN	0.9176	0.9038	0.9106
	Classifier $\rightarrow$ FC Layer	0.9136	0.9019	0.9077
	Full FKAN	0.9480	0.9422	0.9445

The full FKAN achieved the best results on both datasets, with F1-scores of 0.9476 on DAIC-WOZ and 0.9445 on MODMA. Removing FCE reduced the F1-score to 0.9186 and 0.9120, respectively, showing that the coordinate-dependent fractional response helps

model depression-related time–frequency patterns. Replacing FCE with a learnable position embedding improved the results to 0.9259 and 0.9202, but still remained below the full FKAN. This suggests that the gain is not solely due to positional information but also to the proposed fractional-coordinate modelling. Ordinary acoustic tokens alone are less effective at capturing the slow, persistent changes often found in depressed speech. Removing the fractional gate also caused clear degradation, with F1-scores of 0.9307 on DAIC-WOZ and 0.9253 on MODMA. Since the RBF-KAN structure was retained, this gap mainly reflects the contribution of fractional modulation to the edge functions, supporting the need for adaptive non-integer-order responses. The largest drop occurred when RBF-KAN was replaced with MLP-FFN, where the F1-score fell to 0.9057 on DAIC-WOZ and 0.8984 on MODMA. Replacing RBF-KAN with Transformer-FFN improved the F1-scores to 0.9188 and 0.9106, but the FKAN still underperformed under the same setting. This indicates that edge-wise functional representation is more suitable than conventional feed-forward transformation for distinguishing subtle acoustic differences between MDD and HC speech. Replacing the prototype-based classifier with a fully connected layer also weakened performance, with F1-scores of 0.9141 and 0.9077. This suggests that the acoustic state prototypes contribute not only to visualisation but also to classification, as they form a clearer depression-related state space.

5.5. Model Complexity

Table 6 reports the computational complexity of the FKAN. The inferences are based on the CPU mentioned in Section 4.3. The model contains 699 K trainable parameters, indicating that FKAN maintains a compact scale. This is mainly because the convolutional frontend is lightweight, the token grid is reduced in size, and the KAN blocks operate on compact token features rather than the original high-resolution spectrogram. The computational cost was 318 M FLOPs per sample in the test dataset, with an inference time of 13.01 ms and a peak memory usage of 24 MB. These results indicate that the FKAN does not rely on a large backbone or a heavy pre-trained speech model to obtain strong depression detection performance. The class token also avoided expensive global aggregation over the full spectrogram, while the fractional coordinate encoding introduced only a small gating module. Edge-wise KAN functions improved the model’s representational capacity, but the overall cost remained moderate because they were applied after tokenisation.

Table 6. Model complexity analysis.

Item	Value
Number of Parameters	699 K
FLOPs	318 M
Inference Time	13.01 ms
Memory Usage (Peak)	24 MB

5.6. Parameter Sensitivity Analysis

5.6.1. Basis Number

Figure 9 shows the effect of basis number on FKAN performance. For the fractional coordinate basis (K_c in Section 3.3), the best performance was achieved with a basis size of 6, yielding F1-scores of 0.9476 on DAIC-WOZ and 0.9445 on MODMA. When the basis number was reduced to 2, the F1-scores decreased to 0.9217 and 0.9248, respectively, indicating that too few coordinate bases are insufficient to describe depression-related time–frequency changes. Increasing the basis number from 2 to 6 improved performance on both datasets, supporting the motivation for using multiple fractional-coordinate responses to model gradual acoustic variations. When the basis number was further increased to 8 and 10,

the performance started to decline. DAIC-WOZ decreased to 0.9423 and 0.9315, while MODMA decreased to 0.9342 and 0.9267. This suggests that excessive coordinate bases may introduce redundant positional responses and weaken generalisation.

The right part of Figure 9 presents the influence of the RBF basis number on the KAN blocks. The model achieves the highest F1-score when the RBF basis number (K_r in Section 3.4) is 4, reaching 0.9476 on DAIC-WOZ and 0.9445 on MODMA. With only 2 RBF bases, the edge functions had limited flexibility, and the F1-scores decreased to 0.9338 and 0.9189, respectively. When the basis number exceeded 4, performance showed a downward trend, but the two datasets exhibited slightly different sensitivities. On DAIC-WOZ, the F1-score decreased from 0.9392 at 6 bases to 0.9254 at 10 bases. On MODMA, the F1-score changed from 0.9368 to 0.9321 and then dropped to 0.9216. This trend shows that the KAN edge functions need a suitable level of complexity. A small number of bases cannot fully model subtle acoustic differences between MDD and HC speech, while too many bases may make the edge functions overly flexible for limited clinical speech data.

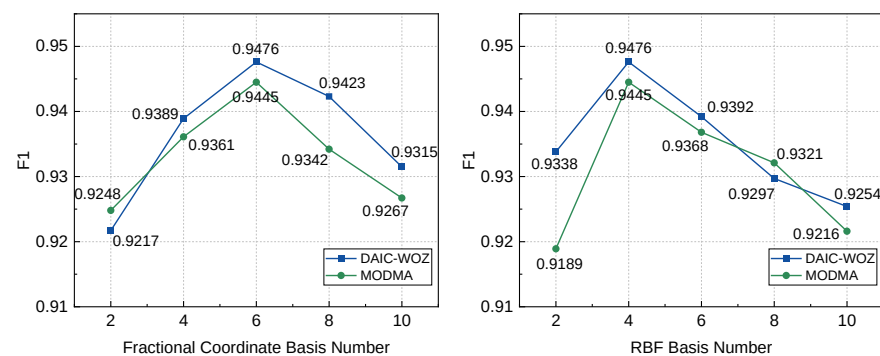


Figure 9. Effect of basis number on FKAN performance across DAIC-WOZ and MODMA.

5.6.2. Fractional Order

Figure 10 presents the effect of the maximum fractional order on FKAN performance. The left part shows the results of $\alpha_{\max}$, which controls the strength of fractional coordinate responses in FCE. When $\alpha_{\max}$ was set to 0.2, the F1-scores were 0.9224 on DAIC-WOZ and 0.9158 on MODMA, indicating that a weak fractional coordinate response is not sufficient to describe depression-related time–frequency changes. As $\alpha_{\max}$ increased, the performance improved steadily and reached the best results at $\alpha_{\max} = 0.8$, with F1-scores of 0.9476 and 0.9445 on the two datasets. When $\alpha_{\max}$ was further increased to 1.0, the F1-scores decreased to 0.9382 and 0.9351. This result shows that the fractional response range should be selected carefully. A moderate upper bound provides sufficient coordinate modulation, whereas an excessively large range may overemphasise positional responses and interfere with the acoustic content of the tokens.

The right part of Figure 10 shows the influence of $\rho_{\max}$, which controls the maximum fractional order in the KAN edge gate. The model performed best when $\rho_{\max} = 0.5$, achieving F1-scores of 0.9476 on DAIC-WOZ and 0.9445 on MODMA. A smaller value, such as $\rho_{\max} = 0.1$, weakened the fractional modulation of edge functions, leading to lower F1-scores of 0.9326 and 0.9249. Increasing $\rho_{\max}$ to 0.3 improved the results, but further increases beyond 0.5 caused a clear decline. At $\rho_{\max} = 0.9$, the F1-scores dropped to 0.9308 and 0.9287. This trend indicates that the fractional gate helps KAN edge functions adapt to depression-sensitive acoustic features, whereas excessive modulation may make edge responses overly sensitive to feature magnitude. The results support the design choice of using a limited fractional order in the proposed FKAN.

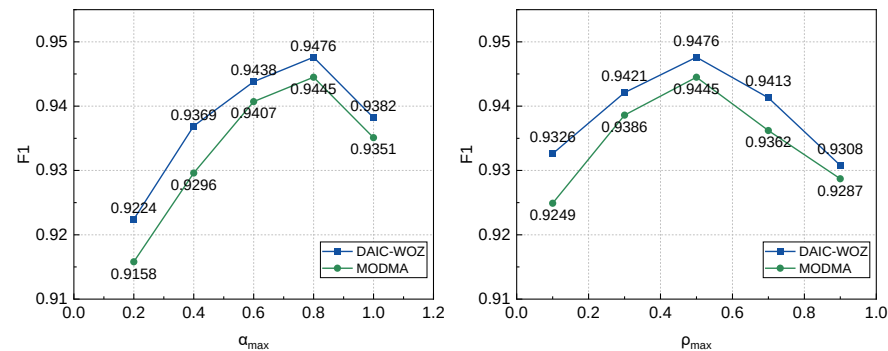


Figure 10. Effect of maximum fractional order on FKAN performance across DAIC-WOZ and MODMA.

5.7. Visualization

5.7.1. t-SNE Analysis

Figure 11 visualises the learned acoustic state representations of the FKAN using t-SNE on DAIC-WOZ and MODMA. In both datasets, the HC and MDD samples formed two largely separated regions, indicating that the proposed model learned discriminative depression-related representations rather than merely fitting the final classification boundary. The class prototypes were also located near their corresponding sample clusters, showing that the prototype-based classifier provides meaningful reference points in the learned state space. On DAIC-WOZ, the HC samples spanned a broader region, while the MDD samples were more concentrated on the right side, with only a few boundary samples between the 2 groups. On MODMA, the two classes were also clearly separated, though the distribution was more scattered, consistent with the domain differences observed in previous experiments.

Beyond calculation and visualisation, these distances provide a sample-level measure of decision confidence. A sample that is close to one prototype and far from the other has a clearer class assignment, whereas similar distances to both prototypes indicate a more uncertain case. This information can support error analysis and future analysis of changes in depression-related speech states.

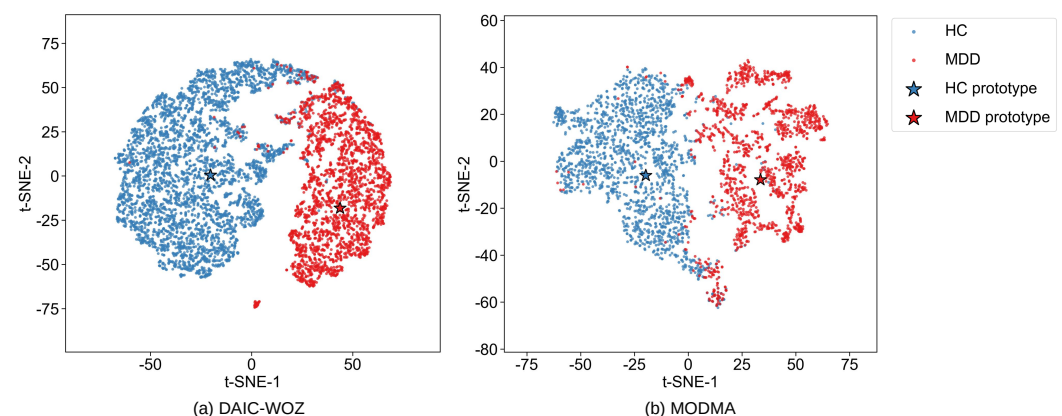


Figure 11. The t-SNE analysis on two datasets.

5.7.2. FCE-Modulated Feature Response

Figure 12 visualises the FCE-modulated feature response on the tokenised time–frequency grid. The plotted variable was obtained from the token features after fractional coordinate encoding, in which the FCE gate softly modulated each acoustic token based on its normalised time–frequency coordinate. The response was averaged over the feature dimension and then reshaped into a 16×4 token grid, so each cell reflects the average modulation strength at a specific time–frequency location. Figure 12a,b show the average

response maps of the HC and MDD samples, while Figure 12c shows their difference. The two groups shared a similar global structure, suggesting that FCE preserves a common acoustic organisation rather than introducing arbitrary positional distortion. At the same time, the difference map shows clear group-dependent variations. Compared with HC, MDD shows relatively stronger responses in the middle-to-high frequency regions and weaker responses in several low-frequency temporal regions. This pattern indicates that FCE captures class-related changes in the distribution of acoustic information across time and frequency. Such behaviour is consistent with the FKAN's motivation, since depression-related speech cues are often expressed as gradual changes in energy, prosody, and spectral balance rather than isolated local events. The visualisation, therefore, provides evidence that FCE contributed to interpretable time–frequency modulation, helping the model learn depression-sensitive acoustic representations.

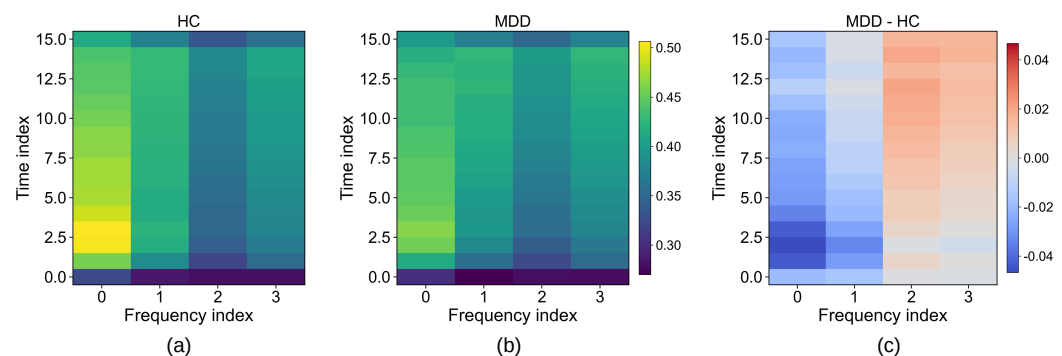


Figure 12. Visualisation of the FCE-modulated feature response on the tokenized time–frequency grid. (a,b) show the average response maps over the HC and MDD groups. (c) is the group difference between MDD and HC.

5.8. Sensitivity Test for Pre-Processing Settings

Table 7 evaluates the sensitivity of the FKAN to several pre-processing settings. For mel resolution, 64 mel bands achieved the best F1-scores on both datasets. Using 40 bands may have removed useful spectral detail, while 80 bands provided no further benefit and slightly reduced performance. For segment length, 10 s gave the best results. Shorter segments contained less temporal context, whereas longer segments may have introduced more within-sample variation and reduced representation consistency. The effect of FFT size differed slightly across datasets. DAIC-WOZ performed best with a 1024-point FFT, while MODMA performed best with a 2048-point FFT. This difference is consistent with their sampling rates and recording conditions. The FKAN maintained competitive performance under the tested pre-processing settings, while the adopted configuration balanced performance and input resolution.

Table 7. Sensitivity of the proposed method to different pre-processing settings. Bold text indicates the settings we have adopted.

Item	Setting	DAIC-WOZ F1	MODMA F1
Mel Bands	40	0.9368	0.9321
	64	0.9476	0.9445
	80	0.9427	0.9394
Segment Length	5 s	0.9342	0.9286
	10 s	0.9476	0.9445
	15 s	0.9409	0.9378
FFT Size	512	0.9349	0.9275
	1024	0.9476	0.9389
	2048	0.9418	0.9445

6. Conclusions

This paper presented the FKAN, a fractional Kolmogorov–Arnold network for interpretable speech-based depression detection. The proposed framework learns acoustic biomarkers from mel-spectrograms by integrating fractional coordinate encoding, fractional-gated edge-wise KAN blocks, and a prototype-based depression-state classifier. The fractional-coordinate encoding introduces non-integer-order time–frequency responses to describe gradual, persistent acoustic variations in speech. The fractional-gated KAN blocks further replace fixed node-wise nonlinear transformations with adaptive edge-wise functional mappings, enabling more flexible modelling of depression-related acoustic patterns. In addition, the prototype-based classifier maps speech samples into a compact acoustic state space, where the relationship between each sample and the MDD or HC prototype can be explicitly analysed.

Experiments on two public depression speech datasets demonstrated the effectiveness of FKAN. On DAIC-WOZ, the FKAN achieved an accuracy of 0.95 and an average F1-score of 0.94. On MODMA, it achieved an accuracy of 0.94 and an average F1-score of 0.94. These results show that the proposed fractional functional modelling strategy delivers strong, stable performance across different speech datasets. The comparison with recent baseline methods further confirmed the FKAN’s advantage in learning discriminative acoustic representations from mel-spectrograms. Ablation studies verified the contributions of fractional coordinate encoding, fractional-gated KAN blocks, and the prototype-based classifier. Sensitivity analysis showed that moderate basis numbers and fractional orders are more effective than overly simple or overly strong settings, suggesting that fractional responses should be carefully controlled when modelling depression-related speech dynamics. Cross-dataset evaluation further indicated that the learned representations exhibit some degree of transferability across different recording conditions and participant populations.

The visualisation results also showed that the FKAN provides useful interpretability for acoustic biomarker analysis. The learned prototype space yielded meaningful distributions of the MDD and HC samples, while fractional modulation patterns and edge-function importance provided additional evidence for understanding class-related acoustic differences. These findings suggest that the FKAN is not only an effective depression-detection model but also a promising framework for connecting fractional functional representations with interpretable acoustic biomarker learning. Although the FKAN achieved promising results on two public datasets, these findings should be interpreted as research evidence rather than clinical validation. Further evaluation on larger, more diverse, and prospectively collected clinical datasets is needed before the method can be considered for practical use. Therefore, future work will further validate the proposed method on larger and more diverse clinical datasets, extend the framework from binary detection to continuous depression severity estimation, and improve its robustness under cross-language, cross-device, and real-world recording conditions. The current model uses mel-spectrograms as its only input. Although this representation provides useful time–frequency information, it does not explicitly preserve detailed dynamics. In addition, the tokenisation stage compresses local time–frequency details. Future work will extend the FKAN to multi-feature speech representations that combine mel-spectrograms with waveform, phase, prosody, and voice-quality features. We will also evaluate the FKAN on larger external datasets and recordings collected under more diverse acoustic conditions, together with clinician-guided validation, to further assess its generalisation and clinical relevance.

Author Contributions: J.Y.: Conceptualisation, Methodology, Software, Validation, Visualisation, Writing—original draft. Y.Z.: Methodology, Software, Validation, Funding acquisition, Supervision, Writing—review and editing. Y.L.: Methodology, Software, Validation, Funding acquisition, Super-

vision, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset available on request from the authors.

Acknowledgments: The authors gratefully acknowledge the providers and maintainers of the DAIC-WOZ and MODMA datasets for granting permission to access and use these datasets in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kupferberg, A.; Hasler, G. The social cost of depression: Investigating the impact of impaired social emotion regulation, social cognition, and interpersonal behavior on social functioning. *J. Affect. Disord. Rep.* **2023**, *14*, 100631. <https://doi.org/10.1016/j.jadr.2023.100631>.
2. Chen, X.D.; Li, F.; Zuo, H.; Zhu, F. Trends in Prevalent Cases and Disability-Adjusted Life-Years of Depressive Disorders Worldwide: Findings from the Global Burden of Disease Study from 1990 to 2021. *Depress. Anxiety* **2025**, *2025*, 5553491. <https://doi.org/10.1155/da/5553491>.
3. Askim, K.; Knardahl, S. The Influence of Affective State on Subjective-Report Measurements: Evidence From Experimental Manipulations of Mood. *Front. Psychol.* **2021**, *21*, 601083. <https://doi.org/10.3389/fpsyg.2021.601083>.
4. Lin, Q.; Wu, X.; Lei, Y.; Cheng, W.; Huang, S.; Wang, W.; Li, C.; Zhao, J. Acoustic signatures of depression elicited by emotion-based and theme-based speech tasks. *BMJ Ment. Health* **2025**, *28*, e301858. <https://doi.org/10.1136/bmjment-2025-301858>.
5. Qin, S.; Zhang, Y.; Ma, Y.; Li, H.; Li, X.; Lian, B.; Cai, W.; Cui, J.; Zhao, X. A cross-linguistic depression detection method based on speech data. *J. Affect. Disord.* **2025**, *390*, 119739. <https://doi.org/10.1016/j.jad.2025.119739>.
6. Sun, C.; Jiang, M.; Gao, L.; Xin, Y.; Dong, Y. A novel study for depression detecting using audio signals based on graph neural network. *Biomed. Signal Process. Control* **2024**, *88*, 105675. <https://doi.org/10.1016/j.bspc.2023.105675>.
7. Zhang, M.; He, J.; Peng, X.; Huang, J.; Zhang, N.; Wang, C.; Jiang, D. Auto-Masked Audio Spectrogram Transformer for depression detection from speech. *J. Affect. Disord.* **2026**, *393*, 120295. <https://doi.org/10.1016/j.jad.2025.120295>.
8. Lee, J.; Hong, J.; Choi, D.; Jung, J. LTD-Conformer: Speech Depression Detection with Speaking and Listening Perspectives. In Proceedings of the 2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS), Madrid, Spain, 18–20 June 2025; pp. 399–404. <https://doi.org/10.1109/CBMS65348.2025.00087>.
9. Schwenzer, M.; Zattarin, E.; Grözing, M.; Mathiak, K. Impaired pitch identification as a potential marker for depression. *BMC Psychiatry* **2012**, *12*, 32. <https://doi.org/10.1186/1471-244X-12-32>.
10. Low, L.S.A.; Maddage, N.C.; Lech, M.; Sheeber, L.B.; Allen, N.B. Detection of Clinical Depression in Adolescents' Speech During Family Interactions. *IEEE Trans. Biomed. Eng.* **2011**, *58*, 574–586. <https://doi.org/10.1109/TBME.2010.2091640>.
11. Akkaralaertsest, T.; Yingthawornsuk, T. Classification of Depressed Speech Samples with Spectral Energy Ratios as Depression Indicator. In Proceedings of the 2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Chiang Mai, Thailand, 30 October–1 November 2019; pp. 1–6. <https://doi.org/10.1109/iSAI-NLP48611.2019.9045167>.
12. Mitra, V.; Tsiartas, A.; Shriberg, E. Noise and reverberation effects on depression detection from speech. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5795–5799. <https://doi.org/10.1109/ICASSP.2016.7472788>.
13. Shinohara, S. Depression: Jitter, Shimmer, Harmonic to Noise Ratio, and Pitch Rate. In *Voice Biomarkers: Current Status and Issues in the Development*; Springer: Singapore, 2025; pp. 81–92. https://doi.org/10.1007/978-981-96-3109-4_7.
14. Verma, A.; Jain, P.; Kumar, T. An Effective Depression Diagnostic System Using Speech Signal Analysis Through Deep Learning Methods. *Int. J. Artif. Intell. Tools* **2023**, *32*, 2340004. <https://doi.org/10.1142/S0218213023400043>.
15. Shi, D.; Lu, X.; Liu, Y.; Yuan, J.; Pan, T.; Li, Y. Research on Depression Recognition Using Machine Learning from Speech. In Proceedings of the 2021 International Conference on Asian Language Processing (IALP), Singapore, 11–13 December 2021; pp. 52–56. <https://doi.org/10.1109/IALP54817.2021.9675271>.

16. Kumari, A.R.; Kumari, A.V.S.R. Adaptive Hybrid Clustering for Automated Teeth Segmentation Framework with Improved Eurasian Oystercatcher-Based Weighted Fused Features. *Cybern. Syst.* **2025**, 1–45. <https://doi.org/10.1080/01969722.2025.2556732>.
17. Mounika, P.; Narayanan, B.; Balmuri, K.R. An Adaptive Trans-ResUnet++-Based Segmentation and Hybrid CNN-Aided Classification for Detecting Breast Cancer from Mammogram Images. *Cybern. Syst.* **2025**, 1–38. <https://doi.org/10.1080/01969722.2025.2540517>.
18. Thamizharasi, M.; Lakshmi, M. Optimizing Alzheimer’s Disease Detection with Dual Attention CNN and Gooseneck Barnacle Search. *Cybern. Syst.* **2026**, 57, 225–258. <https://doi.org/10.1080/01969722.2025.2521703>.
19. Vázquez-Romero, A.; Gallardo-Antolín, A. Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks. *Entropy* **2020**, 22, 688. <https://doi.org/10.3390/e22060688>.
20. Rejaibi, E.; Komaty, A.; Meriaudeau, F.; Agrebi, S.; Othmani, A. MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech. *Biomed. Signal Process. Control* **2022**, 71, 103107. <https://doi.org/10.1016/j.bspc.2021.103107>.
21. Li, Q.; Wang, D.; Ren, Y.; Gao, Y.; Li, Y. FTA-net: A Frequency and Time Attention Network for Speech Depression Detection. In Proceedings of the Interspeech 2023, Dublin, Ireland, 20–24 August 2023; pp. 1723–1727. <https://doi.org/10.21437/Interspeech.2023-296>.
22. Harati, A.; Shriberg, E.; Rutowski, T.; Chlebek, P.; Lu, Y.; Oliveira, R. Speech-Based Depression Prediction Using Encoder-Weight-Only Transfer Learning and a Large Corpus. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7273–7277. <https://doi.org/10.1109/ICASSP39728.2021.9414208>.
23. Zhang, X.; Zhang, X.; Chen, W.; Li, C.; Yu, C. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. *Sci. Rep.* **2024**, 14, 9543. <https://doi.org/10.1038/s41598-024-60278-1>.
24. Yin, F.; Du, J.; Xu, X.; Zhao, L. Depression Detection in Speech Using Transformer and Parallel Convolutional Neural Networks. *Electronics* **2023**, 12, 328. <https://doi.org/10.3390/electronics12020328>.
25. Zhao, Q.; Fan, H.Z.; Li, Y.L.; Liu, L.; Wu, Y.X.; Zhao, Y.L.; Tian, Z.X.; Wang, Z.R.; Tan, Y.L.; Tan, S.P. Vocal Acoustic Features as Potential Biomarkers for Identifying/Diagnosing Depression: A Cross-Sectional Study. *Front. Psychiatry* **2022**, 13, 815678. <https://doi.org/10.3389/fpsy.2022.815678>.
26. Lin, Y.; Liyanage, B.N.; Xu, C.; Zhu, Z.; Liao, Y.; Yang, J.; Tao, Y.; Li, Z.; Shi, C.; Yue, W. Speech-derived acoustic biomarkers for depression: Comprehensive cross-section and longitudinal analyses in different cohorts. *J. Affect. Disord.* **2026**, 403, 121374. <https://doi.org/10.1016/j.jad.2026.121374>.
27. Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljacic, M.; Hou, T.; Tegmark, M. KAN: Kolmogorov–Arnold Networks. In Proceedings of the International Conference on Learning Representations, Singapore, 24 April 2025; pp. 70367–70413.
28. Koudounas, A.; La Quatra, M.; Pastor, E.; Siniscalchi, S.M.; Baralis, E. “KAN you hear me?” Exploring Kolmogorov–Arnold Networks for Spoken Language Understanding. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025; pp. 4123–4127. <https://doi.org/10.21437/Interspeech.2025-1612>.
29. Mai, Y.; Goetze, S. MetricGAN+KAN: Kolmogorov–Arnold Networks in Metric-Driven Speech Enhancement Systems. In Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; pp. 1–5. <https://doi.org/10.1109/ICASSP49660.2025.10890542>.
30. Gratch, J.; Artstein, R.; Lucas, G.; Stratou, G.; Scherer, S.; Nazarian, A.; Wood, R.; Boberg, J.; DeVault, D.; Marsella, S.; et al. The Distress Analysis Interview Corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*; Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., Eds.; European Language Resources Association: Reykjavik, Iceland, 2014; pp. 3123–3128.
31. Cai, H.; Yuan, Z.; Gao, Y.; Sun, S.; Li, N.; Tian, F.; Xiao, H.; Li, J.; Yang, Z.; Li, X.; et al. A multi-modal open dataset for mental-disorder analysis. *Sci. Data* **2022**, 9, 178. <https://doi.org/10.1038/s41597-022-01211-x>.
32. Bell, C.C. DSM-IV: Diagnostic and Statistical Manual of Mental Disorders. *JAMA* **1994**, 272, 828–829. <https://doi.org/10.1001/jama.1994.03520100096046>.
33. Sheehan, D.V.; Lecrubier, Y.; Sheehan, K.H.; Amorim, P.; Janavs, J.; Weiller, E.; Hergueta, T.; Baker, R.; Dunbar, G.C. The Mini-International Neuropsychiatric Interview (MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J. Clin. Psychiatry* **1998**, 59, 22–33.
34. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 2613–2617. <https://doi.org/10.21437/Interspeech.2019-2680>.
35. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980. <http://arxiv.org/abs/1412.6980>.

36. Das, A.K.; Naskar, R. A deep learning model for depression detection based on MFCC and CNN generated spectrogram features. *Biomed. Signal Process. Control* **2024**, *90*, 105898. <https://doi.org/10.1016/j.bspc.2023.105898>.
37. Zhao, M.; Gao, H.; Zhao, L.; Wang, Z.; Wang, F.; Zheng, W.; Li, J.; Liu, C. Decoupled Multi-Perspective Fusion for Speech Depression Detection. *IEEE Trans. Affect. Comput.* **2025**, *16*, 1772–1786. <https://doi.org/10.1109/TAFFC.2025.3538519>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.