Gram-based Attentive Neural Ordinary Differential Equations Network for Video Nystagmography Classification

Xihe Qiu[1]*, Shaojie Shi[1]*, Xiaoyu Tan[2]*†, Chao Qu[2], Zhijun Fang[3],
Hailing Wang[1], Yongbin Gao[1]†, Peixia Wu[4], Huawei Li[4]
[1]Shanghai University of Engineering Science, Shanghai, China
[2]INF Technology (Shanghai) Co., Ltd. Shanghai, China
[3]Donghua University, Shanghai, China
[4]Eye & ENT Hospital of Fudan University, Shanghai, China
yulin.txy@inftech.ai, gaoyongbin@sues.edu.cn

Abstract

Video nystagmography (VNG) is the diagnostic gold standard of benign paroxysmal positional vertigo (BPPV), which requires medical professionals to examine the direction, frequency, intensity, duration, and variation in the strength of nystagmus on a VNG video. This is a tedious process heavily influenced by the doctor's experience, which is error-prone. Recent automatic VNG classification methods approach this problem from the perspective of video analysis without considering medical prior knowledge, resulting in unsatisfactory accuracy and limited diagnostic capability for nystagmographic types, thereby preventing their clinical application. In this paper, we propose an end-to-end data-driven novel BPPV diagnosis framework (TC-BPPV) by considering this problem as an eye trajectory classification problem due to the disease's symptoms and experts' prior knowledge. In this framework, we utilize an eye movement tracking system to capture the eye trajectory and propose the Gram-based attentive neural ordinary differential equations network (Gram-AODE) to perform classification. We validate our framework using the VNG dataset provided by the collaborative university hospital and achieve state-of-the-art performance. We also evaluate Gram-AODE on multiple open-source benchmarks to demonstrate its effectiveness in trajectory classification. Code is available at https://github.com/XiheQiu/Gram-AODE.

*Equal contribution
†Corresponding author

1. Introduction

Benign paroxysmal positional vertigo (BPPV) is a classic type of otogenic vertigo, which is transitory dizziness caused by head movements to a certain position with high prevalence (i.e, the lifetime prevalence of 2.4%) [46]. The characterization of nystagmus in individuals with BPPV is crucial for the diagnosis of BPPV. In clinical practice, information on nystagmus is typically obtained through positional testing and the type of BPPV condition can be determined by irregular rhythmic eye movements, consisting of nystagmus and eye-twisting movements [44]. Video Nystagmography (VNG) can significantly improve the detection rate of BPPV nystagmus [27]. According to the type of nystagmus movement, medical professionals might classify BPPV into different types of therapy. However, distinguishing the different disorders can be a challenging task. During diagnosis, the medical professional must examine the VNG video for nystagmus features such as direction, frequency, intensity, duration, and intensity variation, which is time-consuming and error-prone.

Recent research has demonstrated the success of using deep learning to comprehend human behavior in the video, and automatically classify VNG for BPPV analysis based on relevant image [29, 3] or time-series features [18]. However, these models are proposed from the perspective of video analysis without explicitly considering the disease symptoms and pathology, which provide sufficient classification types in clinical practice.

Therefore, we propose a novel end-to-end data-driven framework to perform **BPPV** diagnosis by considering this problem as an eye **T**rajectory **C**lassification (TC-BPPV) problem due to the disease's symptoms and experts' prior knowledge. This framework can classify the BPPV into ten classes which are commonly recognized in clinical practice. In this framework, we first simplify the complex video

analysis into a trajectory classification task through the eye movement tracking technique. This is achieved by retrieving the pupil center for every nystagmus video in the dataset using the contour detection algorithm [6] and the Hough circle transformation algorithm to produce pupil coordinates [2]. This process dramatically reduces the feature learning space by following the typical clinical diagnosis prior knowledge, whereas the pupil movement trajectory is one of the strongest diagnosis evidence [16].

Then, we perform classification through a **Gram**-based **A**ttentive neural **O**rdinary **D**ifferential **E**quations network (Gram-AODE). We regularize the obtained coordinates to the Gram matrix through a carefully designed production process. This process not only ensures the preservation of the translation equivariant property which is coherent with clinical diagnosis, but also provides detailed relational information in the feature map and overcomes intricate temporal patterns. Subsequently, we process the acquired feature map through a neural ordinary differential equations (ODEs) network to explicitly consider the continuous changes between the trajectory points and implicitly model the underlying dynamics of eye movement. The ablation studies also show that the discrete residual network is ineffective to process the feature image due to the difficulty of choosing the specific number of residual layers, which also supports our utilization of neural ODEs with their superior adaptive resolution capability. Finally, we employ an attention mechanism to perform feature integration in the hidden feature space and output the final classification results. The overview framework is illustrated in Figure 1.

To fully evaluate our proposed framework, we first conduct a comprehensive evaluation of TC-BPPV on the clinical dataset provided by the collaborative university hospital. This dataset consists of real patients' VNG videos that are manually labeled by medical professionals. Due to the trajectory classification nature of our proposed Gram-AODE method, we also evaluate the method on several publicly accessible trajectory classification benchmarks to demonstrate its effectiveness. Based on the experiment results, our proposed TC-BPPV can perform clinical BPPV diagnosis in an end-to-end data-driven manner with human expert-level accuracy. The Gram-AODE method also consistently outperforms some existing state-of-the-art baseline approaches in some open-source trajectory classification benchmarks. Our main contributions are summarized as follows:

- We present TC-BPPV through eye trajectory classification, which is end-to-end and data-driven. According to our knowledge, we are the first to handle the challenging VNG classification as a time-series trajectory classification problem for BPPV diagnosis using machine learning techniques.
- Our proposed model explicitly leverages the inductive bias based on the disease's symptoms and experts' prior knowledge. We consider the BPPV as eye trajectory classification and express the translation equivariant property of the point set.
- We develop a Gram-based attentive neural ordinary differential equations network (Gram-AODE) to perform time-series trajectory classification. This method can overcome intricate temporal patterns and learn the points' inner relation with underlying dynamics in a higher dimensional space to provide an accurate prediction.
- We conduct rigorous evaluations of TC-BPPV on the clinical dataset provided by the collaborative university hospital. The results indicate high feasibility in clinical practice. We also test Gram-AODE in several open-source trajectory classification benchmarks, which demonstrate state-of-the-art performance.

2. Related Work

Deep learning techniques have been applied to automatically classify VNGs for BPPV analysis [17]. [29, 3] extract image-related features for disorder identification, while [24] distinguishes VNGs through multiple ad hoc strategies. However, these approaches are provided from the perspective of video analysis, which lacks consideration of disease symptoms, pathology, and appropriate categorization types for clinical practice. In reality, due to the diseases' symptoms and experts' prior knowledge, the complex video analysis can be simplified into an eye trajectory classification problem through the eye movement tracking technique, which in nature belongs to a time-series classification (TSC) problem. TSC involves monitoring a process at regular intervals and can be divided into univariate time series classification (UTSC) [41, 31, 1] and multivariate time series classification (MTSC) [9, 26] based on the dimensions of the collected data. Approaches towards solving UTSC include symbolic aggregate approximation (SAX) [35, 21, 25] and shapelet-based methods [49, 23]. Deep neural networks such as ShapeNet and TapNet have also shown promise [22, 48]. MTSC can be addressed by combining univariate classifiers such as ROCKET[9] and HIVE-COTE [26]. However, handling the challenging VNG classification as a time-series trajectory classification problem, transforming time-series data into feature images, and capturing torsional nystagmus patterns in fine-grained eye movements for BPPV analysis are underexplored in the literature and pose a significant challenge.

3. Methods

Figure 1 provides an overview of our proposed TC-BPPV framework for VNG diagnosis and Gram-AODE

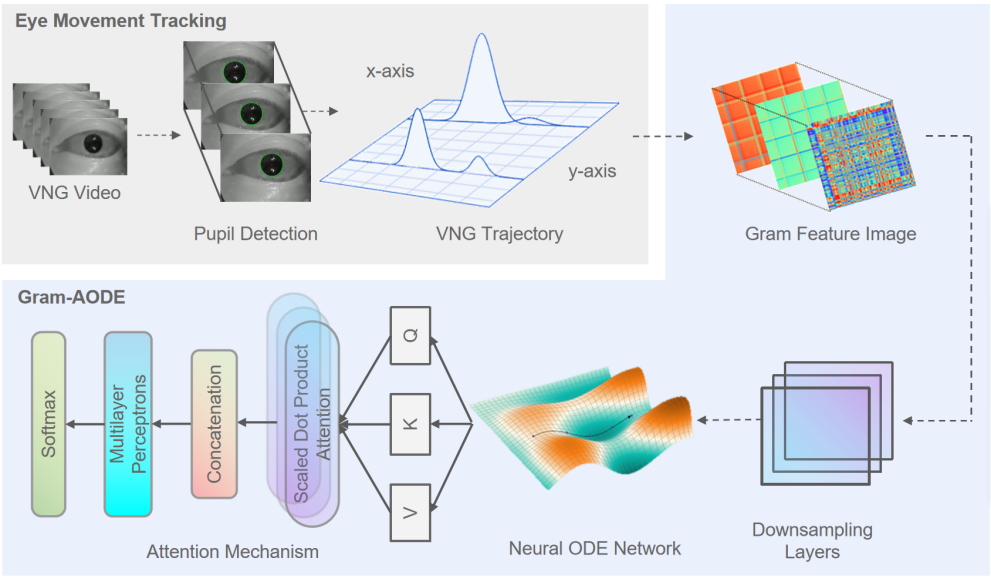

Figure 1. An overview of our proposed framework

method for time-series trajectory classification. In this section, we first present the VNG data pre-processing techniques for detecting eye movement and acquiring trajectory. Then, we introduce the Gram matrix conversion method for converting time-series data into a more stable and low-variance Gram matrix. After that, we describe neural ordinary differential equations network to learn the underlying dynamics and continuous changes of trajectory points. Finally, we elaborate on our designed attention mechanism to integrate hidden features and perform classification. This section presents the VNG data pre-processing techniques for detecting eye movement and acquiring trajectory data.

3.1. VNG data pre-processing

The VertiGoggles R ZT-VNG-II records original VNG videos with a frame size of 640x480 and a frame rate of 60. To prepare the dataset, abnormal and interfering frames were deleted, and pupil position was extracted for each frame using the contour detection and Hough circle transformation algorithms [6], saving it in chronological order with the form of coordinates. This procedure facilitates data augmentation to reduce data dimensionality in low-data regimes through the integration of prior knowledge.

3.2. Gram matrix conversion of time series

Time series data is collected at regular intervals, has temporal characteristics, and presents different regular distributions according to the time order. The classification category of a time series is determined by its temporality, and events of the same type should have similar features. This applies to clinical practice where patients with the same type of BPPV exhibit similar eye movement patterns.

Time series data can be uni-variate or multi-variate [12]. The eye movement trajectory obtained from VNG preprocessing is a multi-variate time series with two coordinates. Unfortunately, the direct classification of such time-series data is challenging due to the intricate temporal variations. Furthermore, to perform more efficient and accurate classification, the property of translation equivariant should be explicitly leveraged in the model learning process as an inductive bias to further reduce learning space complexity and improve convergence. The methodology for diagnosing BPPV via VNG suggests that eye movement patterns and trends are more strongly associated with BPPV type than eye position.

The Gram matrix measures vector similarity and is useful in machine learning tasks such as style transfer and time-series classification [42, 47]. To improve performance, normalization, and wise correlations can be applied. We are inspired by Wang *et al.* [43], who use the Gram matrix to transform time-series data into feature images and CNN for categorization, providing a solution to TSC.

Hence, we perform Gram matrix conversion on the acquired time-series data to represent the inner point relations expressively and reserve the translation equivariance property. A Gram matrix is a Hermitian matrix in the inner product space of a set of vectors, with values determined by the vector inner product of the elements [15].

Set $T = \{t_1, t_2, t_3, \cdots, t_N\}$ for a length of N sequence, where each element is arranged in the order of acquisition time, and $t_i \in \mathbb{R}$ denotes the i-th value. When collecting time series data in real-world scenarios, the collected data may not have the same scale. We follow the normalization formula to convert all elements in set T in or-

der to make the weights of each feature dimension of the data consistent. The normalization formula is given by $t_i' = t_i - t_{\min}/(t_{\max} - t_{\min})$, where t_i' is the normalized i-th value, $t_{\min}$ is the minimum value of T, and $t_{\max}$ is the maximum value of T.

By considering the time-series nature and potential cyclical patterns of eye movement trajectory, we believe that the Polar coordinate can better preserve data temporal correlations compared with the Cartesian coordinate. A given time series data T exists one and only one polar coordinate mapping and inverse mapping. Thus, the time series T' can be mapped according to the following rules:

$$\theta_i = \arccos\left(t_i'\right), r_i = i/N, \tag{1}$$

where θ_i is the angle and r_i is the radial distance. N can be used as a constant factor that adjusts the span of the polar coordinate system and is also numerically equal to the length of the time series data. In polar coordinates, the r_i denotes the temporality, while the θ_i denotes the numerical change of original T'. This property provides a theoretical basis for transforming uni-variant time series data into 2D feature images by using the following Gram matrix for classification.

Typically, the elements of the Gram matrix are determined by the vector inner product, and the Gram matrix generally takes the following form:

$$G = \begin{pmatrix} \langle v_1, v_1 \rangle & \dots & \langle v_1, v_n \rangle \\ \vdots & \ddots & \vdots \\ \langle v_n, v_1 \rangle & \dots & \langle v_n, v_n \rangle \end{pmatrix}, \tag{2}$$

where v represents sequential vectors and $\langle \cdot, \cdot \rangle$ denotes the inner product operation. Here, we replace the inner product operation using $\cos(\theta_i + \theta_j)$ on data point i and j to get a stable transformation with lower variance. It is worth to know that $\cos(\theta_i + \theta_j) = \cos\theta_i \cos\theta_j - \sin\theta_i \sin\theta_j = t_i t_j - \sqrt{1 - t_i^2}\sqrt{1 - t_j^2}$. Hence, it can effectively represent the inner relationships between all data points. After Gram matrix transformation, each element of the Gram matrix can be regarded as a pixel of the feature image and the whole sequence data can be transformed into the image which retains comprehensive sequence information.

3.3. Neural ODEs Network

After we acquire the feature image in a modified Gram matrix transformation, we process the feature image using a neural ODE network to model the underlying dynamics of trajectory points and learn the relation of continuous changes[28]. Neural ODE connects neural networks and differential equations to model continuous time data and produce normalized flows [10, 39]. Augmented neural ODEs represent complex functions and learn complex functions from input to augmented space features [8]. Optimal transmission and regularization strategies reduce neural ODE's computational overhead [11]. Rodriguez *et al.*[30] propose LyaNet for rapid convergence and robustness. Moreover, the neural ODE network can automatically adjust the resolution based on the input feature image, which demonstrates high efficiency[10].

Neural ODE follows typical ODE with general form:

$$h(0) = h_0, \quad \frac{\mathrm{d}h}{\mathrm{d}t}(t) = f_\theta(t, h(t)), \tag{3}$$

where h_0 denotes initial vector, θ represents learning parameters, and f_ω denotes neural networks with parameter ω. Neural ODEs can be viewed as a continuous equivalent of residual networks [45, 13, 33, 7]. In a network with residual structure, the transition of the hidden state between t and $t+1$ layers can be expressed as $h_{t+1} = h_t + f_t(h_t)$. If time t is a continuous variable and the time step $\Delta t = 0$, we can obtain the differential form:

$$\lim_{\Delta t \to 0} \frac{h_{t+\Delta t} - h_t}{\Delta t} = \frac{\mathrm{d}h(t)}{\mathrm{d}t} = f(h(t), t). \tag{4}$$

Thus, the hidden state is transformed into an ODE, the discrete network form is converted into a continuous state, and the hidden state at any time step can be acquired by solving the initial value problem. The hidden state $h(t)$at time t is the feature to be learned by the model. For example, if the start time is t_0, the end time is t_1, and the initial state is $h(t_0)$, then the state $h(t_1)$at the time $t_1$can be expressed as:

$$h(t_1) = h(t_0) + \int_{t_0}^{t_1} f(h(t), t; \theta) dt. \tag{5}$$

In contrast to a typical neural network, the neural ODE network predicts y by solving the ODE, which can continuously adjust the dynamics of the system so that the output is constantly close to ground truth. The neural ODE network is optimized using the ODE Solver [8], which computes gradients via the adjoint method.

3.4. Kernel Attention Mechanism

After the processing of the neural ODE network, we integrate the learned hidden features and provide the final prediction through a designed vision attention mechanism. This mechanism can significantly reduce the computation complexity and provide even more accurate predictions. We start with the typical multi-head attention.

Multi-head attention [40] is a common attention mechanism that allows the model to learn different attention representations in different attention subspaces. Given query $Q \in \mathbb{R}^{n \times d_k}$, keys $K \in \mathbb{R}^{m \times d_k}$, values $V \in R^{m \times d_v}$, and the attention weights W, the common multi-head self-attention is computed as follow:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\mathrm{V}, \tag{6}$$

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \quad (7)$$

$$\text{MultiHead}(Q, K, V) = (\text{head}_1 || \ldots || \text{ head }_h) W_o, \quad (8)$$

where $||$ denotes the concatenation. In order to better apply Multi-head attention to image tasks and reduce computing resource consumption, [4] proposed a low-rank linearized multi-head attention, which can effectively reduce the time complexity and improve the accuracy. We define a kernel function $\phi(\cdot)$ instead of the softmax function. If this kernel function is decomposable, then we can obtain

$$\text{Attention}(Q, K, V; \phi) = \left(\phi(Q)\phi(K)^T\right) V. \quad (9)$$

In this way, we can first compute $\phi(K)^T\ V$, thus changing the time complexity from $O\left(HT^2D/H\right)$to $O\left(HT(D/H)^2\right)$. When head H equals feature dimensions, the time complexity reduces to $O\left(HT\right)$. In our implementation, we utilize the cosine similarity function as our kernel function. The cosine similarity function used in the kernel attention is $\phi(x) = x/\|x\|_2$.

3.5. Gram-AODE

For a sample j with total number of video frames N, in the i frame pupil coordinates for (x_i^j, y_i^j), we use $X_j = (x_1^j, \ldots, x_n^j)$ to represent change of pupil position in x-axis, and $Y_j = (y_1^j, \ldots, y_n^j)$ to represent the change of pupil position in the y-axis. G denotes Gram transformation introduced in Section 3.2, and the input image feature matrix with label z_j can be expressed as $M_j = (G(X_j)||G(Y_j))$ with channel dimension concatenation.

The acquired input image feature is processed using down-sampling layers $g(M_j)$ that reduce the spatial dimensions of the feature map while keeping the channel dimension intact. This is typically performed to decrease computational complexity and compress spatial information into a smaller space while maintaining crucial features [7]. Next, the new feature map is passed through a Neural ODE network block introduced in Section 3.3, which is designed to capture the underlying dynamics and continuous changes in the hidden feature. The Neural ODE block $h_j = f(g(M_j))$ treats the state evolution as a differential equation which allows the model to learn the dynamics of the input data in a continuous and differentiable manner, thereby enabling better generalization and prediction on our task. After that, the hidden vector h_j is linearized and processed using the kernel multi-head attention block introduced in Section 4.4.1 to generate the final prediction $\hat{z}_j = \text{MultiHead}(h_j)$. The attention mechanism allows the model to concentrate on relevant areas of the input feature map and learn to attend to multiple aspects of the feature channels concurrently.

The entire network architecture can be trained end-to-end by implementing an ODE solver [8] and SGD optimizer [5], along with the cross-entropy loss on all K samples, which is commonly used in classification tasks to measure the difference between the predicted and true labels:

$$\mathcal{L}(z, \hat{z}) = -\sum_{j=1}^{K} z_j \log(\hat{z}_j). \quad (10)$$

4. Experiments

In this section, we present an evaluation of our proposed TC-BPPV framework on a dataset comprising VNG videos, which has been provided by the collaborative university hospital. The dataset includes clinical diagnostic results and various diagnostic indicators for each patient. We believe that this dataset will enable a comprehensive evaluation of our proposed framework, and effectively demonstrate its clinical feasibility. For comparison purposes, we have selected three state-of-the-art (SOTA) methods in various applications as baseline models. Firstly, we adopt the 3D-ResNet101 [14, 38], a deep learning model for video classification, as our initial baseline method, which performs video analysis directly on raw video inputs. Secondly, we employ another SOTA method, ROCKET [9], for time-series classification to demonstrate the baseline performance of eye movement trajectory classification. Thirdly, we implement ResNet [14], which is a SOTA image classification method, to show the baseline performance of Gram feature image classification.

Next, we evaluate the capabilities of our proposed Gram-AODE method in time-series trajectory classification tasks. We choose ROCKET [9], HIVE-CORE [26], and ResNet50 [14] with Gram feature image as our baseline methods and test in 25 types of UCR open-source benchmark datasets to demonstrate the general efficacy of our proposed method.

Finally, we provide comprehensive ablation studies to test the influence of attention block, layer number of feature image classification, and various choices of ODESolvers.

4.1. Experiment Setup

The VNG video dataset: The VNG video dataset contains 646 videos of nystagmus movements from different patients, which were recorded in real-world clinics at our collaborative university hospital. The dataset is labeled by medical professionals with respect to the four-movement features of nystagmus: horizontal, vertical, axial, and intensity change, representing the nystagmus pattern of each patient in the video clip. It is noteworthy that numerous complex nystagmus patterns exist; however, only the ten most common nystagmus patterns with clinical relevance are included. The number of occurrences for each nystagmus pattern is 47, 49, 51, 54, 58, 63, 63, 69, 73, and 119,

respectively, with a median of 58 and a mode of 63. To evaluate the performance of our proposed models, we randomly partitioned the 646 VNG videos into 502 videos for training, 72 videos for validation, and 72 videos for testing.

For a specific patient's nystagmus video clip, the following diagnostic indicators are labeled by medical professionals based on their clinical practice: nystagmus intensity variation and horizontal/vertical/axial nystagmus direction. The details are demonstrated in Table 1. For model evaluation, we test the model's performance comprehensively on both the nystagmus type classification and the four diagnostic indicator classifications.

Label	Intensity	Horizontal	Vertical	Axial
0	stronger	left	up	clockwise
1	weaker	right	down	opposite
2	N/A	N/A	N/A	N/A

Table 1. The labels of four types of diagnostic indicators

The UCR datasets: We employ the publicly available UCR dataset, containing uni-variate time series data, for our study. Data normalization is conducted to prepare the input feature matrix, followed by Gram transformation for methods involving feature image classification. The dataset is randomly partitioned into three subsets: training set (80%), validation set (10%), and test set (10%), respectively. For training and evaluation of the models, we employed cross-entropy loss function on all methods.

We evaluate the efficacy of our proposed approach using established metrics typically used in time-series classification tasks. As the VNG video dataset involves multiple class types, we pay particular attention to Top-k accuracy measures, including top-1 accuracy and top-3 accuracy. These measures are essential in mitigating the issue of inaccurate labeling of the data and fully evaluating the performance of our proposed method.

4.2. VNG Experiment Results

Table 2 shows the performance of our proposed TC-BPPV framework on the VNG dataset. We implement ResNet [14] with a single layer of residual blocks on Gram feature image, ROCKET with default setting introduced in [9], and 3D-ResNet101 [14, 38] as our baseline methods.

Our proposed TC-BPPV approach outperforms all other classification methods, including the second-best Gram-ResNet, with an average absolute improvement of approximately 4%. Specifically, our proposed method achieves a Top-1 accuracy of 85.47% and a Top-3 accuracy of 97.64% for the ten-class classification task, facilitating the identification of three possible BPPV types and assisting medical experts in decision-making. By contrast, ROCKET only achieves a Top-1 accuracy of 64.49% in the ten-class classification task, indicating that the intricate temporal patterns in the eye movement trajectory cannot be effectively alleviated using the original dimension without explicitly considering the internal point relations. We observed that the 3D-ResNet101 model only achieves a Top-1 accuracy of 33.66%, potentially due to its large learning space and limited data. The model failed to learn effective patterns in the VNG videos without providing any inductive bias to reduce the learning space. In contrast, our approach reliably classifies the types of BPPVs, exhibiting high recognition accuracy for each type through the use of prior knowledge of eye movement trajectories and equivariant properties, assisting medical professionals in more accurate diagnoses.

4.3. UCR Experiment Results

We evaluate the generalization capability of our proposed Gram-AODE method in time-series trajectory classification by comparing it with HIVE-COTE with settings introduced in [26], ROCKET, and ResNet on Gram feature image which is used in Section 4.2. Since our proposed Gram-AODE method leverages the translation equivariant property which may not be satisfied in other benchmarks, we implement Gram-AODE(2) for evaluation. Compared with Gram-AODE, this method incorporates another average value matrix which replaces the Gram $\langle \cdot, \cdot \rangle$ operation with $\mathrm{mean}(t_i, t_j) = (t_i + t_j)/2$. This matrix will be concatenated with the Gram matrix on the third dimension and processed by the subsequent models.

Table 3 shows that both Gram-AODE and Gram-AODE(2) outperform the prior state-of-the-art classical approaches, HIVE-COTE and ROCKET, on 20 datasets while achieving comparable results on 4 datasets. Our results demonstrate that explicitly representing the points relation of time-series data is generally helpful in mitigating the intricate temporal patterns, which can be learned in a higher-dimensional space. Compared to the classical approaches, our proposed methods are able to effectively model these complex patterns, leading to improved performance. In addition, comparing the results of Gram-ResNet, we can observe that the neural ODE can effectively learn the underlying dynamic changes in the trajectory data, which can generally improve the classification performance. These findings further highlight the potential of our proposed approaches for time-series classification tasks.

4.4. Ablation Studies

4.4.1 Attention Mechanism and the Cosine Operator

In this section, we present an ablation study on the attention mechanism proposed in Section of our framework. We compare the performance of our modified method, which we refer to as Gram-ODE, against our proposed approach, Gram-AODE, by removing the attention mechanism block from our framework. We perform a comprehensive evaluation of our ablated method on all experiments using both the

Model	Top-1 (10 class)	Top-3 (10 class)	Top-1 (5 class)	horizontal	vertical	axial	intensity
TC-BPPV	**85.47**	**97.64**	**93.76**	**93.47**	**97.59**	**98.97**	**96.42**
ROCKET	64.49	77.72	75.23	78.26	79.56	81.34	76.92
3D-ResNet101	33.66	52.51	40.83	55.27	43.01	45.54	41.28
Gram-ODE	82.56	96.56	90.30	91.89	94.38	98.49	95.40
Gram-ResNet(1)	81.05	95.38	89.79	92.63	92.42	95.40	90.30
Gram-ResNet(6)	80.21	96.55	88.77	92.41	91.68	91.83	89.68
Gram-ResNet(18)	78.74	96.10	87.75	91.57	89.28	88.26	87.78

Table 2. Comparative results in accuracy (%) of video nystagmography classification for ten classes, five classes, and four diagnosis indicators in BPPV disorders.

DataSets	HIVE-COTE	Gram-ResNet	ROCKET	Gram-ODE	Gram-AODE	Gram-AODE(2)
ChlorineConcentration	75.49	80.76	72.48	99.32	**99.76**	99.20
Herring	61.31	62.73	63.94	74.60	76.90	**76.92**
DiatomSizeReduction	90.12	90.93	93.64	98.45	99.24	**99.73**
DistalPhalanxTW	70.40	66.61	69.49	66.56	75.90	**79.62**
BirdChicken	94.24	90.28	95.86	99.21	**99.81**	99.27
SonyAIBORobotSurface1	86.72	93.36	82.20	97.25	98.41	**99.10**
Chinatown	96.29	93.93	95.70	99.11	99.57	**99.87**
MoteStrain	94.14	90.82	92.36	96.09	96.37	**98.43**
ItalyPowerDemand	96.56	96.24	96.33	99.09	**99.91**	95.45
MiddlePhalanxTW	58.22	57.19	56.76	53.50	62.50	**73.21**
ProximalPhalanxTW	80.85	78.67	81.11	81.96	83.60	**86.88**
Symbols	95.32	90.51	95.31	95.09	**99.60**	99.06
SonyAIBORobotSurface2	91.19	98.11	92.94	98.97	**99.77**	98.97
SwedishLeaf	94.27	96.05	95.12	86.72	**97.56**	97.34
MiddlePhalanxOutlineAgeGroup	73.53	71.14	72.10	69.64	**75.28**	76.78
GunPoint	97.91	98.34	98.57	99.29	**99.69**	95.00
BeetleFly	91.91	85.79	**100**	75.19	**100**	**100**
BME	99.44	**100**	98.88	72.00	**100**	**100**
ECGFiveDays	**100**	95.43	99.13	99.48	**100**	**100**
FacesUCR	95.56	95.61	96.41	93.45	96.56	**99.55**
MedicalImages	77.84	79.80	74.61	74.07	79.10	**81.73**
PowerCons	**98.89**	91.11	97.69	94.44	97.22	97.23
GunPointAgeSpan	**100**	99.68	99.89	97.83	**100**	97.83
FaceAll	98.30	99.17	98.79	96.76	96.88	**99.55**
Car	87.69	86.24	85.96	50.96	83.34	**97.28**

Table 3. Accuracy (%) classification results of different approaches on various UCR benchmark datasets.

VNG dataset and UCR datasets. Our results are presented in Table 2 and Table 3. Our findings demonstrate that while Gram-ODE achieves comparable results with Gram-AODE in VNG classification tasks, it fails to perform as well in several UCR tasks. These findings underscore the importance of the attention mechanism for feature integration in our proposed framework, highlighting the critical role that attention mechanisms play in our proposed method.

We conduct ablations under identical settings to compare Hydra attention with typical self-attention and two-layer MLPs. The preliminary results in Table 4 demonstrate that Hydra achieves the highest accuracy. Additionally, we highlight the essential role of the cosine operator in our framework.

Method	Hydra	Self-attention	MLPs	W/o. Cosine
Accuracy	**85.47%**	83.24%	80.42%	73.38%
Iter/sec	6.73	6.11	7.56	-

Table 4. Ablation results of VNGs ten-class classification with Top-1 accuracy (%).

4.4.2 Discrete Residual Layers and Neural ODEs

We conducted an ablation analysis to investigate the effect of the number of residual layers on Gram feature image classification. This analysis aims to provide insight into the complexity of the features generated by Gram matrix conversion and the importance of incorporating neural ODE blocks. Specifically, we replace the neural ODE block in Gram-AODE with varying numbers of residual layers (i.e.,

1, 6, and 18 layers) and evaluate the performance in 10-class and 5-class VNG classification tasks. The findings presented in Table 2 reveal that the ResNet model increasingly overfits the dataset as the number of residual layers increases. This suggests that the patterns underlying the Gram feature image cannot be effectively learned using a discrete residual flow, and even a single residual layer fails to achieve optimal performance. In contrast, neural ODEs model the continuous flow and effectively capture the underlying dynamics in the trajectory, resulting in superior classification performance.

4.4.3 The Effect of Simplifying VNG Classification

In this ablation study, we investigate the impact of two different approaches for VNG data classification: (1) direct classification of raw VNG data as eye movement trajectories without Gram transformation, and (2) direct classification of VNG data from the perspective of video analysis without medical prior information. Our aim is to gain insight into the importance of the Gram transformation and the medical prior information for precise VNG data classification. Here, we report the Top-1 results of ten-class classification accuracy.

Method	ResNet50	ROCKET	HIVE-COTE
Accuracy	46.01	64.49	57.48

Table 5. Top-1 accuracy (%) results of time series classification without Gram transformation

Table 5 shows the results of time-series sequence classification on the VNG dataset without the use of the Gram transformation. The experimental evaluation compares the performance of three SOTA baselines: ResNet50 applied to the trajectory images, ROCKET, and HIVE-COTE applied to the trajectory data. The results reveal that it is not appropriate to directly approach the VNG classification task as a time-series trajectory classification problem. Without the Gram transformation, it becomes difficult to distinguish between sequences, and the translation equivariant property is hard to exploit. These findings suggest that the Gram transformation is a necessary step for the VNG classification task, as it can more effectively capture the inner relationships between data points and learn the features inherent in the data.

Method	3D-MobileNetV2	3D-ResNet50	3D-ResNet101
Accuracy	27.93	30.56	33.66

Table 6. The classification results in accuracy (%) for ten classes of BPPV disorders using 3D-CNN methods

Table 6 presents a comparison of the classification performance of three state-of-the-art 3D-CNN[38] approaches, namely MobileNetV2 [34], ResNet50, and ResNet101, in analyzing VNG videos from a video analysis perspective. The lightweight networks MobileNetV2, ResNet50, and ResNet101 with large sum parameters are selected and transformed into 3D-CNNs [38] for evaluation on our original VNG dataset. 3D-CNNs are a traditional algorithm that has achieved success in various video classification applications, especially in addressing the action recognition problem. By performing 3D convolution in both spatial and temporal dimensions, it can effectively capture motion information contained in the video stream.

The classification of VNGs can be viewed as an action recognition problem within the domain of video comprehension. Despite this fact, our experimental results indicate that conventional 3D-CNN approaches (i.e., MobileNetV2, ResNet50, and ResNet101) are inadequate for VNG video classification, as demonstrated in Table 6. In practice, action recognition usually depends on prior knowledge. However, due to limited information in a single VNG video scene, target similarity, and moderate eye movement amplitude, it is challenging to provide useful prior knowledge for training. The VNG dataset used in our study is derived from authentic medical scenarios. Thus, it has fewer data samples compared to other public datasets such as HMDB51 [19], Kinetics [36], and UCF101 [37]. Therefore, training a VNG classification model is more difficult. Consequently, it is noteworthy that video comprehension alone is insufficient for successfully classifying VNG videos.

4.4.4 Different Neural ODE Solver

Various modern ODE solvers are available to tackle real-world problems. We investigate the impact of three variants of the Neural ODE Solver (i.e., Euler, Midpoint, and 5th order Runge-Kutta [32, 20]) on our framework and report their performance in Table 7. Notably, for the VNG classification problem, the Runge-Kutta method demonstrates the highest accuracy but the slowest training speed. In contrast, the Euler method exhibits the fastest training speed, while the midpoint method falls in between the other two methods in terms of performance.

Solver	Accuracy(10)	Accuracy(5)	Duration
Euler	80.84	81.63	39.66
Runge-Kutta	**82.56**	**90.30**	8.59
Midpoint	79.36	87.75	36.40

Table 7. The classification results in accuracy (%) of video nystagmography classification for ten classes, five classes, and training speed (iteration/second) for different ODE solvers

5. Conclusion

In this paper, we conduct an investigation into VNG classification to aid in the clinical diagnosis of BPPV. The proposed framework for BPPV diagnosis (i.e., TC-BPPV), is

an end-to-end data-driven diagnosis framework comprising two stages. The first stage tracks the eye trajectory using an eye movement tracking system, while the second stage involves the classification of the trajectories. To accurately classify the trajectories, we propose Gram-AODE to first perform Gram matrix transformation, and then use neural ODEs network following designed attention mechanism to perform prediction. This method can effectively mitigate intricate temporal patterns in trajectories, consider the equivariant property of disease symptoms, learn the underlying dynamics, and integrate the hidden feature. We extensively evaluate our approach using the clinical dataset provided by our collaborative university hospital, and multiple open-source benchmarks. Our method achieves a substantial improvement in performance compared to SOTA models. This indicates that our proposed approach is applicable in real-world clinical scenarios.

6. Acknowledgement

This work is supported by Shanghai Committee of Science and Technology "Science and Technology Innovation Action Plan" Natural Science Foundation of Shanghai, Shanghai Science and Technology Commission for Social Development Project, and the National Natural Science Foundation of China (Grant No.23ZR1425400, 21DZ1204900, 62102241).

References

[1] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery*, 31:606–660, 2017. 2

[2] Dana H Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern recognition*, 13(2):111–122, 1981. 2

[3] Amine Ben Slama, Aymen Mouelhi, Hanene Sahli, Abderrazek Zeraii, Jihene Marrakchi, and Hedi Trabelsi. A deep convolutional neural network for automated vestibular disorder classification using vng analysis. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 8(3):334–342, 2020. 1, 2

[4] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, and Judy Hoffman. Hydra attention: Efficient attention with many heads. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 35–49. Springer, 2023. 5

[5] Léon Bottou. Stochastic gradient descent tricks. *Neural Networks: Tricks of the Trade: Second Edition*, pages 421–436, 2012. 5

[6] Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000. 2, 3

[7] Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019. 4, 5

[8] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018. 4, 5

[9] Angus Dempster, François Petitjean, and Geoffrey I Webb. Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495, 2020. 2, 5, 6

[10] Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. *Advances in neural information processing systems*, 32, 2019. 4

[11] Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam Oberman. How to train your neural ode: the world of jacobian and kinetic regularization. In *International conference on machine learning*, pages 3154–3164. PMLR, 2020. 4

[12] Yash Garg and Silvestro Roberto Poccia. On the effectiveness of distance measures for similarity search in multivariate sensory data. In *ICMR*, 2017. 3

[13] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34(1):014004, 2017. 4

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 6

[15] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012. 3

[16] Ji-Soo Kim and David S Zee. Benign paroxysmal positional vertigo. *New England Journal of Medicine*, 370(12):1138–1147, 2014. 2

[17] Sheng Kong, Zheming Huang, Weike Deng, Yinwei Zhan, Jujian Lv, and Yong Cui. Nystagmus patterns classification framework based on deep learning and optical flow. *Computers in Biology and Medicine*, page 106473, 2022. 2

[18] Sheng Kong, Zheming Huang, Weike Deng, Yinwei Zhan, Jujian Lv, and Yong Cui. Nystagmus patterns classification framework based on deep learning and optical flow. *Computers in Biology and Medicine*, 153:106473, 2023. 1

[19] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 8

[20] Wilhelm Kutta. *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen*. Teubner, 1901. 8

[21] Thach Le Nguyen, Severin Gsponer, and Georgiana Ifrim. Time series classification by sequence learning in all-subsequence space. In *2017 IEEE 33rd international conference on data engineering (ICDE)*, pages 947–958. IEEE, 2017. 2

[22] Guozhong Li, Byron Choi, Jianliang Xu, Sourav S Bhowmick, Kwok-Pan Chun, and Grace Lai-Hung Wong. Shapenet: A shapelet-neural network approach for multivariate time series classification. In *Proceedings of the*

AAAI Conference on Artificial Intelligence, volume 35, pages 8375–8383, 2021. 2

[23] Guiling Li, Wenhe Yan, and Zongda Wu. Discovering shapelets with key points in time series classification. *Expert systems with applications*, 132:76–86, 2019. 2

[24] Eun-Cheon Lim, Jeong Hye Park, Han Jae Jeon, Hyung-Jong Kim, Hyo-Jeong Lee, Chang-Geun Song, and Sung Kwang Hong. Developing a diagnostic decision support system for benign paroxysmal positional vertigo using a deep-learning model. *Journal of clinical medicine*, 8(5):633, 2019. 2

[25] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Mining and knowledge discovery*, 15:107–144, 2007. 2

[26] Jason Lines, Sarah Taylor, and Anthony Bagnall. Time series classification with hive-cote: The hierarchical vote collective of transformation-based ensembles. *ACM transactions on knowledge discovery from data*, 12(5), 2018. 2, 5, 6

[27] Camilla Martens, Frederik Kragerud Goplen, Karl Fredrik Nordfalk, Torbjørn Aasen, and Stein Helge Glad Nordahl. Prevalence and characteristics of positional nystagmus in normal subjects. *Otolaryngology–Head and Neck Surgery*, 154(5):861–867, 2016. 1

[28] Stefano Massaroli, Michael Poli, Jinkyoo Park, Atsushi Yamashita, and Hajime Asama. Dissecting neural odes. *Advances in Neural Information Processing Systems*, 33:3952–3963, 2020. 4

[29] Tahira Nazir, Aun Irtaza, Ali Javed, Hafiz Malik, Dildar Hussain, and Rizwan Ali Naqvi. Retinal image analysis for diabetes-based eye disease detection using deep learning. *Applied Sciences*, 10(18):6185, 2020. 1, 2

[30] Ivan Dario Jimenez Rodriguez, Aaron Ames, and Yisong Yue. Lyanet: A lyapunov framework for training neural odes. In *International Conference on Machine Learning*, pages 18687–18703. PMLR, 2022. 4

[31] Alejandro Pasos Ruiz, Michael Flynn, James Large, Matthew Middlehurst, and Anthony Bagnall. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 35(2):401–449, 2021. 2

[32] Carl Runge. Über die numerische auflösung von differentialgleichungen. *Mathematische Annalen*, 46(2):167–178, 1895. 8

[33] Lars Ruthotto and Eldad Haber. Deep neural networks motivated by partial differential equations. *Journal of Mathematical Imaging and Vision*, 62:352–364, 2020. 4

[34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 8

[35] Pavel Senin and Sergey Malinchik. Sax-vsm: Interpretable time series classification using sax and vector space model. In *2013 IEEE 13th international conference on data mining*, pages 1175–1180. IEEE, 2013. 2

[36] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*, 2020. 8

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 8

[38] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 5, 6, 8

[39] Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*, 2019. 4

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[41] Luis Vianna, Leliane de Barros, and Scott Sanner. Real-time symbolic dynamic programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. 2

[42] Junlu Wang, Su Li, Wanting Ji, Tian Jiang, and Baoyan Song. A t-cnn time series classification method based on gram matrix. *Scientific Reports*, 12(1):15731, 2022. 3

[43] Zhiguang Wang and Tim Oates. Imaging time-series to improve classification and imputation. *arXiv preprint arXiv:1506.00327*, 2015. 3

[44] Kezhen Wei, Qiancheng Yang, Xiaoguo Yang, and Zhaobang Liu. Application of a pupil tracking method based on yolov5-deeplabv3+ fusion network on a new bppv nystagmus recorder. In *International Conference on Biomedical and Intelligent Systems (IC-BIS 2022)*, volume 12458, pages 948–955. SPIE, 2022. 1

[45] Ee Weinan. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 1(5):1–11, 2017. 4

[46] Niels West, Søren Hansen, Martin Nue Møller, Sune Land Bloch, and Mads Klokker. Repositioning chairs in benign paroxysmal positional vertigo: implications and clinical outcome. *European Archives of Oto-Rhino-Laryngology*, 273:573–580, 2016. 1

[47] Caiyu Wu, Sabor Nabil, Shihong Zhou, Min Wang, Liang Ying, and Guoxing Wang. Gram matrix-based convolutional neural network for biometric identification using photoplethysmography signal. *Journal of Shanghai Jiaotong University (Science)*, 27(4):463–472, 2022. 3

[48] Xuchao Zhang, Yifeng Gao, Jessica Lin, and Chang-Tien Lu. Tapnet: Multivariate time series classification with attentional prototypical network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6845–6852, 2020. 2

[49] Huiyun Zhao, Zhisong Pan, and Wei Tao. Regularized shapelet learning for scalable time series classification. *Computer Networks*, 173:107171, 2020. 2