BENIGN PAROXYSMAL POSITIONAL VERTIGO DISORDERS CLASSIFICATION USING EYE TRACKING DATA

Thang Anh Quan Nguyen[1], *Mohib Ullah*[2], *Azeddine Beghdadi*[1], *Faouzi Alaya Cheikh*[2]

[1]Galileo Institute, Sorbonne Paris Nord University, France
[2]Department of Computer Science, Norwegian University of Science and Technology, Norway.

ABSTRACT

This paper proposes a method that utilizes deep learning techniques in two tasks: (1) An object detection model was constructed to track the movement of the pupil central and dilation, as well as (2) a 1D-convolutional neural network to distinguish between different types of nystagmus and help diagnose underlying conditions. The results of the experiment demonstrate that when given the patient's eye video data, the system is capable of classifying the specific Benign Paroxysmal Positional Vertigo (BPPV) disorder out of the six possible types with accuracy of 88% on the test set. Additional, with the help of model soup, which allows averaging the weights of the models without requiring additional memory or inference time compared to ensemble methods, we can increase this accuracy to 91%. Such system has the potential to be a valuable tool for healthcare professionals and help doctors in the diagnosis and management of vertigo disorders.

***Index Terms*—** BPPV disorders, CNN, model soup, nystagmus, pupil detection, time-series data.

1. INTRODUCTION

Benign Paroxysmal Positional Vertigo (BPPV) is a common vestibular disorder, it is caused by the displacement of calcium carbonate crystals in the inner ear, which can lead to symptoms such as dizziness, and nausea. When patients move their head, these particles are excited within the affected canal, and the conflicting sensory information leads to a sensation of vertigo.

Nystagmus, which is characterized by a rapid, flickering, oscillatory movement of the eyes, is a crucial sign for the differential diagnosis of vertigo. The traces show a distinctive saw-tooth pattern as illustrated in Figure 1, which is composed of two parts: a slow phase and a fast phase. The slow phase has a shallow gradient, meaning that the eyes are moving slowly in one direction. This slow phase can provide important information about the underlying conditions that may be causing the nystagmus. The fast phase, on the other hand, has a steeper gradient, indicating that the eyes are quickly shifting in the opposite direction. The direction in which the eyes move involuntarily is determined by the gradient of the fast phase. Nystagmus can occur in three directions: horizontal, vertical, and torsional. Its detection is widely used in specialty clinics to evaluate patients with vertigo using observation by eye or video nystagmography (VNG). Due to development in camera technology and computer processing capability, VNG has become more popular and can serve as a more reliable method to measure eye movement. This method uses a camera to capture eye images, a computer to record the captured images, and software to detect and track eye movement to obtain a nystagmus waveform for diagnosis. Accurate diagnosis and classification of BPPV are crucial in determining the appropriate treatment approach, as the underlying causes and treatment methods can vary depending on the specific type of disorder.

BPPV can be classified into different types based on the location and orientation of the displaced crystals in the semicircular canals:

- Posterior canal BPPV: This is the most common type of BPPV, accounting for approximately 80% of the cases. It occurs when the displaced crystals are located in the posterior semicircular canal.
- Apo-geotropic lateral canal BPPV: This is a rare form of BPPV that occurs when the displaced crystals are located in the lateral semicircular canal, and the nystagmus is stronger when the affected ear is facing upwards.
- Geotropic lateral canal BPPV: This type of BPPV occurs when the displaced crystals are located in the lateral semicircular canal, and the nystagmus is stronger when the affected ear is facing downwards.

During a nystagmus examination, the patient is instructed to move their head and body in various ways while their eyes are observed. The most two commonly used tests are Dix-Hallpike test and lateral canal test. In the Dix-Hallpike test, the patient is rapidly brought from a sitting to a supine position with their head turned to one side, the vertical-torsional nystagmus is often observed either clockwise or counter-clockwise direction. While in the lateral canal test, the patient is positioned lying down on one side with their head turned to the opposite side so that functioning of the

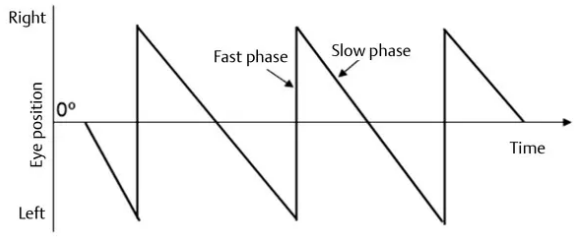

Fig. 1: Nystagmus beating refers to the involuntary movement of the eyes where they repeatedly move back and forth in a specific direction. The slow phase velocity (SPV) and fast phase velocity (FPV) are two important parameters that are used to measure the characteristics of beating.

horizontal canal could be evaluated. By holding each test duration for thirty seconds, the examiner then observes the patient's eyes for any nystagmus that may occur during these movements. The test is then repeated with the head turned to the other side to confirm the diagnosis.

In practice, recognizing nystagmus is difficult in current clinical practice due to a variety of challenges such as a lack of specialists and medical resources. Moreover, it is difficult to evaluate patients with droopy eyelids or eyelashes covering their pupils. As a result, Deep Neural Networks (DNNs) [1,2] have been applied to detect event in time-series data, VNG captures eye movements over time as videos, and these movements are crucial in diagnosing conditions such as BPPV. While 2D Convolutional Neural Networks (CNNs) can be effective in processing images, they are not well-suited for analyzing the temporal aspect of videos. 3D CNNs [3] or other temporal modeling techniques such as recurrent neural networks (RNNs) and optical flow [4] would be more appropriate for analyzing video data. The challenge of using 3D CNNs lies in the processing of spatiotemporal data. They are designed to handle 3D inputs, such as video frames with height, width, and time dimensions. However, they require significant computational resources and require a large amount of data to train effectively. Nevertheless, 1D CNNs and Long Short-Term Memory (LSTM) [5] layers have been shown to provide good results when tasked with detecting abnormal events from Encephalography (EEG) and Electrocardiography (ECG) [6]. Especially, 1D CNNs are particularly well suited to detection tasks in the time domain, specifically where target signals can occur at any time during the full signal [7, 8].

The rest of the paper is organized in the following order. In section 2, the description of the proposed model. The dataset, implementation details are given in section 3. Section 4 lists the quantitative results. The discussion and final remarks are given in section 5 that concludes the paper.

2. METHODOLOGY

There have been several studies exploring the use of deep learning for BPPV diagnosis. For example, Slama et al. [8, 9] considered several hand-crafted features such as gain, absolute preponderance, hypovalence, reflectivity, velocity and simple multilayer perceptron structures for VNG classification. Lim et al [10] represented a grid image of the amplitude of nystagmus at a specific position and then fed to a 2D CNN. Newman et al. [11] applied 2D CNN on composite recognition features, including eye-movement data and three-channel accelerometer data. The authors also addressed novel ensemble learning, cross validation and SMOTE to solve class imbalance problem. Trung X.P's et al. [12] system included relevant information of both eyes from RGB video and optical flow output using a bidirectional gated recurrent unit (Bi-GRU) framework to categorize the input video into two types of posterior canal BPPV and other. The beating feature of the eye was extracted by hand to detect when the beating happened and measure the speed of horizontal beating for the rest four lateral canal classes.

2.1. System overview

The overall framework of our system is shown in Figure 2. The two-stage approach consists of an object detection model to track the movement of the pupil, and a 1D CNN to distinguish between different types of nystagmus and help diagnose underlying conditions. In the first stage, we used a combination of convolution layers and feature extraction algorithms, to identify the pupil central location and dilation in each frame of the video which are important features for accurately measuring and analyzing eye movements. This information was then used to create a time-series data over the course of the video. In the next stage, the network was designed to take the data of the pupil movement extracted in the first step as input, and learn to identify patterns associated with different types of nystagmus.

2.2. Pupil detection

In the original clinic videos, there is no indication of the position of the pupil center. Therefore, the initial step involves finding the pupil's location within each video frame. To achieve this, a pupil location algorithm was utilized to identify and locate the center of the pupil in each video. Conventional approaches to detecting the eye pupil typically utilize image processing techniques such as edge detection, or intensity thresholds, which are specifically designed to identify the circular shape and localized intensity variation that are typically associated with the pupil.

Typically, edge detection detects the sharp contrast between the pupil and the iris, this involves the use of filters to identify edges such as the Canny edge detector or the Sobel operator to the gray scale image of the eye [14]. Since the best

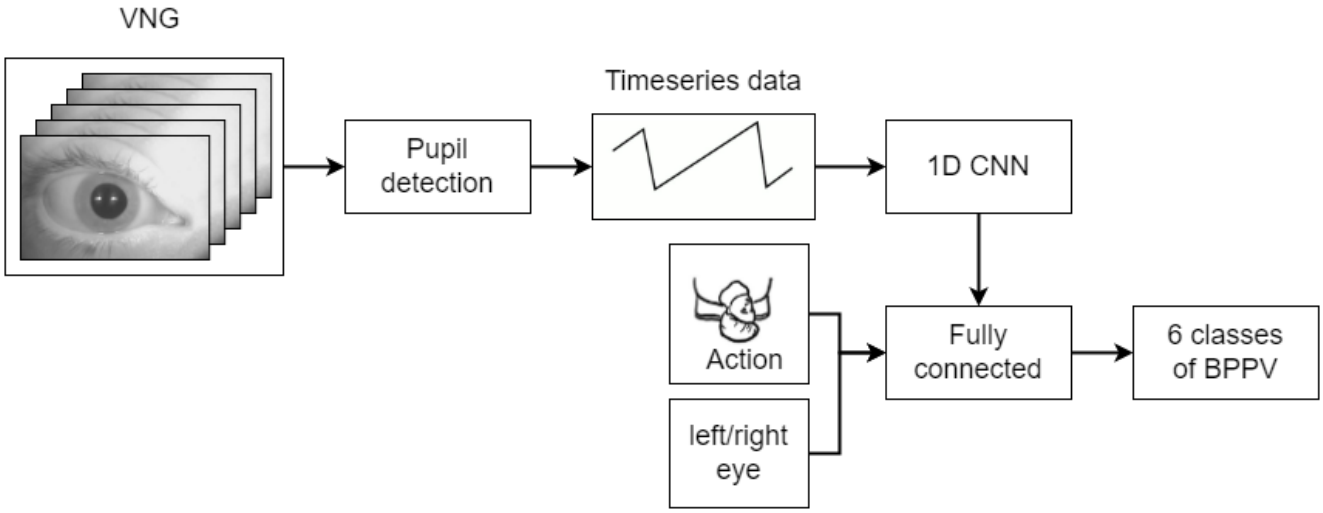

Fig. 2: The architecture for our baseline. Schematic drawing is taken from this paper [13].

shape approximation of the eye pupil image is an ellipse, an ellipse can then be fit to the feature points using several techniques: least-squares fitting of ellipse; voting-based methods, such as Hough transform; and searching-based methods, such as the random sample consensus (RANSAC). However, creating a reliable detector that can handle the many challenges of eye-tracking in everyday situations, including reflections, eyelid occlusion, is still a difficult problem. As the effectiveness of traditional algorithms reaches its limit, there is a shift towards using data-driven approaches. This is demonstrated by the growing trend towards machine-learning-based object detection methods.

The pupil detection model uses a single neural network, consists of CNN and fully connected layers. The model takes input as images of eye then computes the parameters of the pupil, including the center coordinates and radius. EfficientNet [15] models have shown to be highly efficient in terms of accuracy and computational resources. This makes them an excellent choice for fine-tuning on small datasets, where computational resources are limited. Additionally, these models have been pre-trained on large-scale datasets ImageNet, which allows for transfer learning to improve performance on smaller datasets. Therefore, we decided to use the pre-trained weights of Efficient-B0 (Figure 3) as the backbone. Overall, this architecture demonstrate superior performance in terms of both accuracy and efficiency compared to other CNNs, with a significant reduction in parameter size and FLOPS. The Huber loss function is used here due to its robustness to outliers in the data than the MSE loss function and being less sensitive to them than the MAE loss function:

$$\mathcal{L}_\delta(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{if } |y - \hat{y}| \leq \delta \\ \delta|y - \hat{y}| - \frac{1}{2}\delta^2, & \text{otherwise} \end{cases} \quad (1)$$

where the term y represents the the actual distribution and $\hat{y}$ represents the predicted distribution.

2.3. Nystagmus classification

We choose to apply 1D CNN for features extraction in this task, motivated by the successful application of similar network architectures to other event classification tasks. 1D CNNs are effective for recognizing patterns in time-based signals, especially when the features being analyzed can appear at any point in the signal. Moreover, they can automatically learn and extract meaningful features from raw input data, without the need for manual feature engineering. The input to the network consists of 5 channels of time series data, which are processed by three convolutional layers with 64, 128, and 64 filters respectively. Each convolutional layer is followed by a max pooling layer. A dropout layer to reduce overfitting and improving the generalization performance of neural networks. The output of the last pooling layer is flattened and concatenated with additional features (actions and left/right) before being fed into a fully connected layer. The total number of trainable parameters is 250822. The most commonly used loss function for classification problems is the cross-entropy loss:

$$\mathcal{L}_{CE}(y, \hat{y}) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (2)$$

There are five separate data sources available: The first two are eye-movement data, corresponding to horizontal and vertical eye movements. Pupil radius, representing changes in pupil size, either constricted or dilated is also taken into consider. Finally, the velocity of the signal is estimated by using simple difference to produce, which represent the first order

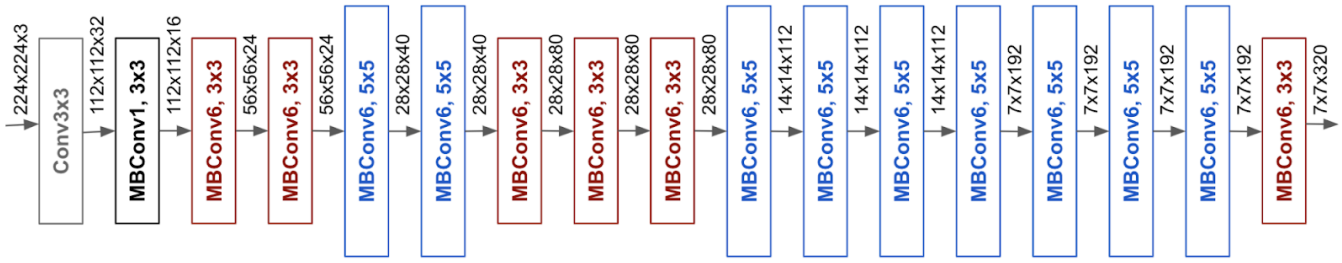

Fig. 3: The EfficientNet-B0 backbone for our detection network [15].

derivative of the signal. Using the velocity signal removes the need to adjust for any DC drift, which is common in time-series data recordings. Periodic spikes corresponding to periods of nystagmus are visible in the velocity signal, and their direction determines the sign of the nystagmus. It is generally understood that the first layers of the network are concerned with detecting lower level features of the target signal, such as signal velocity and acceleration, whereas later layers may learn more subtle, higher level features.

Considering the patient's position during the BPPV test is essential because the position of the patient's head and body can affect the direction of the nystagmus and can be used to determine the affected canal. Additional actions information provided each frame in the dataset such as sitting, head turns left, head turns right, head hanging left, head hanging right are encoded as one-hot vector and added to the network.

2.4. Model soup

The performance of a neural network can be highly dependent on its initial random seed and other hyperparameters such as learning rate, batch size, and regularization. This has led to a growing interest in understanding the impact of random initialization on neural network performance. Ensemble learning [16] is a machine learning technique that combines multiple individual models to improve overall performance. The performance of a neural network can vary depending on the initial random seed used to train it. The idea is to train several models, each with a different initial configuration or algorithm, and then combine their outputs through a weighting or voting strategy to make predictions.

By training or fine-tuning multiple models and then averaging their weights, the resulting model is able to take advantage of the strengths of each individual model and mitigate their weaknesses. This can lead to better overall performance compared to using a single model. The inference time of the resulting model is not increased since the weights are averaged, not concatenated, and we only need one soup model, which could also save storage space compared to ensemble learning method. The term "model soup" [17] comes from the idea of combining different models together like ingredients in a soup. There are several methods for making a soup of models. The first method is to average all the models, which creates a uniform soup. The second method, called the greedy soup, adds each model to the soup one by one. The greedy soup only keeps a model in the soup if it performs better than the other models on a held-out validation set. The models are sorted based on their accuracy on the validation set before the greedy soup is applied. This ensures that the greedy soup is at least as good as the best individual model on the validation set.

3. EXPERIMENTS

3.1. Dataset description

For this research, we combined several datasets. The additional datasets are the publicly available Labelled Pupil in the Wild (LPW) [18] and Eye Tracking XR. These are large-scale datasets designed for pupil detection tasks, containing high-resolution images captured in unconstrained environments, under different lighting conditions, with different camera settings and various head poses. This makes them suitable for evaluating the robustness and generalization of pupil detection algorithms in real-world scenarios. They also contain annotations of pupil positions, providing ground truth data for model training and evaluation. Additional annotations for the radius was conducted using [19] and [20].

In BPPV diagnosing task, the dataset was taken from this paper [12], all the videos were recorded in *.avi format with a resolution of 240 × 320 and at 30 frames per second. Each video consisted of three sub-videos, namely the left eye, right eye, and the entire body of the patient. The tests were conducted with the patient lying on a bed and a doctor performing Dix-Hallpike and lateral canal tests beside them. The videos were taken from unique patients, ranging from 10 to 70 years of age and comprising both genders. These two physical tests were consecutively recorded, resulting in a single video lasting approximately three minutes for each patient. Each class was specified by the typical movement of the eye, there were six classes in total with the corresponding eye movements:

- Left geotropic BPPV: Eye movements are towards the ground, with stronger beating to the left side (when head is turned left).

- Right geotropic BPPV: Eye movements are towards the ground, with stronger beating to the right side (when head is turned right).
- Left apo-geotropic BPPV: Eye movements are towards the sky, with stronger beating to the left side (when head is turned right).
- Right apo-geotropic BPPV: Eye movements are towards the sky, with stronger beating to the right side (when head is turned left).
- Left posterior canal BPPV: Eye movements are slightly up and rotate clockwise (often observed when the head is in a hanging left position).
- Right posterior canal BPPV: Eye movements are slightly up and rotate counterclockwise (often observed when the head is in a hanging right position).

Fig. 4: Example of the dataset. Each frame contains 3 subframes which record the left eye, right eye, and body pose. The black area does not contain information.

3.2. Implementation details

A sliding window with a duration of 100 consecutive samples with no overlap was used to convert the video samples into separate data point, each chunk represents 3.33 seconds of data. Each data point was then labeled with additional information including the BPPV class, which indicates the specific type of BPPV experienced by the patient (e.g. Lt_Apo_BPPV, Rt_Geo_BPPV, etc.), the current action of the patient during the test (e.g. turning head left or right), and whether it is the left or right eye that is considered. That means, we cropped left eye and right eye separately and discarded the body subvideos. This way of labeling is important for providing the model enough information to perform classification.

After detecting the pupil's location, we did a simple thresholding on the bounding box's intensity value to check if the subject's eye is currently closing or not, then we formed a 5-feature vector extracted from a video that contains information about the pupil coordinates, radius and velocities in both horizontal and vertical direction as input to the classifier. We assumed if the patients close their eye, these parameters won't change overtime and we can keep assign previous valid values to the current frame for more stability. Each row of this feature matrix is normalised using a Standard scaling. By doing so, all features will be removed the mean and scaled to unit variance.

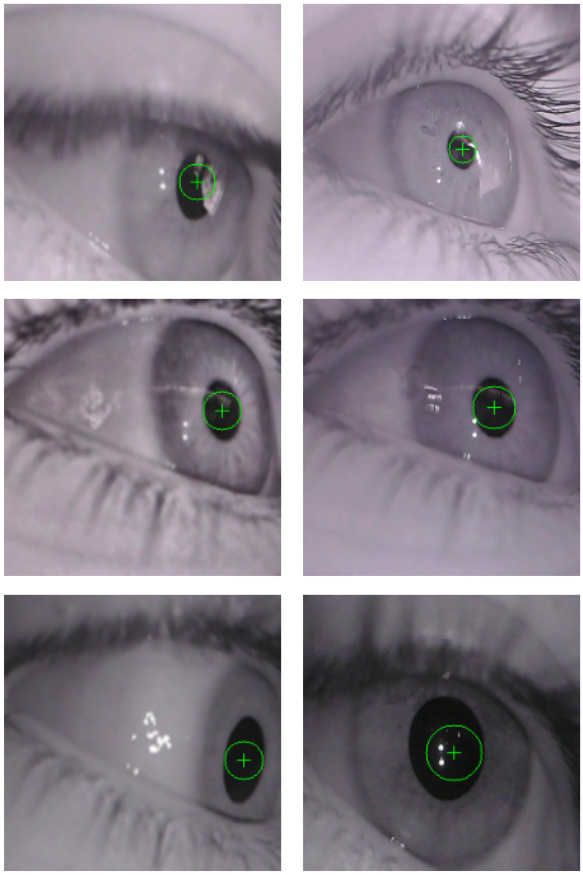

Fig. 5: Several results of the detection model on the LPW dataset.

The limited amount of eye movement data during nystagmus poses a challenge for training machine learning algorithms. Using such a small dataset can result in models that are overfitted and do not generalize well to unseen examples. Large class imbalances can also make it difficult for models to learn distinguishing features, leading to a model that classifies everything as the majority class. To overcome this, two techniques are commonly used: oversampling, which creates new examples of the underrepresented class, and undersampling, which reduces the number of examples in the majority class. Our approach involves the combination of traditional oversampling methods (random duplication of instances from

the minority class) along with novel data augmentation techniques that are based on our understanding of nystagmus including: Time-shifting, by randomly shifting the data points during training, the model can become more robust to slight variations in timing that may occur in real-world scenarios and noise injection which adds varying levels of noise to the signal. Flipping the data along the horizontal or vertical axis to create new examples with mirrored patterns is not considered in this case since we are dealing with classifying the beating direction.

We conducted all of our experiments using the open-source library Pytorch. We used the cloud-based computing platform Google Colab, which provides free access to Nvidia K80 GPU (12GB) for training and testing. First, all the frames were resized to (224x224x3) and normalized to $\mathcal{N}(0, 1)$ to match with EfficientNet-B0's input. The detection model was then trained first on LPW dataset with the fully connected layers for regression are set with randomly uniformed weights initially. Next, we used the above model for BPPV data labeling to train the classifier. The Adam optimizer was used for the learning process in both cases.

Each time the classifier was trained, it was seeded with a random weights value and/or hyperparameters results below:

- Learning rate from 0.001 to 0.005.
- Batch size in $\{4, 8, 16\}$.
- With and without label smoothing.
- With and without regularization.

4. RESULTS

The use of augmentation is to enhance the images from different lightning and varying camera viewpoints conditions. As a result, we applied a comprehensive range of transformation in our primary dataset for eye recordings such as histogram equalization, blurring, noise, cutout and Affine transform. After training the object detection model, the quantitative results, can be seen in Figure 5 are quite impressive. The model is able to accurately detect and identify the pupil in the images. Therefore, it can then be used for automatic labeling the BPPV dataset's nystagmus signal (Figure 6).

To evaluate performance, the following metrics were used: accuracy, precision, recall, and F1-score. Accuracy is calculated as the ratio of the number of correct predictions to the total number of predictions. Precision represents the proportion of predicted positive samples that are actually positive. Recall gives the proportion of actual positive samples that are correctly classified. F1-score is the harmonic mean of precision and recall, which provides a balance between them.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

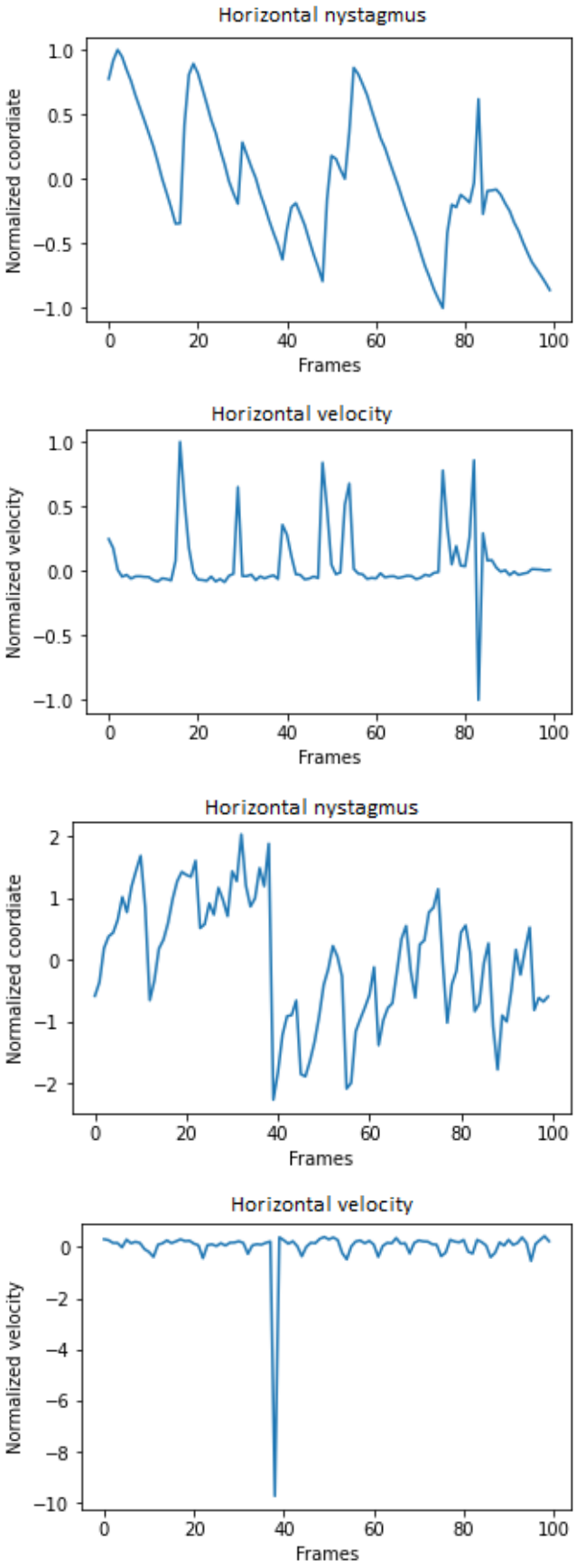

Fig. 6: Several acquired nystagmus diagrams and velocities calculated from pupil's coordinates in the horizontal axis.

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1score = 2\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{6}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. We also re-implemented temporal vision transformer model for this problem to compare with our method.

Table 1: Results on test set for serveral models

	accuracy	precision	recall	F1 score
Individual model	0.8487	0.8800	0.8487	0.8640
Best model	0.8859	0.8885	0.8859	0.8872
Uniform soup	0.4453	0.2123	0.4453	0.2878
Greedy soup	0.9027	0.9096	0.9027	0.9031
Video ViT	0.8601	0.8668	0.8601	0.8605

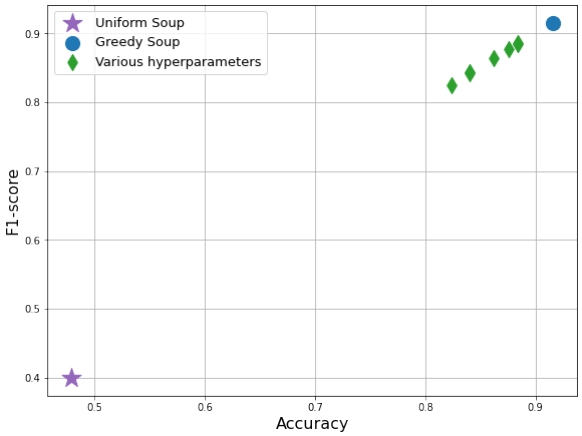

Fig. 7: Greedy soups improve accuracy over all individual models

From the quantitative results listed in Table 1 and Figure 7, we can see that the greedy soup model outperformed the best model using parameters searching, suggesting that this approach is effective for the task.

The confusion matrix on the test set suggests that the model is not suffering from the negative effects of class imbalance, such as being biased towards the majority class. This would be reflected in the matrix by higher values in the true positive and false negative cells for the underrepresented class compared to the true negative and false positive cells in every class.

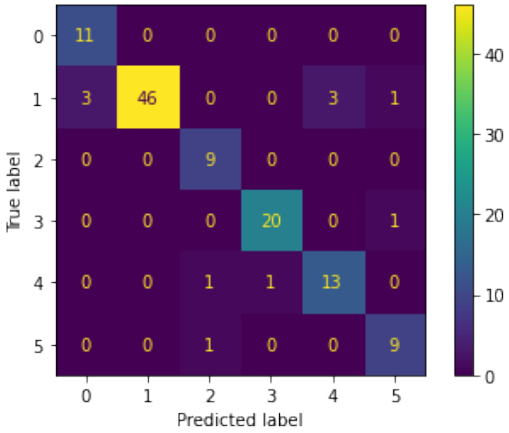

Fig. 8: Confusion matrix of greedy soup model on test set.

5. CONCLUSION

In this work, we have demonstrated techniques for overcoming the limited availability of data for training neural networks to detect nystagmus and proposed a deep learning-based system to support the doctor in automatically diagnosing BPPV disorder effectively based on the visual motions of the patient eyes in the standard maneuver tests. A two-stage architecture is constructed to extract different types of features for accurate classification.

6. REFERENCES

[1] Varad Kabade, Ritika Hooda, Chahat Raj, Zainab Awan, Allison Young, Miriam Welgampola, and Mukesh Prasad, "Machine learning techniques for differential diagnosis of vertigo and dizziness: A review," *Sensors*, vol. 21, pp. 7565, 11 2021.

[2] David Rastall and Kemar Green, "Deep learning in acute vertigo diagnosis," *Journal of the Neurological Sciences*, vol. 443, pp. 120454, 10 2022.

[3] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3d convolutional neural networks for human action recognition," 08 2010, vol. 35, pp. 495–502.

[4] Wanlu Zhang, Haiyan Wu, Yang Liu, Shuai Zheng, Zhizhe Liu, Youru Li, Yao Zhao, and Zhenfeng Zhu, "Deep learning based torsional nystagmus detection for dizziness and vertigo diagnosis," *Biomedical Signal Processing and Control*, vol. 68, pp. 102616, 07 2021.

[5] Lu Shi, ChangYuan Wang, Feng Tian, and HongBo Jia, "An integrated neural network model for pupil detection and tracking," *Soft Computing*, vol. 25, pp. 1–11, 08 2021.

[6] Özal Yıldırım, Paweł Pławiak, Ru-San Tan, and U. Rajendra Acharya, "Arrhythmia detection using deep convolutional neural network with long duration ecg signals," *Computers in Biology and Medicine*, vol. 102, pp. 411–420, 2018.

[7] Jacob L. Newman, John S. Phillips, and Stephen J. Cox, "1d convolutional neural networks for detecting nystagmus," *IEEE*

Journal of Biomedical and Health Informatics, vol. 25, no. 5, pp. 1814–1823, 2021.

[8] Amine Ben Slama, Aymen Mouelhi, Hanene Sahli, Zeraii Abderrazek, 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*, vol. 8, pp. 1–9, 12 2019.

[9] Amine Slama, Aymen Mouelhi, Hanene Sahli, Sondes Manoubi, Rim Lahiani, Mamia Salah, Hedi Trabelsi, and Mounir Sayadi, "A new neural network method for peripheral vestibular disorder recognition using vng parameter optimisation," *International Journal of Biomedical Engineering and Technology*, vol. 27, 08 2018.

[10] Eun-Cheon Lim, Y-S Park, Jeon, Sun Young Kim, Kunwoo Lee, Chang 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*, vol. 8, pp. 633, 05 2019.

[11] Jacob Newman, John Phillips, and Stephen Cox, "Detecting positional vertigo using an ensemble of 2d convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 68, pp. 102708, 07 2021.

[12] Trung Xuan Pham, Jin Woong Choi, Rusty John Lloyd Mina, Thanh Nguyen, Sultan Rizky Madjid, and Chang Dong Yoo, "Lad: A hybrid deep learning system for benign paroxysmal positional vertigo disorders diagnostic," 2022.

[13] Dominik Straumann and Thomas Brandt, "Bedside provocation and liberation maneuvers in patients with benign paroxysmal positional vertigo," *Clinical and Translational Neuroscience*, vol. 4, pp. 2514183X1988189, 02 2020.

[14] Yoanda Syahbana, Yokota Yasunari, Morita Hiroyuki, Mitsuhiro Aoki, Suzuki Kanade, and Matsubara Yoshitaka, "Nystagmus estimation for dizziness diagnosis by pupil detection and tracking using mexican-hat-type ellipse pattern matching," *Healthcare*, vol. 9, pp. 885, 07 2021.

[15] Mingxing Tan and Quoc V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019.

[16] Thomas G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, Berlin, Heidelberg, 2000, pp. 1–15, Springer Berlin Heidelberg.

[17] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," 2022.

[18] Marc Tonsen, Xucong Zhang, Yusuke Sugano, and Andreas Bulling, "Labeled pupils in the wild: A dataset for studying pupil detection in unconstrained environments," 11 2015.

[19] Thiago Santini, Wolfgang Fuhl, and Enkelejda Kasneci, "Pure: Robust pupil detection for real-time pervasive eye tracking," *CoRR*, vol. abs/1712.08900, 2017.

[20] Shaharam Eivazi, Thiago Santini, Alireza Keshavarzi, Thomas C. Kübler, and Andrea Mazzei, "Improving real-time cnn-based pupil detection through domain-specific data augmentation," *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 2019.