Non-Eye Tracking, Deep Learning-enabled Detection of Nystagmus in Dizzy Patients

Narayani Wagle[1] NWAGLE1@JHU.EDU
[1] *Johns Hopkins University*

John Morkos[1] JMORKOS1@JHMI.EDU
Jorge Otero-Millan, PhD[2] JOM@BERKELEY.EDU
[2] *University of California, Berkeley*

Jingyan Liu[1] JLIU230@JHU.EDU
Henry Reith[1] HREITH1@JHU.EDU
Joseph Greenstein, PhD[1] JGREENST@JHU.EDU
Kirby Gong[1]
Daniil Pakhomov[1] DPAKHOM1@JHU.EDU
Sanchit Hira[1] SANCHITHIRA76@GMAIL.COM
Indranuj Gangan[1] IGANGAN1@JHU.EDU
Oleg Komogortsev, PhD[3] OK@TXSTATE.EDU
[3] *Texas State University*

Raimond Winslow, PhD[1] RWINSLOW@JHU.EDU
David Newman-Toker, MD, PhD[1] TOKER@JHU.EDU
David S. Zee, MD[1] DZEE1@JHU.EDU
Kemar E. Green, DO[1] KGREEN66@JHMI.EDU

Editors: Under Review for MIDL 2022

Abstract

Patients with dizziness related to disruption of the ear-brain-eye sensory and neural circuitry often present with a particular pattern of ocular instability called nystagmus. These subtle eye movements can be difficult to detect and interpret at the bedside, and usually require robust eye tracking devices for accurate quantification. Here, we adopted an image processing and deep learning approach to detect nystagmus directly from videos from a small clinical dataset without applying traditional eye tracking techniques. Classification with our best performing model resulted in an AUROC of 0.864. This method may have potential future applications in smartphone and augmented/virtual reality (AR/VR) eye tracking for healthcare purposes.

Keywords: nystagmus, eye tracking, deep , telemedicine

1. Introduction

About 4 million dizzy patients visit the emergency room annually; 5% have a devastating stroke and the rest a benign inner ear disease. The latter can be treated remotely; avoiding hospitalization and expensive diagnostic testing (Cheung et al., 2010; Newman-Toker et al., 2007, 2014, 2008). The estimated national cost of managing dizzy patients in the emergency room is approximately ∼4 billion dollars annually; neuroimaging accounting for ∼12% of the total cost (Saber Tehrani et al., 2013). Studies have shown that current image modalities are insensitive at detecting early signs of strokes, about 35% are diagnosed as a benign inner ear problem which leads to poor clinical outcomes in 40% of the missed cases (Newman-Toker et al., 2014). The brainstem contains neural circuitry for various crucial involuntary function (e .g., cardiovascular and respiratory). Therefore, these strokes can be complicated by rapid death or the need for lifelong artificial cardiac and/or respiratory support.

Nystagmus and other subtle eye movement findings found in dizzy patients are more sensitive and specific than brain imaging in identifying acute strokes affecting the brainstem or the cerebellum (Benarroch, 2018; Kattah et al., 2009). Nystagmus is defined as repetitive, to-and-fro movements of the eyes (**Figure 1**) that is initiated by slow-phases and followed by contra-directional fast phases (Leigh and Zee, 2015). The pattern of nystagmus (direction, velocity, temporal evolution, etc.) can localize with remarkable precision the underlying neural substrate that is damaged (i.e., inner ear or brainstem/cerebellum) and provide rapid diagnostic clues about changes in neurophysiology that occurs in strokes affecting these circuits. Nystagmus detection and interpretation remain a challenge for non-expert front-line providers who are faced with making rapid decisions regarding the care of dizzy patients. Bedside eye movement evaluation by a medical provider is the mainstay in most clinical settings; however, several factors may limit accurate detection (Green et al., 2021; Shaikh et al., 2021; Rucker and Zee, 2021). Yet, only a handful of medical centers have adopted portable eye tracking devices for bedside evaluation of these clinically important eye movements.

Automatic analysis of eye movement recordings is a key step in improving the diagnosis of dizzy patients on the front-line. Our solution to the nystagmus detection problem was to develop a deep-learning model. One obvious barrier to this was the paucity of a nystagmus video datasets given the limited number of clinician and researchers using eye tracking devices to record eye movements in dizzy patients. Additionally, nystagmus recordings from dizzy patients are usually riddled with additional eyelid and non-nystagmus eye movements – making it particularly challenging to work with such datasets. In our method, we detected nystagmus eye movements without using traditional eye tracking techniques. Therefore, instead of generating eye position-time traces from the videos and detecting nystagmus, we constructed a new set of images (filtered images) that contained a motion representation of nystagmus.

2. Related Work

The first attempts at video recognition of action sequences involved 3D CNNs (Bregler, 1997; Goddard, 1992; Tran et al., 2015). Later attempts to make these models more generalizable succeeded using Inception 3D (Carreira and Zisserman, 2017) and ResNet-like versions of Convolutional 3D (Hara et al., 2018). However, these models required significant memory

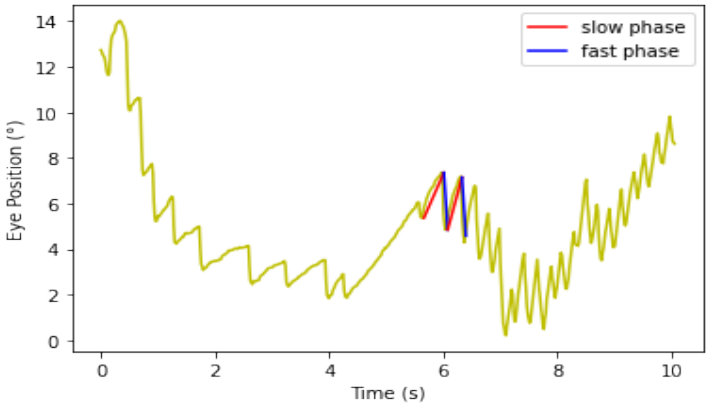

Figure 1: Example of a nystagmus position-time trace

and computational costs. Later developments in CNNs include Channel-Separated Convolutional Networks (CSNs) (Tran et al., 2019). Our methodology of generating filtered images was adapted from (Goddard, 1992; Masoud and Papanikolopoulos, 2003; Cao et al., 2004). In previous iterations of the algorithm, the goal of motion classification was twofold: to predict the type of motion based only on information in the data frame and to recognize types of motion that were not in the training dataset. For our purposes, we created a binary classification problem to differentiate between videos with and without nystagmus which made the second goal irrelevant in our case. Others have attempted to classify nystagmus in the past directly from the waveforms (Punuganti et al., 2019; Phillips et al., 2019; Newman et al., 2020, 2021) or by generating waveforms from recorded videos (Reinhardt et al., 2020; Lim et al., 2019) using various machine/deep learning methods.

3. Method

3.1. Dataset Description

Our dataset consisted of 500 monocular infrared video-oculography (VOG) recordings of dizzy patients from the AVERT (Acute Video-oculography for Vertigo in Emergency Rooms for Rapid Triage) clinical trial – a large multi-center trial comparing the accuracy of eye movements and brain imaging in diagnosing acute dizzy patients. This project was approved by the institutional IRB and informed consent was obtained from all participants. All videos were of the right eye; nystagmus in dizziness is almost always the same in both eyes. All videos were recorded using the Natus/Otometrics ICS Impulse infrared VOG goggles. Recordings were all sampled at 60 Hz and a had a resolution of 160 x 120 pixels. The duration of the videos varied but we only used the first 10 seconds (600 frames). Each video was labelled as "nystagmus" or "no nystagmus" by one expert Neuro-otologist (K.E.G.) based on the presence of two consecutive slow and fast phase alternations (beats) anywhere within the 10-second clip. The "nystagmus" to "no nystagmus" ratio in our dataset was approximately 1:1, and our dataset was split into a train to test ratio of 3:1.

3.2. Image Processing

Based on methods described in (Masoud and Papanikolopoulos, 2003; Cao et al., 2004), the video clip (600 frames) underwent recursive filtering to represent video motion based on the idea that a filtered image ($\mathbf{F}_t$) at time ($\mathbf{t}$) is defined as the absolute value of the difference between a raw video frame ($\mathbf{I}_t$) at time ($\mathbf{t}$) and an intermediate image ($\mathbf{M}_t$) at time ($\mathbf{t}$) that combines content of raw video frames prior to time point t.

$$F_t = |I_t - M_t| \tag{1}$$

$$M_t = (1 - \beta)M_{t-1} + \beta I_{t-1} \tag{2}$$

$$M_0 = I_1 \tag{3}$$

The appearance of the filtered image can then be modified by changing the parameter ($\boldsymbol{\beta}$) that control the weights of the prior context of the intermediate image ($\mathbf{M}$) and the raw video frame ($\mathbf{I}$). Filtered image examples are shown in **Figure 2**.

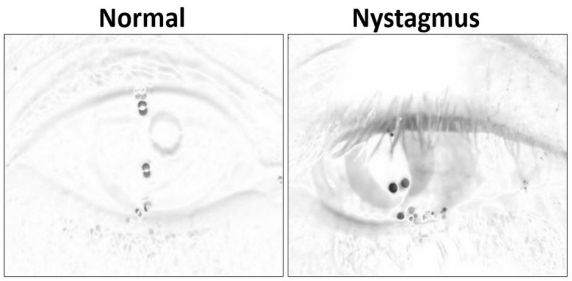

Figure 2: Example of a filtered image for nystagmus and normal cases

3.3. Neural Network Architecture

The proposed motion classification algorithm is a filtered image-based (Cao et al., 2004) approach (**Figure 3**). The filtered image-based motion classification algorithm uses the set of filtered images as video motion data. Each filtered image is labelled according to the label from the video it was generated and used to train a classifier in a supervised fashion. There are two primary methods of developing a filtered-image classifier. The first option involves training classifiers using deep learning approaches to develop a classifier. Classifiers trained on the ImageNet dataset, such as VGG, ResNet, etc which perform better on medical data, can be trained to detect nystagmus from filtered images (Cao et al., 2004). The second option for building a filtered image classifier involves using a support vector machine (SVM). This method performed well for classifying filtered images, specifically for motion classification (Cao et al., 2004). We chose the former method; our network architecture included the ResNet50 network starting with its ImageNet weights. The final layer of the ResNet50 architecture was excluded and five additional layers were

added to the network: 1 global pooling average layer, 3 dense layers with ReLU activation (size=1024, 1024, 512 respectively), and a final dense layer of size 2 with Softmax activation to generate the class probabilities. Our initial model was trained with a batch size of 32 for 5 epochs using the Adam optimizer. This model was trained and tested to yield class predictions of nystagmus or normal for each filtered image in the test set.

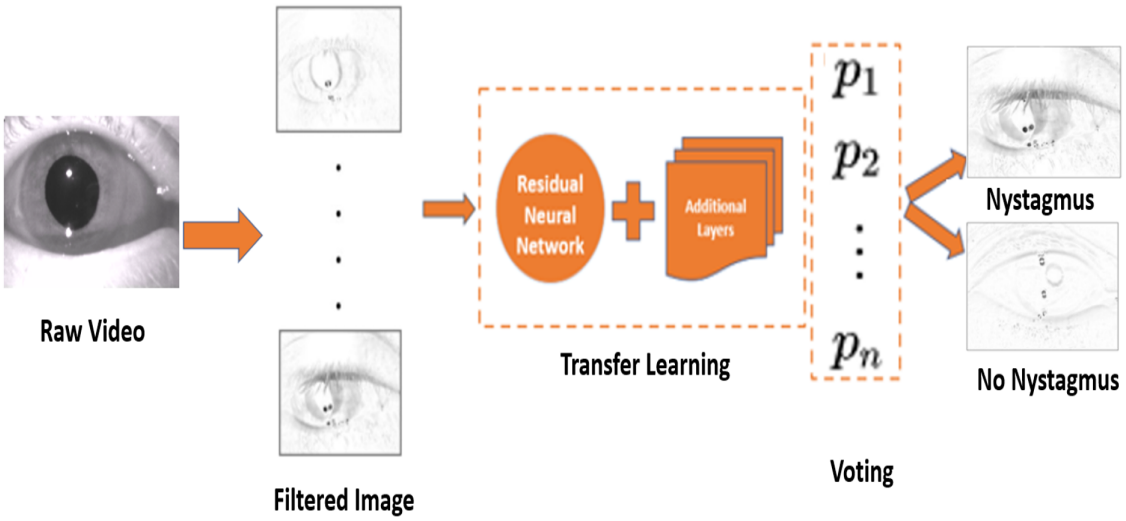

Figure 3: Filtered Image Model Framework

3.4. Voting

Filtered image class predictions were scored using voting methods to obtain a final prediction of "nystagmus" or "no nystagmus" for each video. Two methods of voting were used to score the predictions: soft voting and majority voting. The performance of the models was calculated using the operating point and AUROC with sensitivity and specificity.

Soft Voting The probability of nystagmus for a video ($\mathbf{p}_X$) that yields n filtered images can be defined as a sum of the probabilities of nystagmus for each of the filtered images ($\mathbf{p}_i$)

$$p_X = \frac{\sum_{i=1}^{n} p_i}{n} \tag{4}$$

Majority Voting The probability of nystagmus for a video ($\mathbf{p}_X$) that yields n filtered images can be defined as a sum of the scoring ($\mathbf{f}(x)$) of the probability of nystagmus for each filtered image ($\mathbf{p}_i$). Images are scored a 1 or 0 if they meet a threshold set by the operating point of the receiver characteristic curve generated from image probabilities.

$$p_X = \frac{\sum_{i=0}^{n} f(p_i)}{n} \tag{5}$$

$$f(x) = \begin{cases} 0 & x < threshold \\ 1 & x \geq threshold \end{cases} \tag{6}$$

4. Results

Our original model was adapted from (Cao et al., 2004), and was created with a beta value of 0.5 using majority voting had an AUROC of 0.786 as shown in **Table 1**. Several experiments were aimed at increasing the AUROC as well as sensitivity and specificity. All results are summarized in **Table 1**.

4.1. Filtered-Image Optimization

The filtered image calculation described above includes a free parameter ($\boldsymbol{\beta}$) that controls the temporal dynamics. We tested the performance of the method with 7 different β values (0.001, 0.005, 0.01, 0.05, 0.1, 0.25 and 0.5—see **Figure 4** and **Table 1**). Filtered images obtained with β of 0.25 resulted in the highest accuracy (79.3%); however, specificity was low (0.693) – implying a high false positive rate.

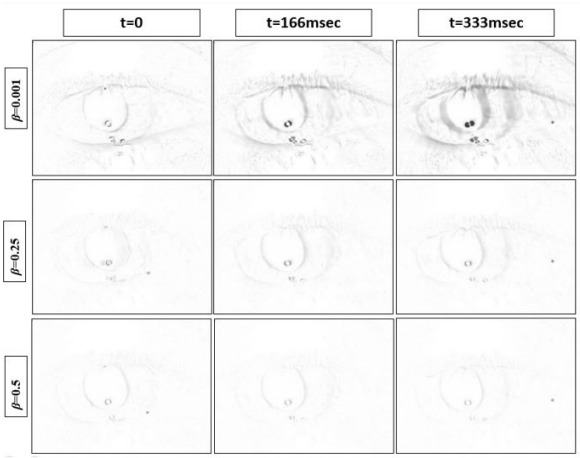

Figure 4: Showing nystagmus in three filtered images 166 milliseconds (msec) apart at β = 0.001, 0.25, 0.5

4.2. Comparison of Voting Strategies

Our neural network classifies individual filtered images as containing or not containing nystagmus. However, to evaluate the results against our labeled data, the entire video was classified as as containing or not containing nystagmus. Hard (majority voting) and soft voting techniques were compared using the value of β previously determined to be best (0.25). While the soft voting model had a slightly higher AUROC (0.847), there was better overall sensitivity (0.724) and specificity (0.838) with majority voting (**Table 1**).

4.3. Transfer Learning

For transfer learning, we used the ResNet50 architecture to build a neural network. We initialized the weights of the network to the weights used by ResNet50 for ImageNet dataset classification and then trained with those weights as our starting point. We also tested this approach with different network architectures that have been used for ImageNet dataset classification, including VGG16, DenseNet121, and InceptionV3. Our experiments indicated that VGG model had a slightly higher AUROC (0.848); however overall sensitivity and specificity were better with the ResNet and Inception models (**Table 1**).

4.4. Hyperparameter Tuning

In terms of hyperparameters, we performed experiments to select the optimal activation function, batch size, number of epochs, and optimizers. The ReLU activation function, five epochs, and a batch size of 32 were selected for all shown experiments. Experimentation with the optimizers showed the best overall performance with Adamax (AUROC = 0.835, sensitivity = 0.783 and specificity = 0.806). FTRL did not produce any meaningful result on our dataset (AUROC=0.500).

4.5. Ensemble

Bootstrap aggregating techniques were applied to the best performing models from the previous experiments producing the overall best performance as shown in **Figure 5** and **Table 1**. The ResNet-soft vote + VGG-hard vote ensemble model had the best performance metrics (AUROC = 0.857, Sensitivity = 0.884; Specificity=0.741)

4.6. Stratified k-fold Cross Validation

The ResNet-soft vote + VGG-hard vote ensemble model (best performing model) was cross-validated using stratified k-fold cross validation. As shown in Table 1, The highest observed AUROC was 0.912 with a mean accuracy of (80.4%) across all three folds. AUROC of each fold were .04 standard deviations away from the mean. As shown in **Figure 6**, the training loss indicates that the model is not overfitted.

4.7. Comparison with Existing Video Classification Method

Existing video classification methods (Karpathy et al., 2014) tend to use some form of frame sampling for model input. Therefore only a subset of the video frames are selected. Two consecutive beats of nystagmus can have a very short duration (as short as ∼500 milliseconds or ∼30 frames) and is likely to be found in tiny chunks of the videos in our dataset (given the relative frequency of noise imparted by eye closure, blinks and other technical issues affecting video recording quality). As a result, implementing existing methods risks eliminating the portions of our video that correspond to the class label. To counteract this, we used a simple LSTM (Hochreiter and Schmidhuber, 1997) and CNN model without any frame sampling. As suspected, the LSTM model performed poorly (AUROC=0.46) as shown in **Table 1**. We believe the results seen may be due to one or both of the following reasons. Since a large number of frames was provided as input per video (600 frames per video), the complexity of the LSTM + CNN network was limited to handle memory/space concerns. Additionally,

video classification methods perform predictions at a video level while our proposed method performs predictions at an image level. With video classification, our dataset for training was ∼500 videos whereas with our method, our dataset for training was ∼300,000 images.

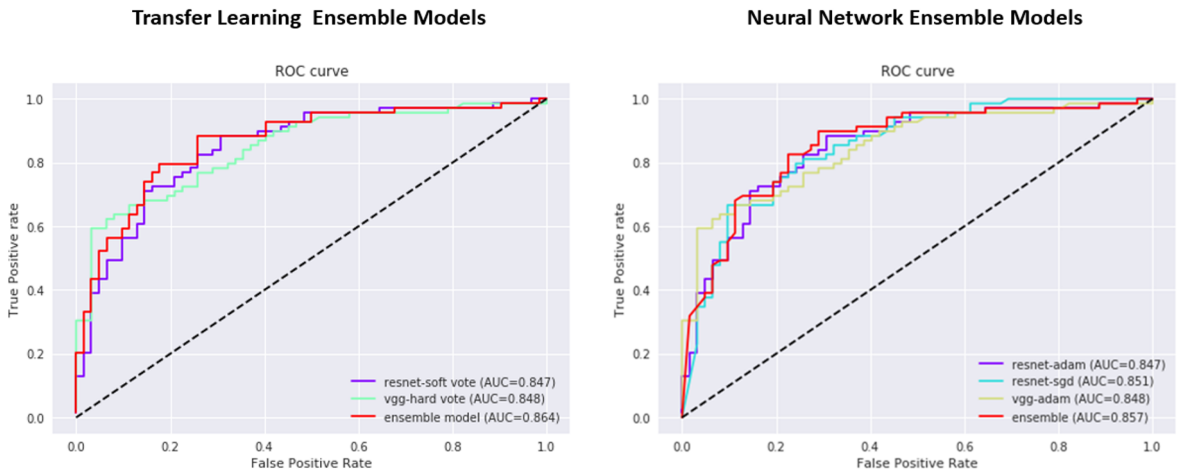

Figure 5: Performance of ensemble models

Figure 6: Training and Validation Loss

5. Conclusion

The results of our model show that it is possible to detect nystagmus from VOG recordings without first extracting eye position or velocity traces. Classifying the images or videos directly has the potential benefit of integrating information from all eye features at once such as pupil, eyelids, iris, etc. Future studies should further optimize the method to improve its performance and confirm if this approach can be more robust to videos including artifacts or different types of no-nystagmus eye movements.

Acknowledgments

We would like to thank the Johns Hopkins Neurology Department and the Neuro-Visual and Vestibular Disorders (NVV) division for supporting the project.

References

Eduardo E Benarroch. Brainstem integration of arousal, sleep, cardiovascular, and respiratory control. *Neurology*, 91(21):958–966, 2018.

Christoph Bregler. Learning and recognizing human dynamics in video sequences. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 568–574. IEEE, 1997.

Dongwei Cao, Osama T Masoud, Daniel Boley, and Nikolaos Papanikolopoulos. Online motion classification using support vector machines. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, volume 3, pages 2291–2296. IEEE, 2004.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

CSK Cheung, PSK Mak, KV Manley, JMY Lam, AYL Tsang, HMS Chan, TH Rainer, and Colin A Graham. Predictors of important neurological causes of dizziness among patients presenting to the emergency department. *Emergency Medicine Journal*, 27(7):517–521, 2010.

Nigel H Goddard. The perception of articulated motion: recognizing moving light displays. Technical report, ROCHESTER UNIV NY DEPT OF COMPUTER SCIENCE, 1992.

Kemar E Green, Jacob M Pogson, Jorge Otero-Millan, Daniel R Gold, Nana Tevzadze, Ali S Saber Tehrani, David S Zee, David E Newman-Toker, and Amir Kheradmand. Opinion and special articles: Remote evaluation of acute vertigo: Strategies and technological considerations. *Neurology*, 96(1):34–38, 2021.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

Jorge C Kattah, Arun V Talkad, David Z Wang, Yu-Hsiang Hsieh, and David E Newman-Toker. Hints to diagnose stroke in the acute vestibular syndrome: three-step bedside oculomotor examination more sensitive than early mri diffusion-weighted imaging. *Stroke*, 40(11):3504–3510, 2009.

R John Leigh and David S Zee. *The neurology of eye movements*. Contemporary Neurology, 2015.

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.

Osama Masoud and Nikos Papanikolopoulos. A method for human action recognition. *Image and Vision Computing*, 21(8):729–743, 2003.

Jacob L Newman, John S Phillips, and Stephen J Cox. 1d convolutional neural networks for detecting nystagmus. *IEEE Journal of Biomedical and Health Informatics*, 25(5): 1814–1823, 2020.

Jacob L Newman, John S Phillips, and Stephen J Cox. Detecting positional vertigo using an ensemble of 2d convolutional neural networks. *Biomedical Signal Processing and Control*, 68:102708, 2021.

David E Newman-Toker, Lisa M Cannon, Matthew E Stofferahn, Richard E Rothman, Yu-Hsiang Hsieh, and David S Zee. Imprecision in patient reports of dizziness symptom quality: a cross-sectional study conducted in an acute care setting. In *Mayo Clinic Proceedings*, volume 82, pages 1329–1340. Elsevier, 2007.

David E Newman-Toker, Yu-Hsiang Hsieh, Carlos A Camargo Jr, Andrea J Pelletier, Gregary T Butchy, and Jonathan A Edlow. Spectrum of dizziness visits to us emergency departments: cross-sectional analysis from a nationally representative sample. In *Mayo Clinic Proceedings*, volume 83, pages 765–775. Elsevier, 2008.

David E Newman-Toker, Ernest Moy, Ernest Valente, Rosanna Coffey, and Anika L Hines. Missed diagnosis of stroke in the emergency department: a cross-sectional analysis of a large population-based sample. *Diagnosis*, 1(2):155–166, 2014.

John S Phillips, Jacob L Newman, and Stephen J Cox. An investigation into the diagnostic accuracy, reliability, acceptability and safety of a novel device for continuous ambulatory vestibular assessment (cava). *Scientific reports*, 9(1):1–11, 2019.

Sai Akanksha Punuganti, Jing Tian, and Jorge Otero-Millan. Automatic quick-phase detection in bedside recordings from patients with acute dizziness and nystagmus. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 1–3, 2019.

Sophia Reinhardt, Joshua Schmidt, Michael Leuschel, Christiane Schüle, and Jörg Schipper. Vertigo–a pilot project in nystagmus detection via webcam. *Current Directions in Biomedical Engineering*, 6(1), 2020.

Janet C Rucker and David S Zee. Cerebellum—editorial regarding consensus paper consensus on virtual management of vestibular disorders: Urgent versus expedited care. shaikh et al., doi. org/10.1007/s12311-020—01178-8, 2021.

Ali S Saber Tehrani, Diarmuid Coughlan, Yu Hsiang Hsieh, Georgios Mantokoudis, Fredrick K Korley, Kevin A Kerber, Kevin D Frick, and David E Newman-Toker. Rising annual costs of dizziness presentations to us emergency departments. *Academic Emergency Medicine*, 20(7):689–696, 2013.

Aasef G Shaikh, Adolfo Bronstein, Sergio Carmona, Yoon-Hee Cha, Catherine Cho, Fatema F Ghasia, Daniel Gold, Kemar E Green, Christoph Helmchen, Richard T Ibitoye, et al. Consensus on virtual management of vestibular disorders: urgent versus expedited care. *The Cerebellum*, 20(1):4–8, 2021.

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.

Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019.

Appendix A. Performance Summary of Models

Table 1: Performance Summary of model experiments. All experiments were developed using Python's Tensorflow and Keras libraries. AUROC - Area under the curve

Image Modification				
	AUC	Sensitivity	Specificity	Accuracy
β=0.001	0.750	0.681	0.774	72.5
β=0.005	0.789	0.811	0.629	72.5
β=0.01	0.828	0.840	0.693	77.1
β=0.05	0.777	0.608	0.806	75.5
β=0.1	0.809	0.710	0.806	75.5
β=0.25	0.846	0.884	0.693	79.3
β=0.5	0.786	0.753	0.741	74.8
Voting(β=0.25)				
Majority Voting	0.839	0.724	0.838	77.8
Soft Voting	0.847	0.884	0.693	79.3
Transfer Learning				
ResNet50	0.839	0.724	0.838	77.8
DenseNet121	0.814	0.753	0.790	77.1
VGG16	0.848	0.594	0.967	77.1
InceptionV3	0.821	0.840	0.725	78.6
Neural Network Optimizer				
Adam	0.840	0.725	0.839	77.9
AdaGrad	0.830	0.623	0.919	76.3
RMSProp	0.794	0.812	0.677	74.8
ADADELTA	0.827	0.667	0.903	77.9
FTRL	0.500	0.000	1.000	47.3
SGD	0.851	0.667	0.903	77.9
Adamax	0.835	0.783	0.806	79.4
Nadam	0.780	0.754	0.758	75.6
Ensemble				
ResNet-Soft Voting	0.846	0.884	0.693	79.3
VGG-Majority Voting	0.848	0.594	0.967	77.1
ResNet-Soft Voting + VGG-Majority Voting*	0.864	0.884	0.741	81.6
ResNet-Adam	0.847	0.884	0.693	79.4
ResNet-SGD	0.851	0.667	0.903	77.9
VGG-Adam	0.848	0.594	0.968	77.1
ResNet-SGD + ResNet-Adam + VGG-Adam	0.857	0.898	0.710	80.9
Stratified k-fold Cross Validation				
Fold 1	0.912	0.721	0.970	84.7
Fold 2	0.817	0.836	0.730	78.3
Fold 3	0.849	0.853	0.729	78.4
Average across folds	0.859	0.803	0.809	80.4
Comparison with Existing Video Classification Model				
LSTM + CNN	0.460	1.00	0.020	48.4