

1D Convolutional Neural Networks for Detecting Nystagmus

Jacob L. Newman, John S. Phillips, and Stephen J. Cox

Abstract—Vertigo is a type of dizziness characterised by the subjective feeling of movement despite being stationary. One in four individuals in the community experience symptoms of dizziness at any given time, and it can be challenging for clinicians to diagnose the underlying cause. When dizziness is the result of a malfunction in the inner-ear, the eyes flicker and this is called nystagmus. In this article we describe the first use of Deep Neural Network Architectures applied to detecting nystagmus. The data used in these experiments was gathered during a clinical investigation of a novel medical device for recording head and eye movements. We describe methods for training networks using very limited amounts of training data, with an average of 11 mins of nystagmus across four subjects, and less than 24 hours of data in total, per subject. Our methods work by replicating and modifying existing samples to generate new data. In a cross-fold validation experiment, we achieve an average F1 score of 0.59 (SD = 0.24) across all four folds, showing that the methods employed are capable of identifying periods of nystagmus with a modest degree of accuracy. Notably, we were also able to identify periods of pathological nystagmus produced by a patient during an acute attack of Ménière’s Disease, despite training the network on nystagmus that was induced by different means.

Index Terms—1D convolutional neural networks, Biomedical signal processing, Dizziness, Electronystagmography, Nystagmus, Time series classification, Vertigo, Vestibular diseases.

I. INTRODUCTION

VERTIGO is a specific type of dizziness in which an individual perceives that they or their environment are moving, even though they are not [1]. Patients with vertigo can experience unpredictable attacks of severe spinning, and this can last for several hours at a time [2], during which they may be completely incapacitated. Dizziness and vertigo can impact significantly on many areas of a patient’s life, so quick access to a diagnosis and treatment is desirable. There are a range of clinical tests available for assessing balance disorders, such as dizziness and vertigo [3], but they are all performed in clinical environments and it is rare for them to take place whilst a dizzy

This work is funded by the UK Medical Research Council (MR/P026265/1).

J. L. Newman and S. J. Cox are with the School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, United Kingdom (e-mail: jacob.newman@uea.ac.uk; sjc@uea.ac.uk).

J. S. Phillips is with the Department of Ear, Nose, and Throat Surgery, Norfolk & Norwich University Hospitals NHS Foundation Trust, Norwich, NR4 7UY, United Kingdom (e-mail: john.phillips@mac.com).



Fig. 1. The CAVA device consists of five electrode pads contained within two, detachable mounts, and an electronic logging unit which sits behind the left ear. Two electrodes placed near the temples on either side of the face capture horizontal eye movement. Two electrodes above and below the left eye record vertical eye movement. A fifth electrode beneath the right ear provides a reference voltage. The device also contains a push button for patients to log events of interest, such as the onset of an attack of dizziness.

or vertigo attack is in progress. Dizziness is usually episodic and is often unpredictable [4], and some forms of dizziness can be induced by movement of the head. There are many possible causes of dizziness and vertigo [3], this means that forming a diagnosis is made even more challenging [5]. As such, patients often consult a number of clinicians from different specialities before receiving a definitive diagnosis or treatment [6], [7].

The Continuous Ambulatory Vestibular Assessment (CAVA) system has been developed to overcome the limitations of conventional balance assessments which only take a snapshot of a patient’s symptoms and in a clinical setting where it is rather unlikely that a dizziness or vertigo attack will take place. CAVA provides a continuous record of a patient’s vestibular function and is intended to be worn for thirty days, for twenty-three hours a day [8]. Hence it is highly likely to record any attacks of dizziness or vertigo that the patient experiences during this period. The data provided by the CAVA device is intended to be analysed by computer algorithms before presenting the outcome to a clinician to confirm and assess the results in the context of the patient’s other signs and test results, as it would be infeasible for clinicians to inspect many days of data manually. The development of these algorithms is the focus of the work presented here.

Vertigo is accompanied by a flickering eye-movement called nystagmus and therefore, observation of eye movement is crucial to clinicians for confirming whether a patient is experiencing true symptoms of vertigo. The CAVA device (Fig. 1) records horizontal and vertical eye-movements by way of the corneo-retinal potential produced by the eyeballs. Nystagmus is visible in eye-movement traces as a saw-tooth like signal,

made up of a slow phase (a waveform with a shallow gradient) and a fast phase (a waveform with a steeper gradient). The polarity of the gradient of the fast phase defines the direction of the nystagmus: a positive gradient corresponds to “right-beating” nystagmus, a negative gradient “left-beating”. The slow phase is clinically relevant because it corresponds to involuntarily drifting of the eyes because of a vestibular malfunction.

Previously, we undertook a clinical investigation involving healthy volunteers who wore the CAVA device continuously for up to thirty days [8], [9]. On eight days of their trial, each participant watched a nystagmus-inducing video on a VR headset. The data captured during this investigation was randomised prior to an automated computer analysis, the purpose of which was to identify the days on which nystagmus had been induced. The algorithms we developed for that study achieved a high level of diagnostic accuracy (sensitivity of 99.1% and specificity of 98.6%), demonstrating that very short periods of clinically useful information could be confidently identified from within days of normal eye-movement data.

Following this work, we continued to evaluate our device and algorithms on pathological nystagmus that was provided by patients experiencing vertigo as a symptom of specific inner-ear diseases, or was induced as a result of a routine balance test known as caloric testing. This data has provided some novel challenges in classification because of a number of differences between it and our artificially induced nystagmus data. The induced data was characterised by high-amplitude, highly regular sawtooth-like waves, that were always thirty seconds in duration. By contrast, pathological nystagmus has a much lower and much more variable amplitude, the signal-to-noise ratio is therefore lower, the fundamental frequency of the signal changes with time, and the total duration of the episodes is also highly variable. Furthermore, in our previous work, we were able to train models to detect nystagmus using a relatively small dataset of artificially induced data, which contained only a few minutes of nystagmus data. In order to train robust models capable of detecting a broad range of pathological nystagmus, much more data is required. Capturing adequate amounts of representative data is costly, time-consuming and generally challenging to obtain, as even symptomatic patients may only capture a few minutes of dizziness over the course of a month.

Our specific method of data capture also makes the task of identifying nystagmus more challenging. CAVA collects data in real-world environments, where patients are expected to apply the device to themselves, without expert supervision. Thus, user-error could negatively impact upon the quality of data collection, as could motion artefacts, or interference from household sources of electromagnetic radiation. The long-term duration of data capture also increases the chance of capturing unseen or rare examples of eye movement data, making classifiers more susceptible to making false positive detections. The large quantity of data could also make the classification process computationally slow. Thus, the variability of physiological nystagmus, the availability of representative training data, and the issues surrounding real-world data capture are the three main challenges posed by this task. The objectives of

the work presented here are to overcome these limitations by developing algorithms that can outperform traditional machine learning techniques, as a step towards an automated nystagmus detection system. To this end, we will soon undertake a blinded recognition task in which our algorithms will be presented with hundreds of data files, each representing a day’s worth of eye movement data. The algorithms will then automatically determine which of those files contains a period of nystagmus. Our ultimate aim is for the system to be able to provide automatic diagnosis as well as detection of nystagmus.

Apart from our previous work in [9], there are no previous studies that specifically focus on detecting nystagmus within long-term electrooculography data. However, several algorithms have been developed to identify nystagmus within short-term data [10]–[15]. Many of these systems adopt a heuristic approach to nystagmus detection, usually involving the identification of peaks in signal velocity, which can indicate the presence of a fast phase. For example, [15] used a peak detector followed by a clustering approach in order to identify fast phases within short duration, bedside recordings made from subjects with positional vertigo. Such approaches, while effective when applied to short-term data that are known to contain nystagmus, can be slow to process large quantities of data and may produce many false positive detections when applied to highly variable long-term data. 1D Convolutional Neural Networks (CNNs) have also been used to classify diseased versus healthy induced nystagmus signals captured using video goggles in clinical settings [16]. Despite this technique not being used to identify or confirm the presence of nystagmus, it is reassuring that it achieved a classification accuracy of 96.36% for discriminating signals from healthy people with those from patients suffering from Vestibular Neuritis and Ménière’s disease. Deep Neural Networks (DNNs) have also been applied to event detection in Encephalography (EEG) and Electrocardiography (ECG). Networks incorporating convolutional layers [17]–[19] and Long Short-Term Memory (LSTM) [20]–[22] layers have been shown to provide good results when tasked with detecting abnormal events from long-term EEG and ECG data.

In this article, we develop our algorithm’s capability to detect pathological nystagmus and present details of approaches taken to overcome the limited availability and imbalance of representative nystagmus data. We evaluate a Deep Neural Network (DNN) designed to detect periods of pathological nystagmus from within horizontal eye-movement data. Firstly, in section II, we describe more details of the CAVA device (II-A), followed by details of an ongoing clinical investigation (II-B), which is the source of the dataset described in section II-C. In section II-D, the experimental setup is explained, followed in section II-E by a detailed description of our approaches to overcoming limited training data and the DNN developed for this task. The results of our experiments are provided in section III, with a discussion in section IV. The manuscript concludes in section V.

II. METHODS

A. The CAVA Device

The CAVA device contains five ECG electrode pads that are strategically placed on the face to record the corneo-retinal potentials produced by the eyes (Fig. 1). The corneo-retinal potential is conventionally used as a proxy for eye-movement when use of cameras is deemed infeasible. Using this technique, also known as electrooculography or electronystagmography, the device records horizontal and vertical eye movement. The device also contains an accelerometer for recording 3-axis acceleration of the head. Vertical and horizontal eye movement data are sampled at approximately 42 Hz and 3-axis acceleration of the head at approximately 20 Hz. The device has been designed to require minimal intervention from the patient or the study team while deployed on trial, and so patients are not required to charge, download data or otherwise maintain their device. Patients are taught to apply and remove the device by themselves, to activate the device’s event marker and to interpret the device’s status LED. For more information about the CAVA device, please see [8].

B. Clinical Investigation

We are currently undertaking a clinical investigation involving patients suffering from pathological dizziness, such as individuals with Ménière’s disease, Vestibular Migraine and Benign Paroxysmal Positional Vertigo. We are in the first *training* phase of this investigation, in which patients are recruited to provide training and development data for our computer algorithms. This will be followed by a second phase in which patient data will be used as part of a blinded analysis. During the trial, patients are required to wear the CAVA device in the community, for twenty-three hours a day, for thirty days. Thus, patients wear the device during their normal daily activities and crucially during any dizzy attacks they experience. Typically, data captured in this way is 24 hours in duration and contains tens of minutes of nystagmus. The beat direction of the nystagmus can be left or right beating, depending on the patient’s specific condition and which ear(s) are affected.

At the end of each patient’s thirty day trial, they undergo caloric testing in a clinical setting. In practice, a patient may undergo many additional tests before receiving a firm clinical diagnosis, but only caloric testing is undertaken here, as it used as source of data collection rather than to facilitate a diagnosis. During this procedure, which lasts about twenty minutes, warm and then cool water are introduced into the inner ear canal, causing momentary dizziness, usually for a couple of minutes. In healthy people, warm water is expected to produce nystagmus beating towards the irrigated ear, whilst cool water produces nystagmus which beats in the opposite direction. For patients with vestibular malfunction, the nystagmus response may be weaker when the diseased ear is irrigated. Thus, the beat direction of nystagmus induced through caloric testing is controlled through the test itself. The experiments described in this article use a combination of data captured during caloric testing (3 out of 4 patients) and data captured during an attack of vertigo in the community (1 patient).

TABLE I

A SUMMARY OF THE DATASET USED IN THE NYSTAGMUS DETECTION EXPERIMENTS DESCRIBED IN THIS MANUSCRIPT. THE DURATIONS PRESENTED ARE THE TOTAL DURATIONS OF NYSTAGMUS AND NON-NYSTAGMUS DATA FOR EACH SUBJECT.

#	Nystagmus (mm:ss)	Non-Nystagmus (hh:mm:ss)	Source
1	10:53	00:51:55	Caloric Test
2	08:58	00:54:00	Caloric Test
3	06:14	16:50:24	Caloric Test
4	17:26	23:35:50	Ménière’s Attack

C. Dataset

The dataset used in the following experiments consists of data captured from four individuals (Table I). Here, we only use the data corresponding to horizontal eye-movement, as the nystagmus we are aiming to detect occurs almost entirely in the horizontal plane. The data was sampled with 12-bit precision and at a rate of approximately 42 Hz. The data from three of these individuals was captured during a caloric testing procedure, during which four separate periods of nystagmus are expected, each lasting up to three minutes. The difference in total data duration for each patient is mainly due to the duration that each patient wore their device. Patients 1 and 2 donned the CAVA device shortly before the caloric test started, whereas patient 3 was wearing their device for several hours before the test. The data from patient 4 represents a full day of data, during which the patient reported experiencing an acute Ménière’s attack, over a period of about three hours. All data was hand-labelled at the sample level with either a 0 (normal eye movement) or a 1 (nystagmus), based on a clinical expert’s interpretation on the presence of nystagmus in each signal.

D. Experimental Setup

The main classification task was to automatically classify each frame (where a frame is the data extracted from a moving window) as either a positive example of nystagmus, or not. The best frame duration was determined by experimentation and the results are presented in section III. To evaluate our system, we employed per-subject cross-fold-validation. Using this approach, the data is divided into a number of testing and training “folds”. Each testing fold contains data from a single subject and the data from the remaining subjects is used to train the neural network: this means that the system is always tested on data from a patient it has never “seen” before. In addition, we also withhold 20% of data from each training fold to provide development data which was used to determine the optimal network configuration for this task. Table II shows the quantity of data within each of the four folds, including the proportion of nystagmus data both before and after data augmentation and class balancing steps were applied (see section II-E.2 for more details).

E. Nystagmus Detection System

The nystagmus detection system is described in the following sections. The feature extraction process applied to the training and testing data is described in section II-E.1.

TABLE II

SHOWS THE NUMBER AND PROPORTION OF TRAINING FRAMES FOR EACH SUBJECT'S TESTING FOLD, BEFORE AND AFTER DATA AUGMENTATION AND SMOTE CLASS BALANCING. THIS DATA RELATES TO EXPERIMENTS CONDUCTED USING A FRAME SIZE OF 400 FRAMES, BUT THE PROPORTIONS ARE VALID FOR OTHER FRAME DURATIONS.

#	# (%) Nyst. Frames Before Augmentation	# (%) Nyst. Frames After Augmentation	# (%) Nyst. Frames After SMOTE	# Non-Nyst. Frames
1	155 (1.2%)	620 (4.8%)	12407 (50%)	12407
2	174 (1.4%)	696 (5.3%)	12387 (50%)	12387
3	191 (2.5%)	764 (9.1%)	7604 (50%)	7604
4	131 (2.3%)	524 (8.6%)	5579 (50%)	5579

The methods by which we address the imbalance in class data are described in section II-E.2. In section II-E.3, we provide details of the DNN architecture we use. The machine learning elements of the system were developed in Python, using the Keras software package [23]. Post-processing and data visualisation was performed using MATLAB. Lastly, in section II-E.4 we discuss the classification process, including a smoothing step applied to the DNN output.

1) *Feature Extraction*: A non-overlapping sliding window is used to segment the time-series data (Fig. 2). No filtering or pre-processing is applied to the data. We estimate the first order derivative (velocity) of the signal by simple differencing, producing vectors which we term *frames*. Using the velocity signal negates the need to remove any DC drift in the signal, which is common in electrooculography recordings. In the velocity signal, periods of nystagmus are visible as periodic spikes, whose sign depends on the direction of the nystagmus. Each frame of data is normalised to be a unit vector. The original data is labelled at the sample-level, and the class label (“nystagmus present” or “nystagmus not present”) of each frame is determined by majority vote of the samples from which it was derived. For example, for a frame duration of 400 samples, a frame containing 100 nystagmus samples and 300 non-nystagmus samples would be assigned a “nystagmus not present” label. In the case of a tie, frames are labelled as “nystagmus not present”.

2) *Balancing Class Data*: The small amount of nystagmus eye movement data available is a significant challenge when training machine learning algorithms for this task. Although some patients report episodes of dizziness lasting up to several hours, our data shows that when they do occur, periods of nystagmus are sporadic and last for a few minutes at most. Even if patients were to experience daily attacks, this would still correspond to less than 1% of the total eye-movement data collected. Training with such a small set of nystagmus data leads to overfitted models that do not generalise well to unseen examples of nystagmus [24]. Large class imbalances can prevent models from learning discriminative features, as the optimal model becomes close to one that simply classifies everything as the majority class.

There are two predominant techniques for overcoming class imbalances: oversampling and undersampling. Oversampling aims to create new examples of the underrepresented class, whilst undersampling reduces the number of examples in the majority class. Experimentally, oversampling has been

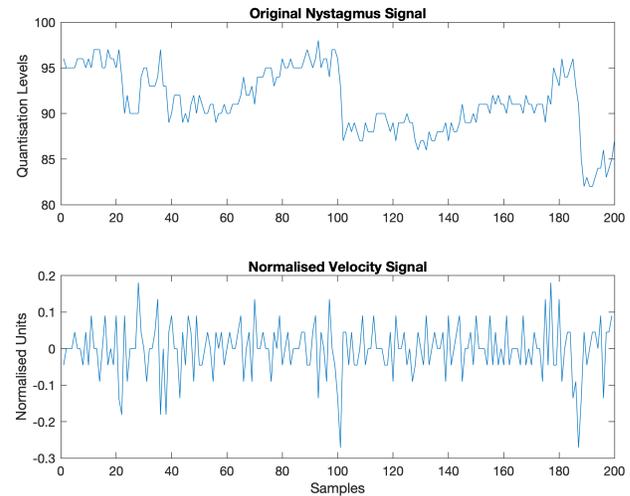


Fig. 2. The top panel shows a 200 sample frame corresponding to an example of left-beating nystagmus. The bottom panel shows the velocity signal extracted from the example shown in the top panel. The signal has been normalised by transformation to a unit vector, such that the magnitude of the vector is equal to 1. The fast phases of the nystagmus are visible as prominent peaks of negative velocities.

shown to outperform undersampling [25], [26], especially when applied to large class imbalances and when training neural networks. A number of oversampling techniques have previously been described for rebalancing class data, including random duplication of examples from the minority class [27], Synthetic Minority Oversampling Technique (SMOTE, [28]), which generates new examples by interpolating the feature space between neighbouring data points, or by exploiting an understanding of the data, such as by mirroring or translating signals [29].

To address these issues, we have employed a number of techniques designed to create new training examples of nystagmus from the limited number of examples available in each training fold (Fig. 3). Our approach combines conventional oversampling techniques with data replication methods based on our intuition about nystagmus. The techniques are applied separately for each fold of the cross-validation. First, each nystagmus frame is duplicated and reversed in time. This step results in nystagmus that beats in the opposite direction to the original example. Next, all examples are duplicated and multiplied by -1 , which again reverses the direction of the nystagmus but by reversing in the y-axis (e.g. a velocity of 1 becomes a velocity of -1). Three new examples of nystagmus are produced for each original frame of nystagmus. These data augmentation steps do not require knowledge of the beat direction for the original nystagmus signal, as we are not currently concerned with balancing the quantities of left and right-beating nystagmus. Finally, we use SMOTE to balance the number of examples in the nystagmus and non-nystagmus classes.

3) *Neural Network*: Fig. 4 shows the Deep Neural Network (DNN) architecture developed for use in these experiments. One network was trained for each fold of the cross-validation using an Nvidia GeForce GTX 1080 Ti GPU-enabled graphics

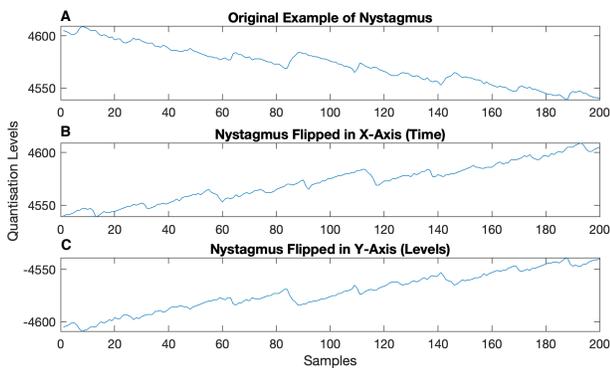


Fig. 3. The methods of data manipulation used to generate new examples of nystagmus from existing ones. **(A)** A 200-sample frame displaying an original timeseries waveform. The waveform is a positive example of nystagmus. **(B)** Nystagmus examples are duplicated and flipped in the x-axis, and **(C)** also duplicated and flipped vertically. These steps produce four times the original volume of nystagmus training data.

card, taking approximately thirty-seconds per epoch (an *epoch* is a single pass of the training data through the network, during training). Our DNNs use 1D Convolutional layers, hence they are Convolutional Neural Networks (CNN). In a 1D-CNN, it is generally accepted 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. We opted to use CNNs because they have shown to work well for event detection in other types of 1D signal, such as arrhythmia detection in Electrocardiography (ECG) data [17], [18]. 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. The arrangement of our CNN architecture was adapted from examples of networks successfully applied to ECG event detection. The parameters used in our networks, such as the kernel size and number of filters per layer, were determined by way of preliminary parameter searches. The values selected provided a good balance between classification accuracy and time taken to train the networks.

The network consists of 11 layers in total. The input layer has 199 dimensions, corresponding to the dimensionality of the velocity features in the data frames. This is followed by two 1D convolutional layers, with a kernel size of 3, which are intended to learn the basic features of the data. A 20% dropout layer is used to improve the generalisability of the network, followed by two more 1D convolutional layers. A 1D pooling layer reduces the network dimensionality to 128. A dropout layer precedes two Dense layers, followed by the final output layer. The total number of trainable parameters was 72, 953. To train the network, the Adam’s optimiser and a learning rate of 0.001 was used, with a batch size of 20. All networks were trained using 30 epochs, which was found to be the optimal duration for classification of the development data. Binary cross-entropy was selected as the loss function and accuracy was the chosen performance metric.

4) Classification: Unseen test data was treated using the same feature extraction process as applied to the training data

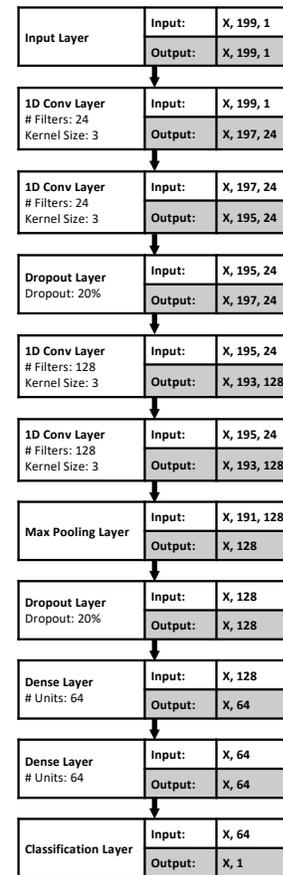


Fig. 4. Deep Neural Network architecture containing 11 layers, inclusive of input and output layers. ‘X’ denotes a sample / frame.

(Section II-E.1). Testing data was classified on a frame-by-frame basis by a fold-specific Deep Neural Network (DNN), as described in section II-E.3. After this classification stage, each frame was represented by a binary classification, indicating whether that frame contained nystagmus or not.

A sequence of classified frames typically has some frames labelled “nystagmus” and some “non-nystagmus”. A single frame classified as nystagmus, surrounded by non-nystagmus frames, is not likely to be a genuine episode, as episodes of nystagmus are typically much longer than the duration represented by a single frame (14 sec is the longest frame duration tested here). Similarly, a frame classified as non-nystagmus that is found within a number of positively classified frames is likely to be a false negative detection. Therefore, we used a *sieve* filter to smooth the output from the classification. For a full description of the operation of a sieve filter, please see [30], but to summarise, the sieve essentially operates by removing very short durations of negative or positive classifications.

In addition to the DNN classifier described here, we also performed baseline experiments using a Support Vector Machine (SVM) classifier and neural networks containing only non-convolutional layers. The SVM classifier and one of the non-convolutional networks used the same velocity features as the DNN classifier. We did not normalise the recognition features for the SVM classifier, as this classifier is not capable of

extracting temporal patterns, and normalisation could destroy some potentially discriminative aspects of the data. Parameter optimisation was used to select the best configuration of SVM classifier for each training fold. A further non-convolutional neural network baseline used frequency domain recognition features (Fast Fourier Transform) instead of velocity features, and was configured in a similar manner to [9]. All experiments were evaluated using the same cross-fold validation approach, and the same training data was used for comparable experiments. These baseline experiments were performed using all class balancing techniques (augmentation and SMOTE), but we did not apply the sieve filter, as the results are generally too poor to benefit from post-processing.

III. RESULTS

The first experiment sought to find the optimal frame duration for the subsequent experiments. Table III shows the results of varying the frame duration from 100 samples (2.3 s) to 600 samples (14.1 s). These results were obtained using both data augmentation and SMOTE simultaneously. The average F1 score was lowest for a frame duration of 100 samples, suggesting that this duration is not long enough to capture a sufficient number of nystagmus beats in order to train a reliable network. A frame duration of 400-samples produced the highest average performance across all metrics except for sensitivity, which was only marginally lower than the highest value obtained. Therefore, all subsequent experiments are performed using a 400-sample frame duration.

Table IV shows the results of the nystagmus detection task across eight different experiments: First, three baseline experiments using an SVM and two non-convolutional neural networks, followed by five different system configurations of Deep Neural Network (DNN). For the five DNN systems, the first uses networks trained without using any class balancing techniques. The second is a system where the class data is replicated by the augmentation methods described in section II-E.2, but not using the SMOTE method or any post-processing of the classification. The third a system uses SMOTE without the other data replication techniques. The networks in the fourth system are trained using all class balancing techniques, including data replication and SMOTE, but no sieve filter. In the final system, all data replication approaches were used, including the sieve filter. We mostly consider the F1 scores when comparing results from the different classifiers, as this metric is commonly used in Computer Science to summarise the results of binary classification tasks. More detail regarding the F1 score can be found in [31], but in summary it provides the harmonic mean of precision and recall.

The results for all baseline experiments showed poor performance compared to the DNN approaches. The results from the SVM classifier were the lowest of the three baselines, with poor results across all metrics, except for accuracy. However, the values shown for classification accuracy are misleadingly high for all experiments, which is a common issue when evaluating classification performance on a vastly imbalanced dataset, where high accuracies can be achieved simply by

classifying all examples as belonging to the majority class. The non-convolutional networks offered improved results over the SVM, with the network trained using velocity features outperforming the network trained using frequency domain recognition features. The average F1 score for each baseline experiment was worse than for all configurations of DNN. A McNemar’s test confirmed that the difference in performance was statistically significant for all configurations of DNN compared to all other systems ($p < 0.0001$). We achieved qualitatively similar results to the SVM using Random Forest, K-Nearest Neighbour and XGBoost classifiers.

TABLE III

EXPERIMENTAL RESULTS SHOWING THE EFFECT OF VARYING FRAME SIZE (IN SAMPLES) ON THE FRAME-LEVEL CLASSIFICATION PERFORMANCE OF OUR DEEP NEURAL NETWORK SYSTEM. THE RESULTS FOR EACH SUBJECT WERE OBTAINED USING HOLD-ONE-OUT CROSS-VALIDATION, IN WHICH EACH SUBJECT WAS HELD-OUT FOR TESTING AND THE REMAINING SUBJECTS WERE USED FOR TRAINING. EACH CLASSIFIER USED ALL CLASS BALANCING TECHNIQUES (AUGMENTATION AND SMOTE), BUT NO SIEVE POST-PROCESSING.

#	tp	tn	fp	fn	Sens. (%)	Spec. (%)	F1	MCC	Acc. (%)
100 samples									
1	118	1035	259	159	43	80	0.36	0.20	0.73
2	122	1235	111	106	54	92	0.53	0.45	0.86
3	111	20578	4677	40	74	81	0.04	0.11	0.81
4	219	32956	2401	257	46	93	0.14	0.17	0.93
Mean:					54.3	86.5	0.27	0.23	0.83
200 samples									
1	85	546	99	55	61	85	0.52	0.41	0.80
2	91	558	113	25	78	83	0.57	0.50	0.82
3	60	11321	1305	17	78	90	0.08	0.17	0.90
4	171	14913	2749	83	67	84	0.11	0.17	0.84
Mean:					71.0	85.5	0.32	0.31	0.84
300 samples									
1	50	385	43	45	53	90	0.53	0.43	0.83
2	64	348	98	14	82	78	0.53	0.46	0.79
3	43	7533	882	10	81	90	0.09	0.18	0.89
4	138	8994	2769	43	76	76	0.09	0.15	0.76
Mean:					73.0	83.5	0.31	0.31	0.82
400 samples									
1	46	283	37	26	64	88	0.59	0.50	0.84
2	53	262	71	7	88	79	0.58	0.52	0.80
3	32	5746	564	9	78	91	0.10	0.19	0.91
4	86	7805	1009	58	60	89	0.14	0.19	0.88
Mean:					72.5	86.8	0.35	0.35	0.86
500 samples									
1	27	230	25	32	46	90	0.49	0.38	0.82
2	40	213	52	9	82	80	0.57	0.49	0.81
3	27	4701	346	7	79	93	0.13	0.23	0.93
4	82	5897	1147	40	67	84	0.12	0.17	0.83
Mean:					68.5	86.8	0.33	0.32	0.85
600 samples									
1	34	181	30	16	68	86	0.60	0.49	0.82
2	28	174	48	12	70	78	0.48	0.38	0.77
3	23	3893	312	6	79	93	0.13	0.22	0.92
4	67	4994	872	39	63	85	0.13	0.18	0.85
Mean:					70.0	85.5	0.34	0.32	0.84

*tp = true positive, tn = true negative, fp = false positive, fn = false negative, MCC = Matthews Correlation Coefficient.

For the different combinations of DNN system, the results from each combination of class balancing and sieve filtering are better than the baseline DNN, in terms of classification sensitivity and average F1 score. The differences are all statistically significant, according to a McNemar’s test. A combination of all techniques, including the sieve filter, provides

the highest F1 scores across three out of four subjects. For the best set of results, the sensitivity ranges from 25% for patient 4 to 81% for patient 2, and specificity near to 100% for all patients. Examination of the columns labelled tp, tn, fp and fn in Table IV shows that the number of false positive detections is extremely low compared to the number of true negative detections, producing a high level of specificity. Some systems showed a decrease in F1 score for patient 3 compared to the baseline. This was due to an increase in the number of false positive detections. However, inclusion of the sieve filter was sufficient to reduce these short and isolated misclassifications.

TABLE IV

RESULTS OF A FRAME-LEVEL CLASSIFICATION TASK. THE FIRST THREE ROWS SHOW THE RESULTS OF THREE BASELINE EXPERIMENTS OBTAINED USING AN SVM CLASSIFIER AND TWO NON-CONVOLUTIONAL NEURAL NETWORKS. EACH BASELINE USED ALL CLASS BALANCING TECHNIQUES BUT NO SIEVE POST-PROCESSING. THE FOURTH RESULT WAS OBTAINED USING A DEEP NEURAL NETWORK (DNN), WITHOUT CLASS BALANCING OR A SIEVE FILTER. ALL SUBSEQUENT RESULTS RELATE TO EXPERIMENTS USING DNNs AND VARIOUS COMBINATIONS OF CLASS BALANCING.

#	tp	tn	fp	fn	Sens. (%)	Spec. (%)	F1	MCC	Acc. (%)
SVM baseline									
1	0	320	0	72	0	100	0.00	0.00	0.82
2	0	333	0	60	0	100	0.00	0.00	0.85
3	0	6310	0	41	0	100	0.00	0.00	0.99
4	0	8814	0	144	0	100	0.00	0.00	0.98
Non-convolutional network baseline									
1	32	278	42	40	44	87	0.44	0.31	0.79
2	41	277	56	19	68	83	0.52	0.43	0.81
3	11	5218	1092	30	27	83	0.02	0.02	0.82
4	75	5749	3065	69	52	65	0.05	0.05	0.65
Non-convolutional network baseline using FFT features									
1	19	265	55	53	26	83	0.26	0.09	0.72
2	4	234	99	56	7	70	0.05	-0.19	0.61
3	0	6298	12	41	0	100	0.00	-0.00	0.99
4	120	3719	5095	24	83	42	0.04	0.07	0.43
DNN. No class balancing. No sieve filter									
1	0	320	0	72	0	100	0.00	0.00	0.82
2	0	332	1	60	0	100	0.00	-0.02	0.84
3	10	6296	14	31	24	100	0.31	0.32	0.99
4	0	8814	0	144	0	100	0.00	0.00	0.98
Data replication. No SMOTE or sieve filter									
1	18	315	5	54	25	98	0.38	0.39	0.85
2	31	330	3	29	52	99	0.66	0.65	0.92
3	22	6242	68	19	54	99	0.34	0.36	0.99
4	66	8631	183	78	46	98	0.34	0.33	0.97
SMOTE. No data replication or sieve filter									
1	33	311	9	39	46	97	0.58	0.54	0.88
2	34	321	12	26	57	96	0.64	0.59	0.90
3	32	6055	255	9	78	96	0.20	0.29	0.96
4	92	8193	621	52	64	93	0.21	0.26	0.92
All balancing techniques. No sieve filter									
1	46	283	37	26	64	88	0.59	0.50	0.84
2	53	262	71	7	88	79	0.58	0.52	0.80
3	32	5746	564	9	78	91	0.10	0.19	0.91
4	86	7805	1009	58	60	89	0.14	0.19	0.88
All balancing techniques, plus sieve filter									
1	35	320	0	37	49	100	0.65	0.66	0.91
2	49	321	12	11	82	96	0.81	0.78	0.94
3	30	6287	23	11	73	100	0.64	0.64	0.99
4	44	8647	167	100	31	98	0.25	0.24	0.97

*tp = true positive, tn = true negative, fp = false positive, fn = false negative, MCC = Matthews Correlation Coefficient.

In Fig. 5 we present the Receiver Operator Curves (ROCs) for each fold of the cross-validation experiment using all balancing techniques. These curves were generated using the

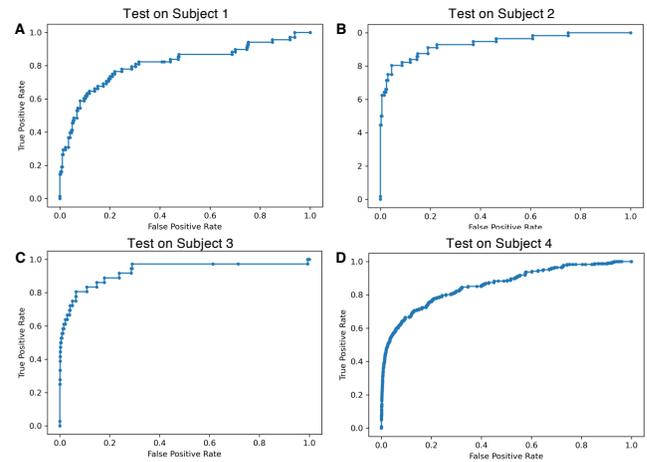


Fig. 5. These Receiver Operator Curves (ROCs) correspond to the classification outputs from each of the four cross-validation folds. They were generated using the direct outputs from the neural networks trained using all balancing techniques, but prior to the smoothing stage. **A** shows the ROC for subject 1 and has an Area Under Curve (AUC) of 0.81. **B** is for subject 2 and has an AUC of 0.93. **C** is for subject 3 and has an AUC of 0.93. **D** is for subject 4 and has an AUC of 0.87.

classification probabilities produced each fold-specific neural network. All plots show that the networks perform well across a range of classification thresholds. The Area Under Curve (AUC) statistic for each plot ranges from 0.85 to 0.93, demonstrating consistent discriminative capabilities across all testing folds.

IV. DISCUSSION

The results presented in the previous section have highlighted the problem of classifying events that are rather variable and occur as less than 1% of the available data. It is encouraging that we were able to use nystagmus data from patients undergoing caloric testing to train a network to detect pathological nystagmus. This is promising for future research as until there is widespread wearing of the CAVA device, caloric testing is the only reliable way to obtain vestibular-induced nystagmus data for analysis and diagnosis.

We have shown that 1D Convolutional Neural Networks (CNNs) are well-suited to this task and vastly outperform other machine learning approaches, such as Support Vector Machines (SVMs) and simpler non-convolutional neural network architectures. It is well known that 1D CNNs work well when applied to pattern recognition problems involving time-series signals such as Electrocardiography data [17], [32], particularly where the features of interest can occur at any point in time in a given signal. By contrast, conventional distance metrics and machine learning techniques do not perform well when the position of the target signal is highly variable, as confirmed by the results presented here. Therefore, it is far more common to apply traditional machine learning techniques to derived features that are independent of time, such as frequency domain recognition features. However, by using a similar technique to our previous work [8], [9], we have also shown that a combination of Fast Fourier Transform (FFT) features and non-convolutional neural networks are still

outperformed by networks using simpler velocity features. This disparity in performance is likely due to the increased variability of pathological nystagmus obscuring informative frequency components. This explanation is supported by previous work [33], where it was also suggested that common sources of signal noise can mask or imitate the presence the nystagmus.

Although neural networks have previously been applied to several tasks involving eye-movement signals, such as classifying normal versus abnormal nystagmus during caloric tests [16] and detecting saccades [34], this study is the first example of 1D CNNs applied to the task of detecting entire nystagmus waveforms from within hours of normal eye-movement data. While heuristic approaches to detecting optokinetic nystagmus have been shown to yield high levels of classification accuracy (89.13% sensitivity and 98.54% specificity in [10], and 93% accuracy in [12]), these results are not comparable with our study as the data was captured during optokinetic tests and are extremely short in duration (8 seconds each in [10], compared to up to 24 hours in our longest example and almost an hour in the shortest). While it is impressive that [10] were able to extract and analyse eye-movement signals from young children in a laboratory setting, the constrained detection task described is very different to identifying nystagmus within many hours of eye-movement data.

Another factor separating our study from others is that over half of the data used was captured in the community, rather than a clinical setting. Capturing data in ‘real world’ conditions may be affected by motion artefacts, incorrect donning of the monitoring device, by measurable differences between spontaneous and induced nystagmus, or by the increased variability of continuous, long-term eye movements. By contrast, nystagmus captured during caloric testing is usually uninterrupted, the data capture process is monitored by a professional, and is not subject to the same sources of real-world ‘interference’. Therefore, our results are a first step towards reliable detection of nystagmus in long-term eye-movement data, although there is evidently much room for improvement.

The performance we demonstrate for subject 4, the subject who wore the device for 24 hours, is the lowest of all test subjects presented. For the experiments giving the highest average F1 score overall, we were able to identify nearly a third of subject 4’s nystagmus (44 frames), but at the expense of nearly four times the number of false positive detections (167 frames). At first glance, this might seem like a disappointing result, however, a further 8647 true negative detections were made. Thus, we were able to identify a significant proportion of pathological nystagmus buried within vast and highly variable eye movement data, with only a small proportion of true positive detections. It should also be noted that even an apparently low F1 score of 0.24 actually represents performance that could not be obtained through guessing.

The two lowest F1 scores were produced by the two longest data files, suggesting that performance, specifically the number of false positive detections, is correlated with the total duration of eye-movement. To explore this further, we visualised the false positive and false negative detections

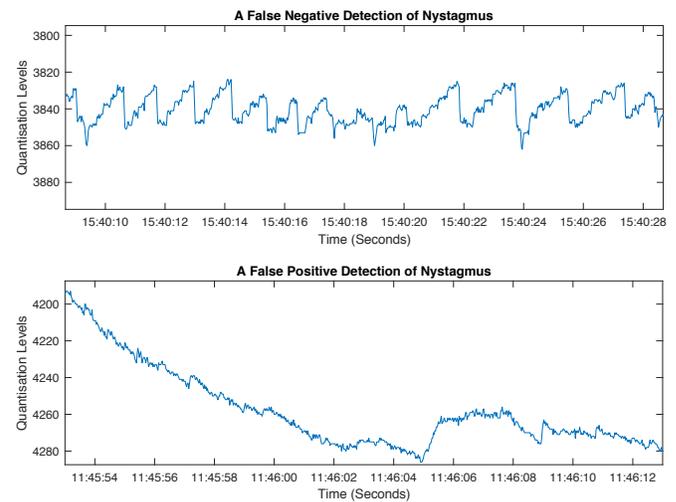


Fig. 6. The top panel shows a false positive detection, in which the signal appears nystagmus-like, but with noisy and ‘stepped’ slow phases that are likely to have been produced by the subject reading. The bottom panel shows an example of a false negative classification of right-beating nystagmus. The signal contains two short periods of low amplitude nystagmus, the first ending at about 11:46:00 and the next starting approximately nine seconds later.

(Fig. 6). False negatives, such as the example shown in the bottom panel of Fig. 6, were subtle, containing low amplitude nystagmus concealed by relatively high levels of background noise. Analysis of one of the false positive detections for subject 4 (top panel of Fig. 6), revealed a period of reading that was misidentified as nystagmus and which is redolent of some examples genuine nystagmus, such as that shown in Fig. 2. This signal is very similar to that of nystagmus, except that the slow phase is characterised by short saccadic motions, moving from left-to-right, corresponding to the eyes reading each word on a line of text. We expect that correctly identifying examples such as these may be possible by training the network with more representative training data. These results highlight the challenges posed by real world data compared to data obtained in a laboratory setting, and suggest a sensible focus for future work.

Our experimental framework was designed around a blinded recognition experiment that we will undertake at the end of an ongoing clinical investigation. In this experiment, our algorithm will be presented with around 400 separate data files, each file containing a day’s worth of eye movement data, and will determine which of these files contain positive examples of nystagmus. Each day will be classified as containing a positive example of nystagmus if any frames within that day are positively classified as nystagmus. Therefore, for this task, higher specificity for frame-level classification is preferred, since any number of false positive frames would lead to a false positive ‘day’. The ROCs for each testing fold (Fig. 5) showed that all classifiers performed well across a range of classification thresholds, showing that the system could be configured to favour sensitivity or specificity, depending on the requirements of a given task. For example, initial screening tests usually favour sensitivity, while increased specificity is more appropriate for invasive follow-up procedures.

V. CONCLUSION

In this article we have demonstrated techniques for overcoming the limited availability of data for training neural networks to detect nystagmus. This is the first reported application of the use of deep neural networks for this task. The results have shown that despite very limited amounts of training data, it is possible to overcome large class imbalances by generating new examples of training data from existing examples. Although we only achieved moderate frame-level accuracy, tuning our system to provide higher levels of sensitivity is likely to provide adequate results for a potential screening application.

Although these techniques have proven capable of achieving moderate levels of accuracy for detecting nystagmus, our next goal is to evaluate them on a much larger dataset, and also to compare the current results to those obtained when training networks using larger quantities of genuine data. Over the remainder of our current clinical investigation, we will capture a wealth of data from patients suffering from dizziness and vertigo. That data will be subject to a blinded analysis, where the task will be to automatically detect the days on which patients reported experiencing episodes of dizziness or vertigo. The models used for that analysis will be similar to those described here, thus providing a challenging and thorough evaluation of these techniques. An additional challenge posed by this task is the inclusion of patients with Benign Paroxysmal Positional Vertigo (BPPV), whose nystagmus may contain a large component of vertical eye movement. Although in our previous clinical investigation we established that CAVA was capable of capturing vertical eye movements, it has been shown that the voltage resolution of vertical electrooculography data is lower than for the horizontal channel [35]. Therefore, it will be interesting to evaluate how this impacts upon our algorithm's capabilities to detect nystagmus in the vertical plane.

In parallel to our clinical investigation, we intend to explore and evaluate a range of other contemporary machine learning approaches for this classification task. For example, we wonder whether Generative Adversarial Networks (GANs) could be used to augment our limited volumes of training data, perhaps in place of SMOTE. GANs essentially work by pitching two neural networks against each other; one to generate artificial examples of the positive class (the "generator"), and one to learn to distinguish between genuine examples from those produced by the generator (the "discriminator"). In doing so, GANs could learn to produce new yet realistic examples of nystagmus with which to train our DNNs. There are also a number of variations to neural networks which we would like to evaluate and which have shown to provide incremental improvements when applied to other classification problems. For example, ResNet and DenseNet are approaches to neural networks which seek to overcome the vanishing gradient problem, whereby network weights can become so small that all or part of a network will stop training. 2D convolutional neural networks have also been used in cardiac arrhythmia detection with good results.

Following the completion of our clinical investigation, we will have a large dataset of patient data available to us

with which we can further evaluate and develop the methods described here. A longer term aim is to apply this system to vertigo resulting from a variety of defined inner-ear conditions, and to quantify the characteristics of nystagmus from them, with a view to determining whether different pathologies can be discriminated on the basis of the nystagmus signals they produce. Our ultimate aim is to develop a complete medical system to allow clinicians to assess dizzy patients purely on the data provided by the CAVA system. In this regard, we also intend to extend our system to provide a more detailed analysis of a patient's nystagmus, by automatically extracting parameters such as slow and fast phase velocity. This innovation has the potential to improve the speed and accuracy of diagnosis for patients reporting dizziness and vertigo, by providing an objective record of a patient's dizzy episodes over the course of a month.

ACKNOWLEDGMENT

The authors would like to acknowledge colleagues in the School of Computing Sciences at UEA for constructive conversations. We would also like to acknowledge Wright Design Limited who worked with us to design and develop the CAVA device. The MRC reviewed the study design but were not involved with any other aspects of this work.

REFERENCES

- [1] Sonia Sandhaus. Stop the spinning: Diagnosing and managing vertigo. *The Nurse Practitioner*, 27(8), 2002.
- [2] Tsutomu Nakashima, Ilmari Pyykkö, Megan A. Arroll, Margaretha L. Casselbrant, Carol A. Foster, Nauman F. Manzoor, Cliff A. Megerian, Shinji Naganawa, and Yi-Ho Young. Meniere's disease. *Nature Reviews Disease Primers*, 2(1):16028, 2016.
- [3] Leif Erik Walther. Current diagnostic procedures for diagnosing vertigo and dizziness. *GMS Curr Top Otorhinolaryngol Head Neck Surg*, 16:Doc02, 2017.
- [4] Yangming Leng, Bo Liu, Renhong Zhou, Jingjing Liu, Dongdong Liu, Su-Lin Zhang, and Wei-Jia Kong. Repeated courses of intratympanic dexamethasone injection are effective for intractable meniere's disease. *Acta Oto-Laryngologica*, 137(2):154–160, 2017. PMID: 27650470.
- [5] John Phillips, Neil Longridge, Arthur Mallinson, and Gordon Robinson. Migraine and vertigo: A marriage of convenience? *Headache: The Journal of Head and Face Pain*, 50(8):1362–1365, 2010.
- [6] Debbie Fife and John E Fitzgerald. Do patients with benign paroxysmal positional vertigo receive prompt treatment? analysis of waiting times and human and financial costs associated with current practice ¿reciben tratamiento oportuno los pacientes con vértigo postural paroxístico benigno? análisis del tiempo de espera y del costo humano y financiero asociado con la práctica actual. *International Journal of Audiology*, 44(1):50–57, 2005.
- [7] Jill Trato and Eric G Johnson. Differential diagnosis and management of a patient with peripheral vestibular and central nervous system disorders: a case study. *J Man Manip Ther*, 18(3):159–165, Sep 2010.
- [8] 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):10452, 2019.
- [9] Jacob L. Newman, John S. Phillips, Stephen J. Cox, John FitzGerald, and Andrew Bath. Automatic nystagmus detection and quantification in long-term continuous eye-movement data. *Computers in Biology and Medicine*, 114:103448, 2019.
- [10] Mehrdad Sangi, Benjamin Thompson, and Jason Turuwenua. An optokinetic nystagmus detection method for use with young children. *IEEE J Transl Eng Health Med*, 3:1600110, 2015.
- [11] Tomasz Pander, Robert Czabanski, Tomasz Przybyla, and Dorota Pojda-Wilczek. An automatic saccadic eye movement detection in an optokinetic nystagmus signal. *Biomed Tech (Berl)*, 59(6):529–543, Dec 2014.

- [12] Jason Turuwhenua, Tzu-Ying Yu, Zan Mazharullah, and Benjamin Thompson. A method for detecting optokinetic nystagmus based on the optic flow of the limbus. *Vision Research*, 103:75 – 82, 2014.
- [13] T. Charoenpong, P. Pattrapisetwong, and V. Mahasithiwat. A new method to detect nystagmus for vertigo diagnosis system by eye movement velocity. In *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, pages 174–177, May 2015.
- [14] A. Ben Slama, A. Mouelhi, M. A. Cherni, S. Manoubi, C. Mbarek, H. Trabelsi, and M. Sayadi. Features extraction for medical characterization of nystagmus. In *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 292–296, March 2016.
- [15] Sai Akanksha Punuganti and Jorge Otero-Millan PhD. *Detection of Saccades and Quick-Phases in Eye Movement Recordings with Nystagmus*. PhD thesis, 2020.
- [16] 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.
- [17] Özal Yıldırım, Paweł Plawiak, Ru-San Tan, and U. Rajendra Acharya. Arrhythmia detection using deep convolutional neural network with long duration ecg signals. *Computers in Biology and Medicine*, 102:411 – 420, 2018.
- [18] Özal Yıldırım, Ulas Baran Baloglu, and U. Rajendra Acharya. A deep convolutional neural network model for automated identification of abnormal eeg signals. *Neural Computing and Applications*, 2018.
- [19] C. Park, G. Choi, J. Kim, S. Kim, T. Kim, K. Min, K. Jung, and J. Chong. Epileptic seizure detection for multi-channel eeg with deep convolutional neural network. In *2018 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–5, 2018.
- [20] U. Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli. Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals. *Computers in Biology and Medicine*, 100:270 – 278, 2018.
- [21] S. Chauhan and L. Vig. Anomaly detection in ecg time signals via deep long short-term memory networks. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–7, 2015.
- [22] M. Zihlmann, D. Perekrestenko, and M. Tschannen. Convolutional recurrent neural networks for electrocardiogram classification. In *2017 Computing in Cardiology (CinC)*, pages 1–4, 2017.
- [23] François Chollet et al. Keras. <https://keras.io>, 2015.
- [24] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6(5):429–449, October 2002.
- [25] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29, June 2004.
- [26] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249 – 259, 2018.
- [27] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang, editors, *Advances in Intelligent Computing*, pages 878–887, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.
- [28] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002.
- [29] J. Margeta, A. Criminisi, R. Cabrera Lozoya, D.C. Lee, and N. Ayache. Fine-tuned convolutional neural nets for cardiac mri acquisition plane recognition. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 5(5):339–349, 2017.
- [30] Andrew Bangham, Richard Harvey, Paul D. Ling, and Richard Aldridge. Morphological scale-space preserving transforms in many dimensions. *Journal of Electronic Imaging*, 5, 11 1997.
- [31] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.
- [32] D. Li, J. Zhang, Q. Zhang, and X. Wei. Classification of ecg signals based on 1d convolution neural network. In *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6, 2017.
- [33] Sai Akanksha Punuganti. *AUTOMATIC DETECTION OF NYSTAGMUS IN BEDSIDE VOG RECORDINGS FROM PATIENTS WITH VERTIGO*. PhD thesis, Johns Hopkins University, 2019.
- [34] Raimondas Zemblys, Diederick C. Niehorster, and Kenneth Holmqvist. gazenet: End-to-end eye-movement event detection with deep neural networks. *Behavior Research Methods*, 51(2):840–864, 2019.
- [35] A. B. Usakli and S. Gurkan. *Design of a Novel Efficient Human-Computer Interface: An Electrooculogram Based Virtual Keyboard*. PhD thesis, 2010.