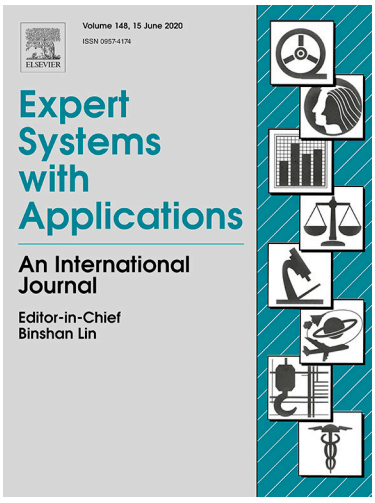

Artificial Neural Networks for classifying the time series sensor data generated by medical detection dogs

Lucy Withington, David Diaz Pardo de Vera, Claire Guest, Clara Mancini, Paul Piwek

PII: S0957-4174(21)00970-2
DOI: https://doi.org/10.1016/j.eswa.2021.115564
Reference: ESWA 115564

To appear in: *Expert Systems with Applications*

Received Date: 12 March 2020
Revised Date: 24 June 2021
Accepted Date: 4 July 2021

Please cite this article as: Withington, L., Diaz Pardo de Vera, D., Guest, C., Mancini, C., Piwek, P., Artificial Neural Networks for classifying the time series sensor data generated by medical detection dogs, *Expert Systems with Applications* (2021), doi: https://doi.org/10.1016/j.eswa.2021.115564

Artificial Neural Networks for classifying the time series sensor data generated by medical detection dogs

Lucy Withington, The Open University, Milton Keynes, UK. tsc.aci@outlook.com

David Diaz Pardo de Vera, Animal-Computer Interaction Lab, The Open University, Milton Keynes, UK. daviddpdv@gmail.com

Claire Guest, Medical Detection Dogs, Great Horwood, UK, Claire.Guest@medicaldetectiondogs.org.uk

Clara Mancini, Animal-Computer Interaction Lab, The Open University, Milton Keynes, UK. clara.mancini@open.ac.uk

Paul Piwek, The Open University, Milton Keynes, UK. paul.piwek@open.ac.uk

Corresponding author

Lucy Withington

tsc.aci@outlook.com

+44 7771 826523

Address:

Care of: Dr Paul Piwek
Jennie Lee Building
The Open University
Walton Hall
Milton Keynes
Buckinghamshire
MK7 6AA
UK

Ethics statement

The Canine Interfaces for Bio-Detection Dogs project, including data collection, was conducted in full compliance with the European Directive 2010/63/EU, On the Protection of Animals Used for Scientific Purposes. Additionally, the project complied with the ethics protocol that informs all research conducted at the Animal-Computer Interaction Laboratory (Mancini, 2017), which requires that animal participants be given the same level of protection and care given to any human participant. The project, including its data collection methods, were approved by The Open University's Animal Welfare Ethical Review Body.

Abstract

The aim of this research was to discover if artificial neural networks can be used to classify pressure sensor data generated by medical detection dogs as they sniff biological samples.

A detection dog can be trained to recognise the odour emitted by one of a wide range of diseases such as prostate cancer, malaria or, potentially, COVID-19. The dog searches a row of sample pots and indicates a positive sample by sitting in front of it. This offers a non-invasive means of diagnosing the specific cancer or disease that the dog has been trained to recognise. For this study, pressure sensors were attached to the sample pots to generate time series data pertaining to the dog's searching behaviour as they press their nose against the sample pot to sniff its content. Automatic classification could provide a second form of indication, to support or refute the dog's explicit signal (to sit at a positive sample), which is not always correct. Ultimately, classification software could eliminate the need for the dog to perform an indication gesture, making the dog's task easier and training quicker.

Four different neural network architectures were evaluated: multilayer perceptron (MLP), a convolutional neural network (CNN), a fully convolutional network (FCN) and ResNet (a deep convolutional neural network). Each model was trained to classify the pressure data generated by medical detection dogs. To achieve a useful level of accuracy, it was found that the models needed to be trained using only those data samples where the dog had correctly classified the scent sample.

Model hyperparameters were tuned to improve accuracy. We found that the best performing model was MLP. When tested on previously unseen data, where the dog was not always correct, the classification performance of the MLP approached that of the medical detection dogs. For our particular dataset, the model's true positive rate (i.e. recall) was 59%, matching that of the dogs. The model's true negative rate was 79%, compared to the dogs' 91%.

Keywords

- Animal Computer Interaction
- Time series classification
- Classification algorithms
- Intelligent system
- Convolutional neural network
- Sensors

1. Introduction

Medical detection dogs are trained to recognise the odour of volatile organic compounds that are emitted by biological samples provided by patients with certain diseases. Dogs have

been trained to recognise prostate cancer from urine samples, malaria from clothing worn by carriers of the disease (Guest et al., 2019) and, most recently, dogs are being trained to detect COVID-19 (London School of Hygiene & Tropical Medicine, 2020; Straiton, 2020). The detection dog is taught to indicate a positive sample by sitting in front of it[1]. This requires the dog to translate their spontaneous response to the odour into a human signalling convention (the sit). This translation limits the signal's reliability (Mancini, Harris, Aengenheister, & Guest, 2015). In studies by Johnston-Wilder et al. (2015) and Mancini et al. (2015), pressure sensors were attached to the sample pots and collected data on the sniffing behaviour of medical detection dogs. A visual inspection of the data suggested that it may be possible to classify a sample as positive or negative from this data alone, without the need for the dog to sit (Johnston-Wilder et al., 2015). This proposal is investigated in the research presented here. Our research aims to find out whether the dogs' spontaneous response can be classified directly, using artificial neural networks, from the univariate time series data of a pressure sensor on the sample pot.

Neural networks are computational models inspired by the biological neuron. They have been adopted for time series classification (Nweke, Teh, Al-Garadi, & Alo, 2018 and Wang, Yan, & Oates, 2017b) and have a wide range of applications: human activity recognition from smartphone data can aid assisted living (Ronao & Cho, 2016); animal behaviour classification from accelerometer data informs the study of the effect of environmental factors on wild animals (Brewster et al., 2018; Nathan et al., 2012). To our knowledge, this is the first study to apply artificial neural networks to automatically interpret detection dogs' responses to scent samples based on pressure sensor data.

A range of other sensors and instruments have been used, sometimes in combination with other statistical or machine learning techniques, to capture canine behavioural and physiological responses to olfactory stimuli: the use of accelerometers and gyroscopes to identify sniffing activity based on nose position, with k-Nearest Neighbours algorithm for data analysis (Ladha et al., 2013); hot-film probes to measure airflow during sniffing and inform a computational fluid dynamics simulation (Craven et al., 2010); stethoscopes on the rib cage and throat to extract ECG and acoustic features of sniffing behaviour with manual statistical analysis and inspection of data visualisations (Brugarolas et al., 2019); manual video analysis and statistical analysis to study sniffing frequency and duration (Concha et al., 2014); and magnetic resonance imaging (MRI) to identify differences in the activation of different brain regions in response to low and high odour intensity (Jia et al., 2014). However, these methods are too coarse (unable to characterise response patterns), too

[1] An informative video, by OUresearch on YouTube, can be viewed at https://youtu.be/MyHjq8Od-Xg

intrusive (disruptive of sniffing activity) or too impractical (unusable in-situ). Instead, we used pressure sensors to capture performance-based signals (i.e. signals whose intensity is directly determined by the activity being performed) that are sufficiently fine-grained and can be captured in situ without constraining or interfering with the dogs' detection activity. Our early work with a prototype system that uses pressure sensors to capture the dogs' spontaneous physical interaction with the samples (Mancini et al., 2015) suggested that highly detailed data patterns might be identified which correlate with sample concentration (Johnston-Wilder et al., 2015). Careful inspection of close-up HD video recordings of the dogs' sniffing behaviour revealed that every contact perceptible on video was reflected in intricate pressure signals obtained from the sensors. The work presented here is the first to apply artificial neural networks to the analysis of this data towards the precise identification of such patterns. The neural network architectures recommended by Wang et al. (2017b) were applied to this novel dataset.

This research examined whether pressure sensor data is an effective indicator of a detection dog's response and therefore an indicator of a positive sample. It has the potential to improve the accuracy of canine bio-detection, strengthening its viability as a clinical diagnostic test. In the case of prostate cancer, for example, detection via a urine sample would be preferable to the invasive biopsy that is currently used for diagnosis.

The rest of this paper is organised as follows: Section 2 reviews time series classification neural networks in the literature and select architectures suitable for this application. Section 3 provides background information on the process of training and working with medical detection dogs. In Section 4 the method of collecting the detection dogs' data is described and the selected neural networks are detailed. In Section 5 the accuracy of the neural networks in classifying the data samples is presented. A discussion of the insights afforded by this research then follows in Section 6. Finally, a summary of the results, our conclusions and recommendations for further work are given in Section 7.

2. Related work

Time series classification (TSC) has long been performed using features-based machine learning methods (Nweke et al., 2018). This requires features such as the mean, zero-crossings, signal magnitude area and dominant frequency (Figo, Diniz, Ferreira, & Cardoso, 2010) to be extracted from the data and used as an input to a classification algorithm. There is no universal set of features that is appropriate for every domain and often the features do not generalise well between data samples from different sources (Nweke et al., 2018). In the case of medical detection dogs, different dogs might generate different features. In contrast, a neural network automatically discovers the discriminating features, and the same

procedure can be applied to a wide variety of datasets (LeCun, Bengio, & Hinton, 2015). In the past few years, artificial neural networks have proved superior to conventional machine-learning techniques in a diversity of fields including classification of images and of speech (LeCun et al., 2015). Inspired by this, researchers have started to apply neural networks to TSC.

A public repository of time series datasets, the University of California, Riverside (UCR) TSC archive (Dau et al., 2018) is extensively employed by researchers to evaluate time series classifiers. In the more established field of non-neural-network TSC, the accepted baseline is the 1-Nearest Neighbour classifier (1-NN) and the current state-of-the-art classifier is The Collective of Transformation-Based Ensembles (COTE) (Bagnall, Lines, Bostrom, Large, & Keogh, 2017). The accuracy of these classifiers was compared to that of artificial neural networks (ANNs) across the range of UCR TSC datasets by two independent research teams, Wang et al. (2017b) and Fawaz, Forestier, Weber, Idoumghar, & Muller (2018). Both found that the accuracy of the best ANN classifier was comparable to that of COTE. However, an ANN has the advantage that, once trained, it is faster than COTE, which needs to revisit its entire training set to perform a nearest neighbour classification (Fawaz et al., 2018).

Different types of ANN architecture for TSC have been proposed and evaluated, but there is no undisputed best choice and the research is ongoing. Moreover, neural networks have not previously been applied to data generated by medical detection dogs. Therefore, this research has selected, evaluated and further developed ANNs suitable for use with this novel dataset.

Wang et al. (2017b) evaluated three ANNs on 44 datasets from the UCR TSC archive: multilayer perceptron (MLP), fully convolutional network (FCN) and ResNet (residual network). MLP is a dense neural network where every node in a layer is connected to every node in the next layer. In contrast, FCN and ResNet are types of convolutional neural network (CNN), which have more sparse connectivity. CNNs have been used successfully for image classification (LeCun et al., 2015). The convolution operations in a CNN are equivariant to a translation in the input data; a shift in the input data shifts the output in the same way. Thus, a CNN is insensitive to the location of features in an image or in time series data. The FCN, by Long, Shelhamer, & Darrell (2015), unlike other types of CNN, does not contain intermediate pooling layers, which would aggregate sections of the time series data as it progresses through the network. ResNet, developed by He, Zhang, Ren, & Sun (2016), is a CNN with shortcut connections that enable data to skip over one or more layers. This enables deeper neural networks (over 100 layers) to be successfully trained.

Wang et al. found that the FCN gave the best accuracy across the datasets, with ResNet giving similarly good results. MLP performed less well but was comparable to the benchmark non-ANN classifier, 1NN-DTW (1-Nearest Neighbour with distance calculated using Dynamic Time Warping).

Fawaz et al. (2018) implemented nine ANN architectures, including the same MLP, FCN and ResNet as Wang et al. (2017b). Fawaz et al. evaluated the ANNs on 85 UCR TSC datasets and found ResNet to be the most accurate, closely followed by the FCN. Given the results of these two studies, the FCN and ResNet, with the hyperparameters used by Wang et al., were selected for application to the medical detection dogs' dataset. A conventional CNN (with intermediate pooling layers but no shortcut connections) was also applied. In addition, the MLP of Wang et al. was evaluated because, as Wang et al. propose, it is a suitable baseline ANN, being the simplest architecture.

In the studies by Wang et al. (2017b) and Fawaz et al. (2018), the ANNs' hyperparameters (number of layers, neurons and training steps, etc.) were not tuned. Indeed, Wang et al. elected not to tune the ANNs' hyperparameters in order to give a "most unbiased baseline" and to simplify deployment. However, neural network performance depends heavily on the hyperparameter settings (van Rijn & Hutter, 2017). Therefore, as part of this study, a hyperparameter search was performed to optimise the ANNs for classification of the detection dogs' data.

3. Background

Volatile organic compounds (VOCs) are released by cancer tissue and are a biomarker of cancer (Willis, Britton, Harris, Wallace, & Guest, 2011), and VOCs emitted by the human body change in response to certain diseases (London School of Hygiene & Tropical Medicine, 2020). Dogs can be trained to recognise and signal the presence of such biomarkers found in bodily fluids such as urine, sweat or breath. The dogs participating in this study were from Medical Detection Dogs, a UK charity with over ten years' experience in selecting and training detection dogs.

During their training, medical detection dogs learn to associate the odour of a VOC with a reward (usually food or play), whereby the odour becomes of interest through a process of classical conditioning. The spontaneous response a dog has when they search a sample and detect an odour of interest is the dog's "stimulus response". In order to communicate that they have detected the odour, the dogs are also trained to sit in front of the sample emanating the odour (positive sample) through a process of operant conditioning. Sitting (or a similar trained signal) is the dog's "operant response" (Mancini et al., 2015). The

requirement for the dog to translate their stimulus response into an operant response puts additional cognitive load on the dog and limits the accuracy of the dog's positive/negative signal (Mancini et al., 2015). To address this issue, the pressure sensors are placed behind the sample pots to record the dogs' spontaneous interaction with the pots as they investigate the content, thus capturing the dogs' stimulus response. Using solutions of amyl acetate in mineral oil allowed us to control VOC concentration levels, which is harder or impossible to do with biological samples. Whatever the VOC, though, the dogs learn to focus on the compound's odour through classical conditioning and until the olfactory stimulus is conditioned (through consistent association with a reward during training) the odour has no particular relevance for them. Once an odour has become relevant, the strength of the dogs' stimulus response depends on the VOC's odour concentration (although the relation is not necessarily linear) as well as other possible confounding odours, which can be controlled for.

4. Methods

4.1. Data collection

Figure 1 shows a dog searching a sample pot through a perforated plate. The pot is held on a pivoted arm which moves against a pressure sensor when the dog makes contact. Figure 2 links to a video of a medical detection dog searching the line of sample pots. He indicates correctly at the second sample and the run ends.

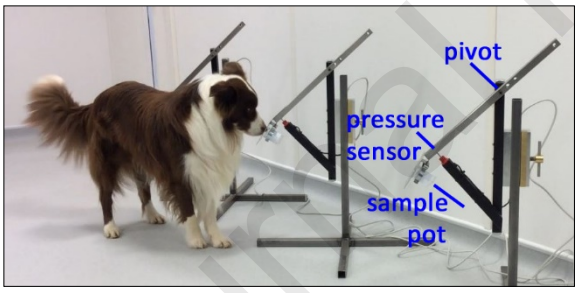

Figure 1 Medical detection dog searching the three samples

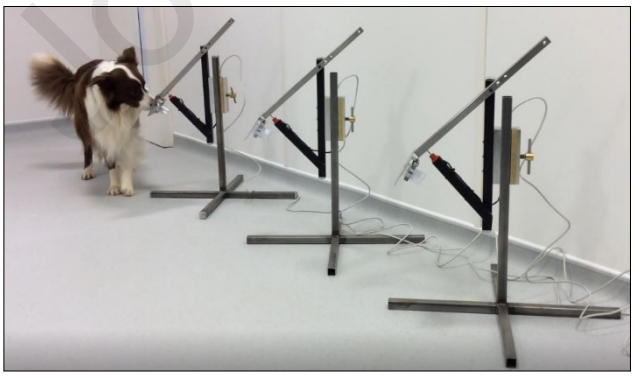

Figure 2 Link to a video showing a medical detection dog searching the line of sample pots.

A detection session consists of a series of passes of the dog down a line of three sample pot stands, separated approximately one metre from each other. The dog is not allowed to return to a sample inspected previously in the same pass. The dog is not usually allowed to continue in a pass after they indicate at a sample, even if they do so at the wrong sample. Occasionally, however, trainers do allow continuation to give the dog a chance to inspect the actual target sample further down the line. Different combinations of three samples are used throughout a session, typically from a pool of eleven control samples (samples containing only mineral oil) and three target samples each containing different concentrations of amyl acetate in mineral oil solution. Amyl acetate was used instead of biological samples because it afforded greater control (Johnston-Wilder et al., 2015). Different concentrations are used to reflect the variability that would be found in biological samples. The concentration of amyl acetate in mineral oil ranged from 1/1,000 to 1/60 million. Each of these combinations may consist of two controls and a target sample (placed at different positions in the line) or three control samples. A pass with a target sample is considered successful if the dog indicates at the target position. A pass with no target sample is considered successful if the dog inspects all three positions in succession and does not indicate at any of them. If a pass has not been successful, the dog may be sent down the line again, for another attempt. The set of all successive passes performed by the dog for a specific combination of samples is called a run. At the trainer's discretion, and always in case of a successful pass, the current run ends and a new run is set-up with a new combination of samples.

The accuracy of the dogs' operant response is not 100%; they produce some false positives and false negatives. Therefore, there are samples where the dog's search data does not relate to its class and could mislead a neural network during training. To overcome this and improve neural network training, a "dog-correct balanced dataset" was created, containing only those samples where the dog indicated correctly. This dataset was used for training. The dataset used for testing the neural network included samples where the dog indicated incorrectly, as would be the case in a clinical application of this system.

Figures 3 to 5 plot the three pressure sensors as a single dog searched each sample pot in turn. The sampling frequency of the pressure sensor system is 500Hz. In the first two seconds he searched pot 1, he moved to pot 2 at 2 seconds and searched pot 3 from 4 to 8 seconds. The plots show these time windows. Figures 3 and 4 show pressure sensor data relating to negative samples. These plots are narrow. Figure 5 relates to a positive sample. This plot is wider and has four separate peaks. The time series data of pressure sensor data shown in these plots is the input to the neural network classifiers that are introduced in section 4.3.

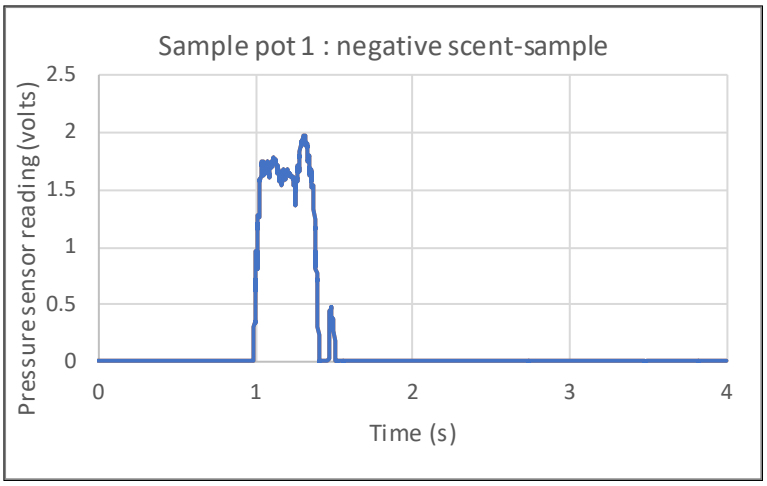

Figure 3 Plot of pressure sensor 1 data for negative scent-sample

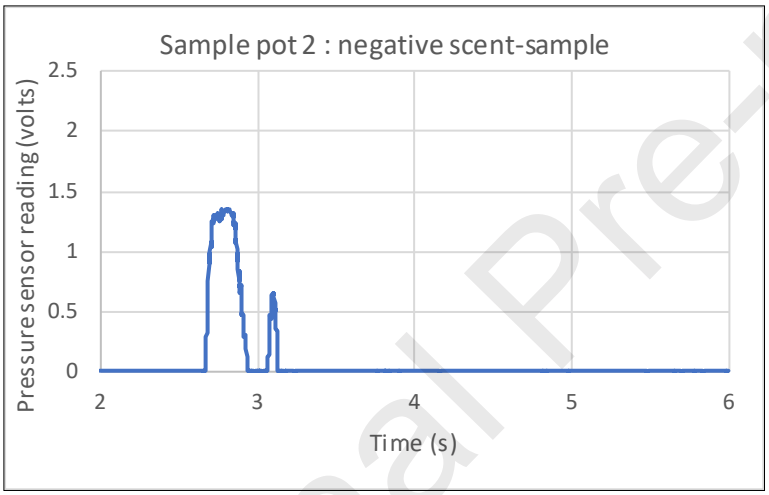

Figure 4 Plot of pressure sensor 2 data for negative scent-sample

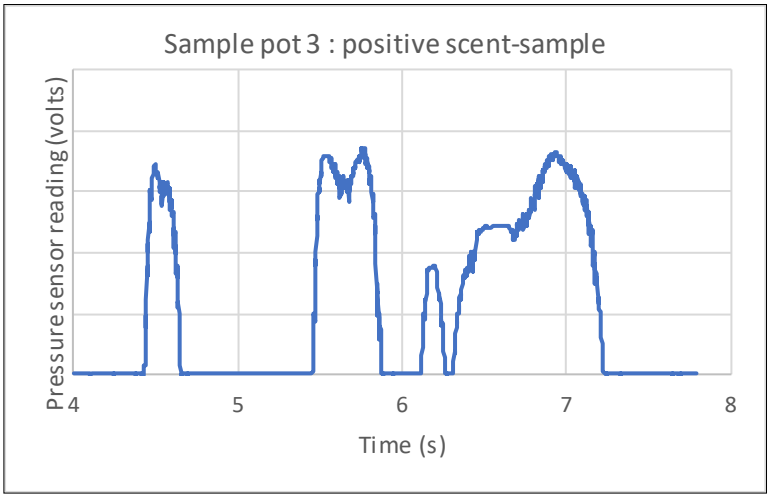

Figure 5 Plot of pressure sensor 3 data for positive scent-sample

The right-hand side of Figure 6 shows the introduction of an ANN classifier. The dog's stimulus response drives the pressure sensor and this data is input to the ANN. The ANN classifier provides a positive/negative result, which would usefully supplement, or even supplant, the dog's operant response. Thus, in this proposed system, the detection dog is still required to sniff and recognise the scent-sample but the ANN replaces the need for the dog to perform an indication when they detect a positive, such as sitting. Dispensing with the dog's operant response would reduce the time it takes to train a detection dog.

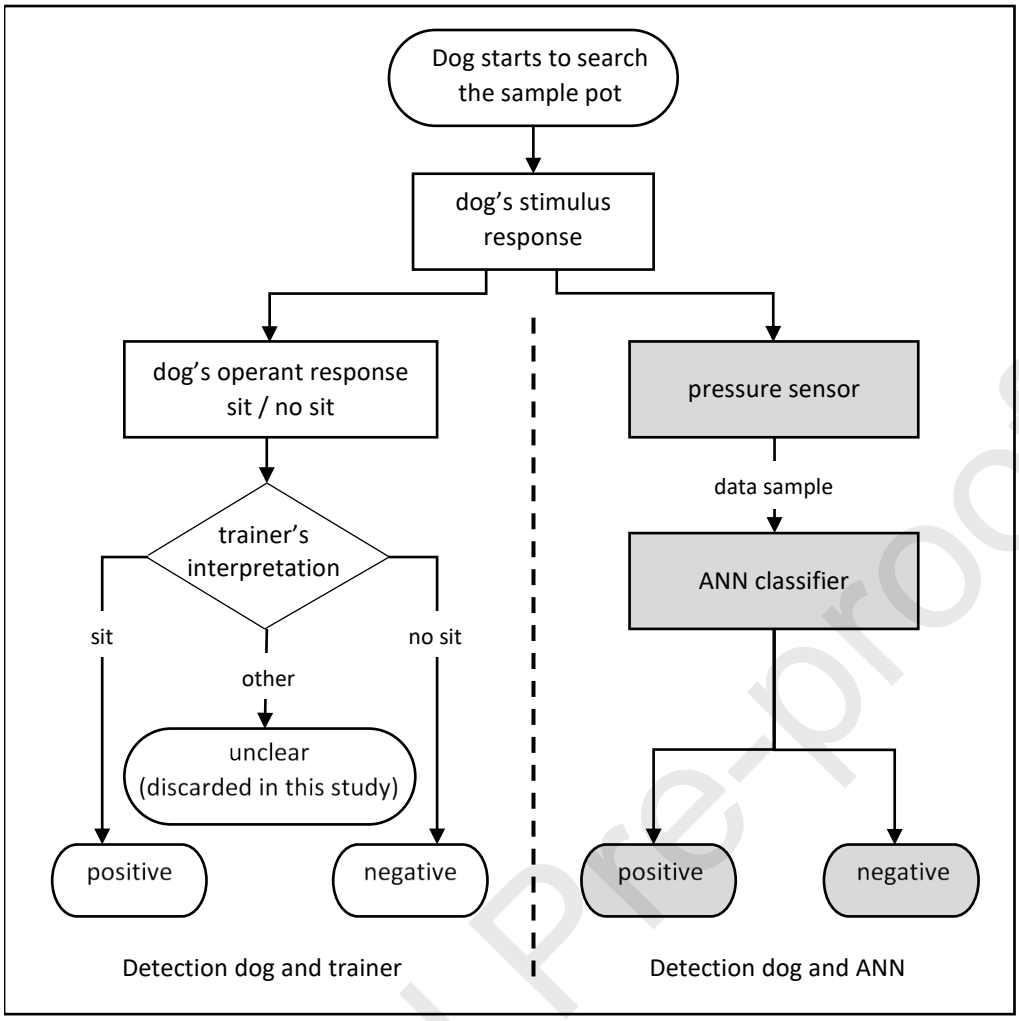

Figure 6 Flow chart showing how the pressure sensor and ANN change the classification process (right) as compared to the current process (left). It also shows the dependency of the results on the accuracy of the dog's stimulus response.

In this research, the ground truth as to whether a sample pot contains a positive or negative sample is known. The detection dog's indication of positive/negative (as interpreted by the dog trainer) is also known, thus the accuracy of the dog's operant response can be calculated. This can then be compared to the accuracy of the ANN classifier's positive/negative result. Both of these results depend upon the accuracy of the dog's stimulus response, which is unknown.

4.2. Data preparation

The duration of a dog's search varied but was typically far shorter than 1,000 data points (2 seconds at 500Hz). Occasionally, the dog's neck hits the sample pot as he sits at a positive sample or he continues to sniff the sample after he has sat, leading to a longer pressure recording. This was the case in Figure 5. The dog pushed his nose against sample pot 3 after he had indicated a positive by sitting. A review of the video of many more runs provided

confidence that the dogs in this study either move on or sit at a sample pot in under two seconds. Therefore, we elected to automate finding these two second pressure sensor event windows and create time series data samples of length 1,000, as MLP, CNN, FCN and ResNet operate on input data of a fixed size.

There is a wide range of techniques for event detection in time series data (Guralnik & Srivastava, 1999). We selected a simple method, finding the first time at which the data sample clearly exceeds the sensor's noise level. The start of an event is defined as the first time, t_i, at which

$$\frac{\sum_{t_i}^{t_{i+n}} x(t_i)}{n} > 0.1$$

where $x(t)$ is the pressure sensor value at time t and $n = 50$. That is, when the average pressure value is greater than 0.1 for 50 data points, which is 0.1 seconds. The event window starts at this detected time and continues for 2 seconds. Thus, the event window is defined as $t_i \leq t < t_{i+1000}$. Figure 7 shows an example. Truncating and retaining only the event window also aligns the data samples; the dog's search commences 0.1 seconds into each 2 second data sample. This 2 second time series of 1,000 data points is a single sample for input to the neural network.

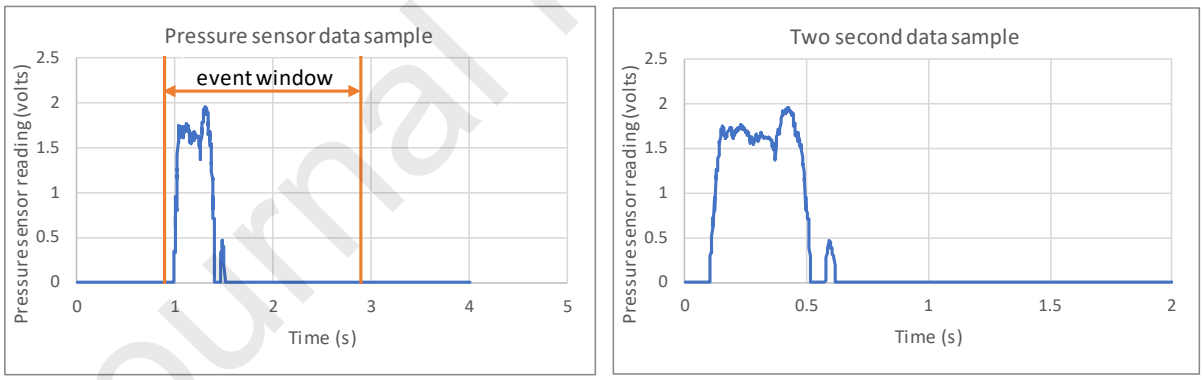

Figure 7 Identification of the event window within a data sample (left). Truncation creates a two second data sample (right)

A final test set of 223 data samples was reserved and not used during model development. A balanced dataset of 312 positive samples and 312 negative samples (624 in total) was used for model development. With a balanced dataset, the accuracy that can be achieved by chance is 50%, which provides a simple baseline for evaluating the accuracy of the classifier. The accuracy of the dogs' operant responses on this balanced dataset was 74%. Data relating to a dog's false positive or false negative could mislead the model during training. To overcome this and improve model training, these data samples were removed,

leaving a “dog-correct balanced dataset” of 284 samples for use during model development. There is, of course, little practical value in a model that can only classify data where we know that the dog has already correctly classified this sample. Therefore, the models were also tested on a subset of the development dataset where the dog did not always indicate correctly (154 samples). We refer to this simply as a “balanced dataset”. Samples in the balanced dataset do not appear in the dog-correct balanced dataset and were not used for training the model.

Iterated k-fold cross validation was employed during model validation (Chollet, 2017, Chapter 4). Typically, 4 iterations of 3-fold cross validation were used, generating 12 test results. To compare models, Wilcoxon’s Signed Rank test was used (Japkowicz & Shah, 2010, Chapter 6). This tests the null hypothesis that the k-fold cross validation results from two models could be drawn from the same distribution. That is, it tests if the performance of one model is different to that of another model. The conventional threshold of 0.05 was used: if the Wilcoxon p-value is greater than 0.05 then the null hypothesis cannot be rejected.

The best performing model configuration was selected as the “final model”. As per best practice, the final model was trained using all available development data (Goodfellow, Bengio, & Courville, 2016, Chapter 7), in this case a dog-correct balanced dataset of 380 samples. This final trained model was then tested using the previously unused final test set of 223 samples (imbalanced, dog not always correct).

4.3. Neural network architectures

We tested the MLP, FCN and ResNet architectures with the hyperparameters used in the software made available by Wang, Yan, & Oates (2017a). These are referred to as the “untuned” models. A CNN was also tested. Its initial architecture was inspired by Ackermann’s (2018) CNN for human activity recognition. The networks are presented in Figures 8 and 9. Each ANN outputs the probability that the input data sample belongs to class 1 (a positive sample). Since this is a binary classification problem, a sigmoid function was used instead of the softmax function that Wang et al. (2017a) needed for the multi-class problems in the UCR TSC archive. As is usual, a threshold probability of 0.5 was applied; if the output probability is greater than 0.5 then class 1 was assigned, otherwise class 0 (a negative sample).

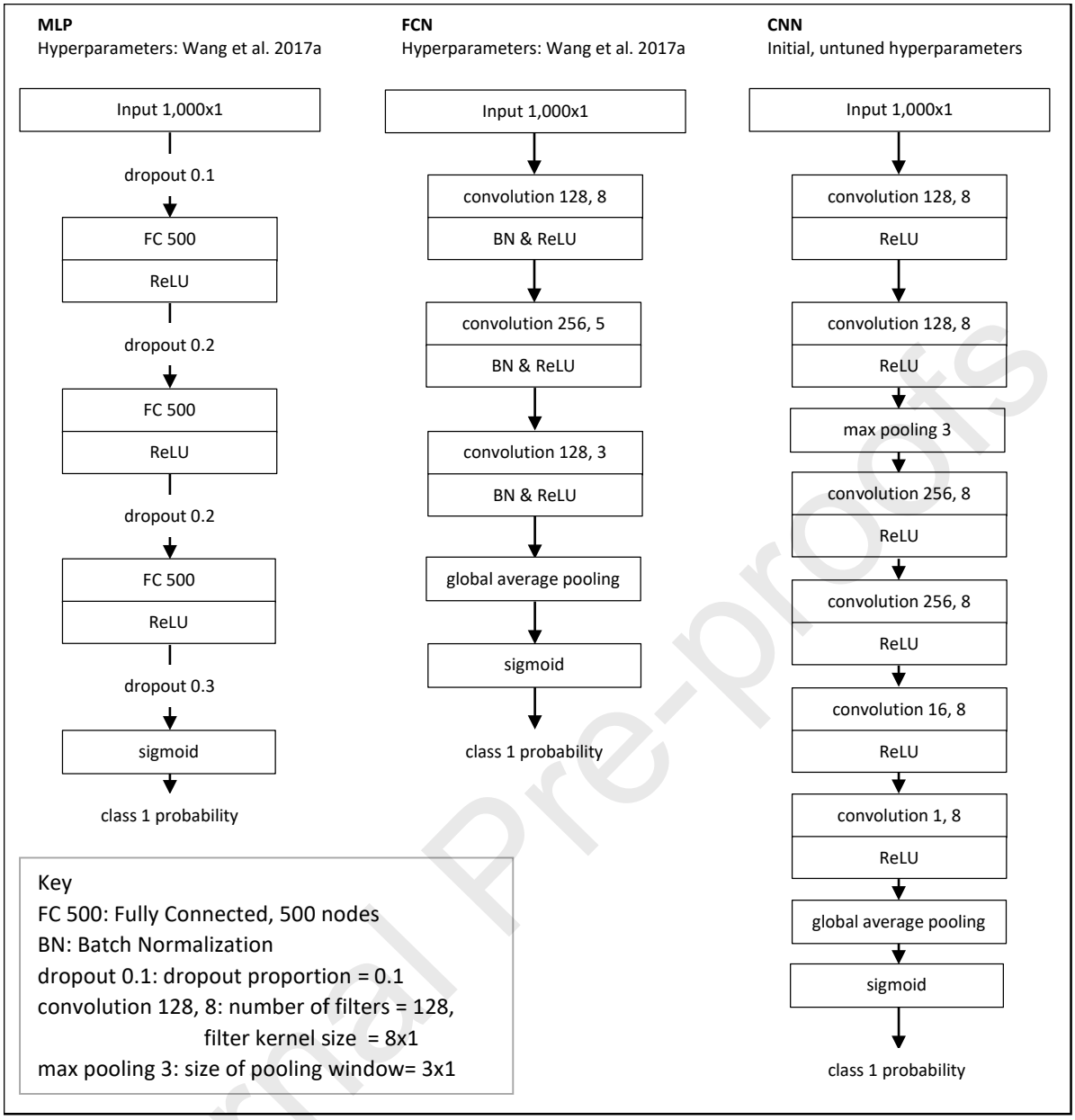

Figure 8 Network architectures of MLP, FCN and CNN (untuned)

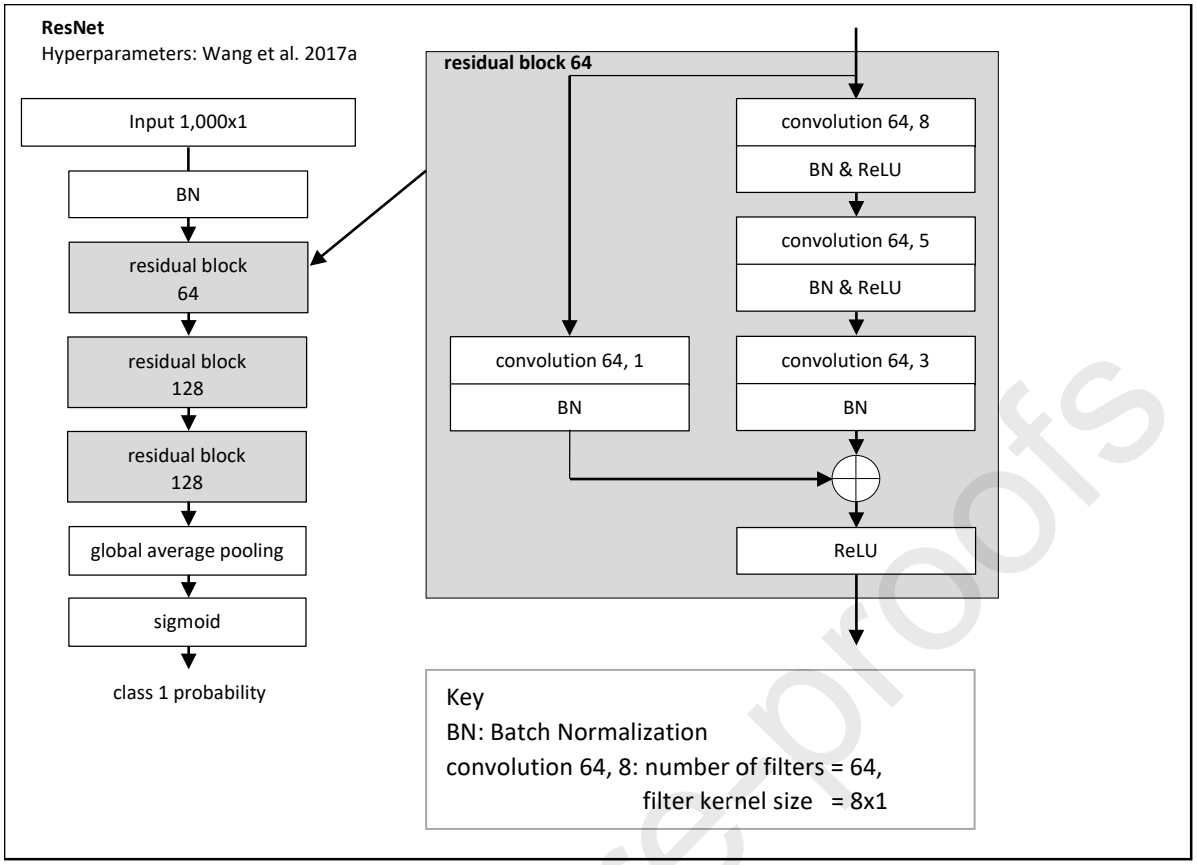

Figure 9 Network architecture of ResNet (untuned)

The input data was standardised using the mean and standard deviation of the training dataset. The ANNs were trained using the Adam optimiser (Kingma & Ba, 2015) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-7}$. The loss function was binary cross entropy. The initial learning rate was $\alpha = 0.001$ and this was halved if there was no reduction in loss after 50 epochs, to a minimum of $\alpha = 0.0001$. In each training run, the model with the lowest loss on the training dataset was selected and its validation accuracy was recorded.

First, the untuned MLP, FCN and ResNet were trained and validated on the dog-correct balanced dataset. Then the MLP, FCN, ResNet and CNN hyperparameters were tuned to maximise accuracy on the same dataset. Hyperparameters that were tuned included the number of layers, number of nodes per layer, number of epochs, convolution filter size and number of convolution filters. In addition, L^2 parameter regularisation (Goodfellow et al., 2016, Chapter 7) was added to MLP, which reduced overfitting.

5. Results

This section shows the result of training and testing the untuned Wang et al. (2017b) models on the dog-correct balanced dataset. The results from the tuned neural networks models,

including the CNN model, follow. The dogs are respectively referred to as dog 0, dog 1 and dog 2.

5.1. Untuned models

Wang et al. (2017b) proposed their models as a default choice for real world applications, without necessarily requiring further tuning for that application. Therefore, we present the accuracy of those untuned models in this section and of the tuned models in the following sections.

Figure 10 shows the accuracy of the models of Wang et al. (2017b) plus the 1-Nearest-Neighbour (1-NN) baseline model when trained and validated using the dog-correct balanced dataset (284 samples). Four iterations of 3-fold cross validation were performed, using 189 samples for training and 95 for validation testing. MLP performed best, with a mean validation accuracy of 77% and sample standard deviation of 3.7%. FCN had a high sample standard deviation (SSD), at 8.9%, and one of the twelve trained FCN models achieved only 50% accuracy on the validation data. ResNet's SSD was lower, at 4.2% but it took around 12 times longer than MLP and FCN to train and validate. These results are detailed in Table 1.

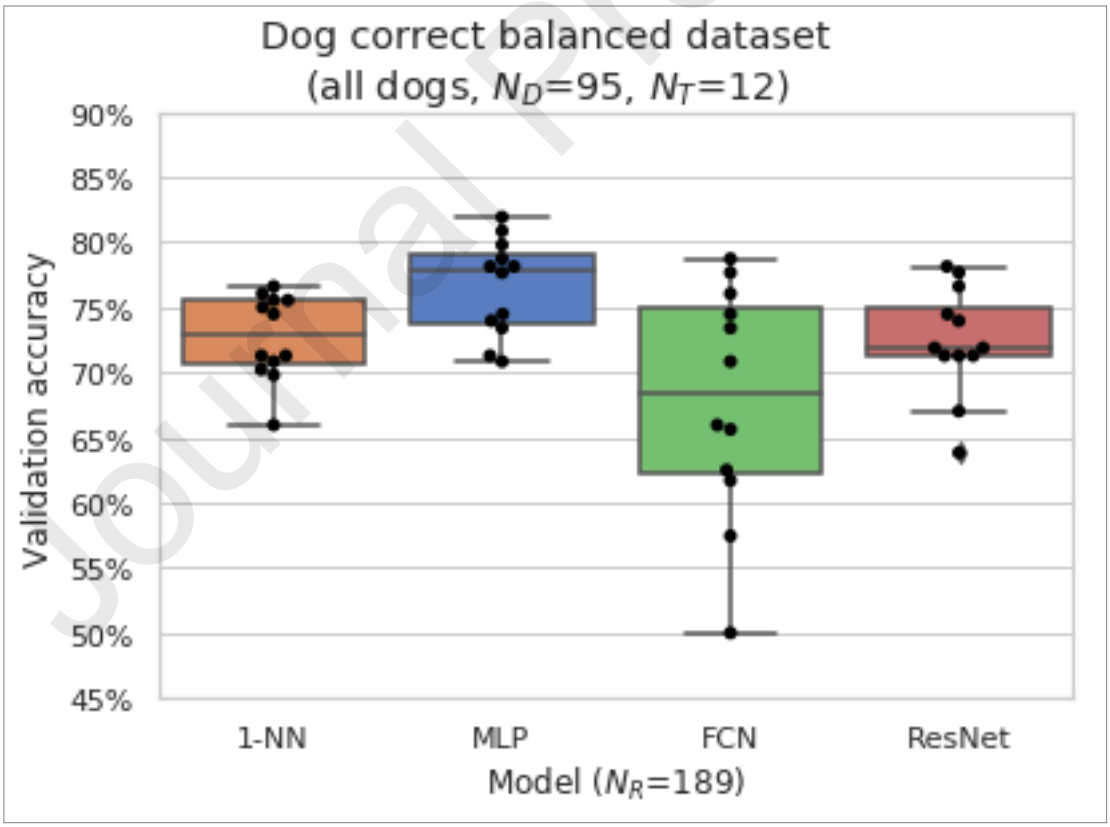

Figure 10 Validation accuracy of the four models. Models trained using data from all dogs, where the dog indicated correctly

The models of Figure 10 were then tested on a balanced dataset of 154 samples that were not used for training or validating the models and where the dogs achieved 76% accuracy. Each type of model was trained and tested twelve times under the four iterations of 3-fold cross validation; Figure 11 and Table 1 show the results, for each model, of the twelve tests on the balanced dataset. For these, the dataset remained the same, unlike the tests of Figure 10 where the dataset varied. This contributed to the lower sample standard deviations seen in Figure 11.

MLP was the most accurate, achieving 65% mean accuracy with an SSD of 2.4%. ResNet accuracy was similar, at 62%. However, the higher SDD (3.7%) gives less confidence that any single ResNet would be accurate.

The Wilcoxon's Signed-Rank test revealed that the MLP's accuracy results are significantly different to those from the other models. In contrast, 1-NN results could be drawn from the same distribution as those of FCN or ResNet. Similarly, FCN results could be drawn from ResNet's results distribution.

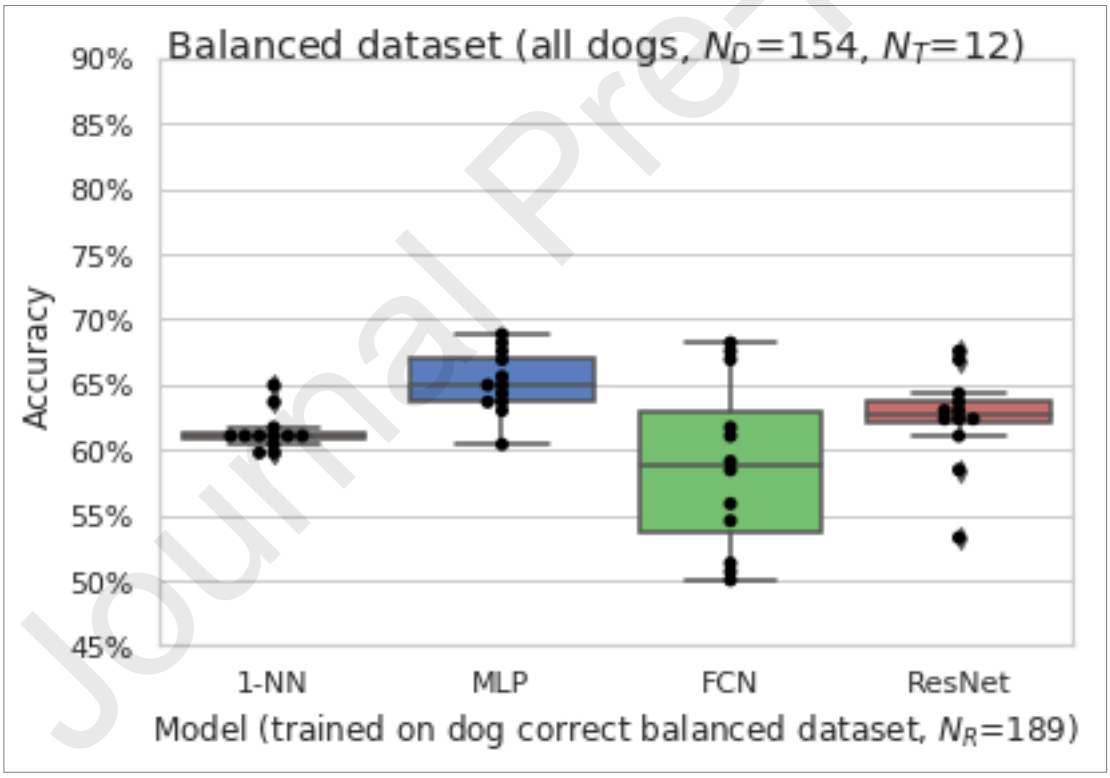

Figure 11 Accuracy of the four models. Models trained using the dog-correct balanced dataset (as per Figure 10) and tested on a balanced dataset where the dog was not always correct

The results shown in Figures 10 and 11 are collated in Table 1.

Table 1 Untuned models trained on the dog-correct balanced dataset (all dogs)

		Dog-correct balanced dataset (all dogs, N_D=95, N_T=12)		**Balanced dataset dog accuracy 76%** (all dogs, N_D=154, N_T=12)	
Model (N_R=189)	**Hyperparameters**	**Mean validation accuracy**	**Sample standard deviation**	**Mean accuracy**	**Sample standard deviation**
1-NN	**untuned**	73%	3.3%	61%	1.5%
MLP	**untuned**	77%	3.7%	65%	2.4%
FCN	**untuned**	68%	8.9%	59%	6.5%
ResNet	**untuned**	72%	4.2%	62%	3.7%

5.2. Tuned models

Model architecture and hyperparameters were changed with the aim of finding a tuned MLP, FCN and ResNet with higher validation accuracy and lower variability than the untuned models. At this stage, the CNN was also introduced and tuned.

The performance of the tuned models (MLP, FCN, ResNet and CNN) on the balanced dataset is shown in Figure 12 and the detail is in Table 2. The models were trained on the dog-correct balanced dataset for all dogs.

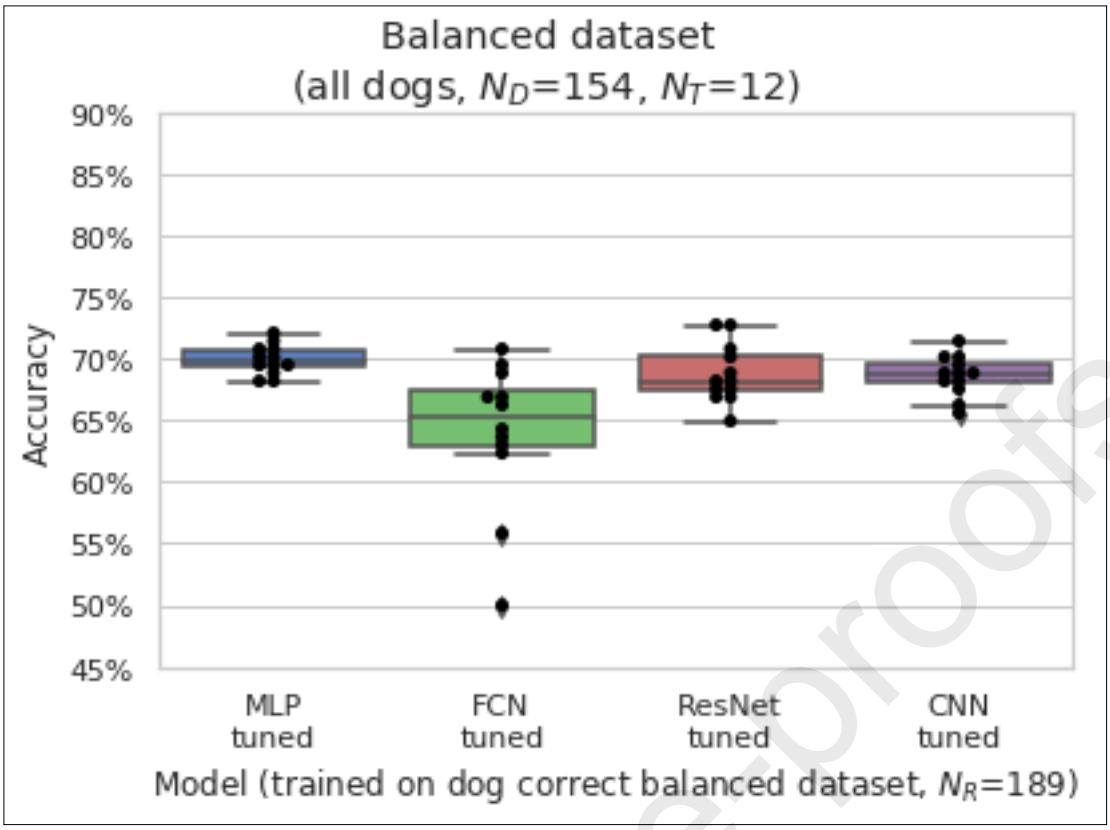

Figure 12 Accuracy of the four ANNs. Models trained using the dog-correct balanced dataset and tested on the balanced dataset

The tuned MLP, FCN and ResNet performed significantly better than their untuned counterparts, which was confirmed by the Wilcoxon's Signed-Rank test.

The average accuracy of the tuned MLP on the balanced dataset was 70% on average and ranged from 68-72%. In comparison, the dogs' accuracy on this dataset was 76%. Tuned MLP was the most accurate, with tuned ResNet a close second. Their Wilcoxon's p-value was 0.18, which is above the 0.05 threshold; there was no significant difference between the two models. The tuned CNN, with a mean accuracy of 68.6%, was only slightly less accurate than ResNet, at 68.8%. There was, however, a great difference in training and testing time with ResNet taking nearly ten times as long as MLP, and CNN taking four times as long as MLP. Therefore, the tuned MLP was selected as the most promising model.

Table 2 Mean accuracy and sample standard deviation of the models trained using the dog-correct balanced dataset (all dogs)

		Dog-correct balanced dataset (all dogs, N_D=95, N_T=12)		**Balanced dataset dog accuracy 76%** (all dogs, N_D=154, N_T=12)	
Model (N_R=189)	**Hyperparameters**	**Mean validation accuracy**	**Sample standard deviation**	**Mean accuracy**	**Sample standard deviation**
1-NN	**N/A**	73%	3.3%	61%	1.5%
MLP	**untuned**	77%	3.7%	65%	2.4%
MLP	**tuned**	83%	2.4%	70%	1.2%
FCN	**untuned**	68%	8.9%	59%	6.5%
FCN	**tuned**	73%	6.8%	64%	5.9%
ResNet	**untuned**	72%	4.2%	62%	3.7%
ResNet	**tuned**	78%	3.4%	69%	2.4%
CNN	**tuned**	77%	3.7%	69%	1.6%

The tuned models and their hyperparameters are shown in Figure 13 and Figure 14. Precise detail along with the model development steps and results is provided in Withington (2019).

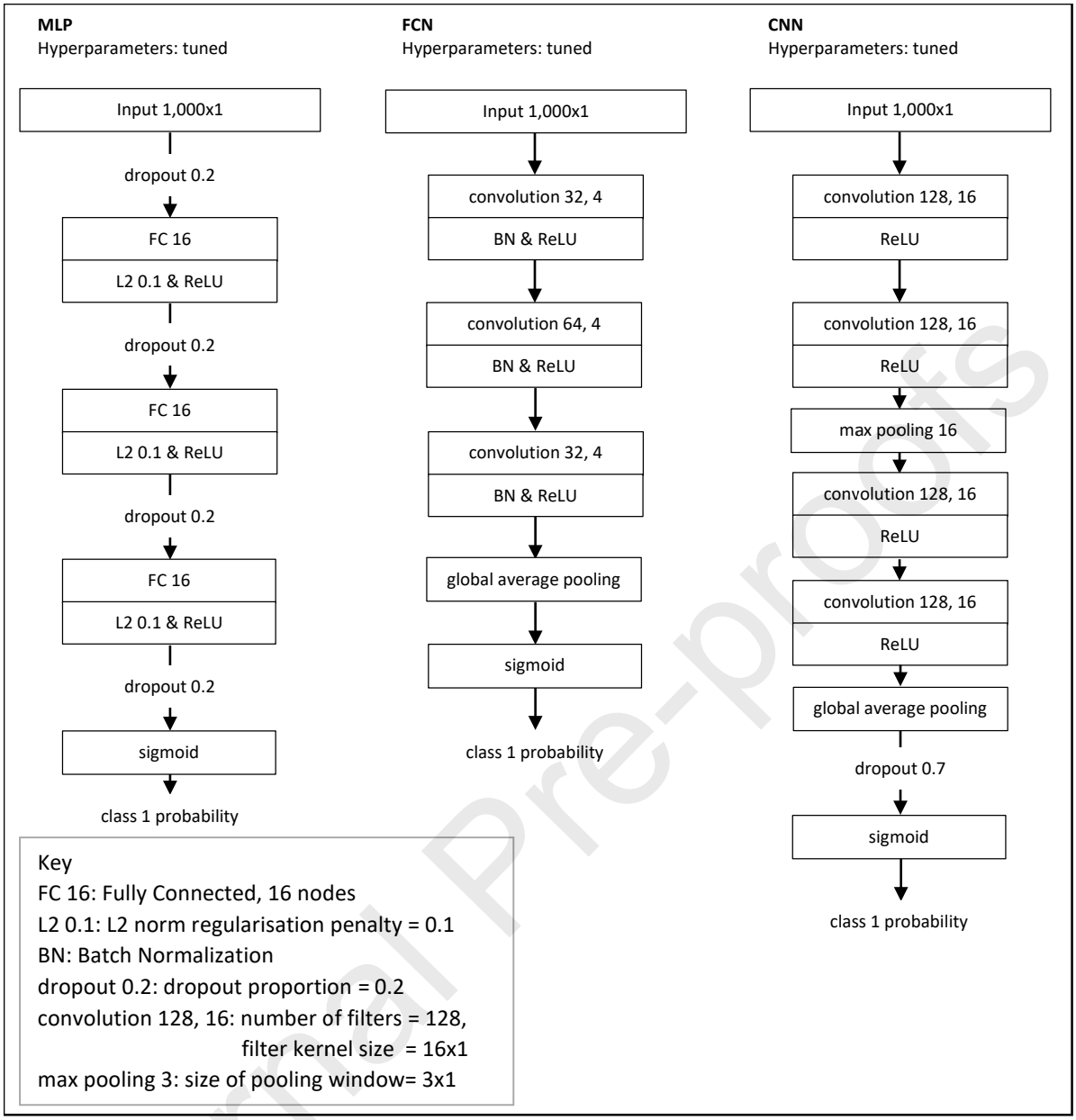

Figure 13 Tuned MLP, FCN and CNN and their hyperparameters

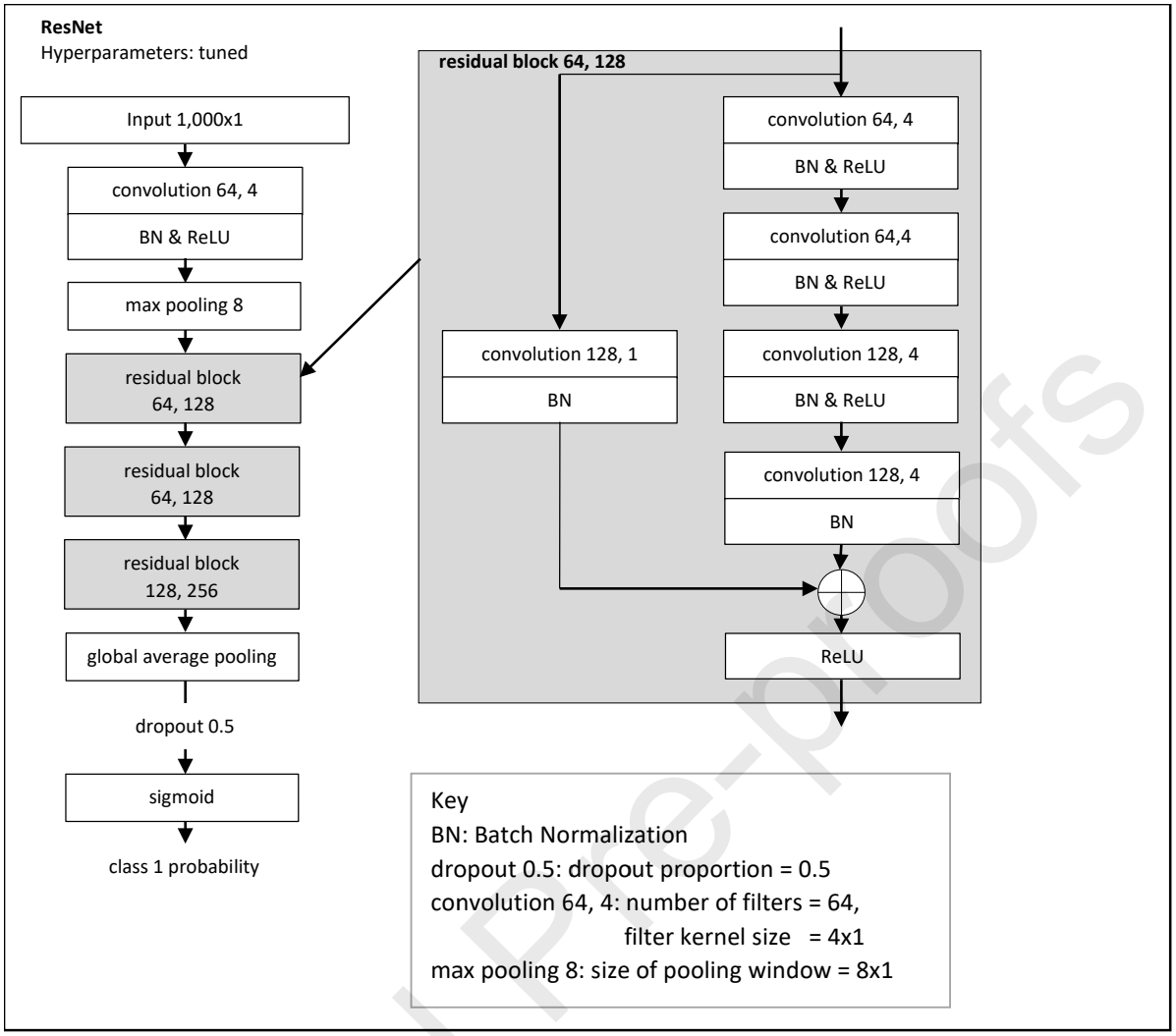

Figure 14 Tuned ResNet and its hyperparameters

5.3. Tuned MLP operating point

The confusion matrix of the selected model in classifying the balanced dataset is shown in Figure 15. It is compared to the dogs' confusion matrix. The dogs' false negative rate is high, 38%, but the MLP's is worse, at 49%. Both MLP and the dogs' true negative rate is high, at 91% and 90% respectively.

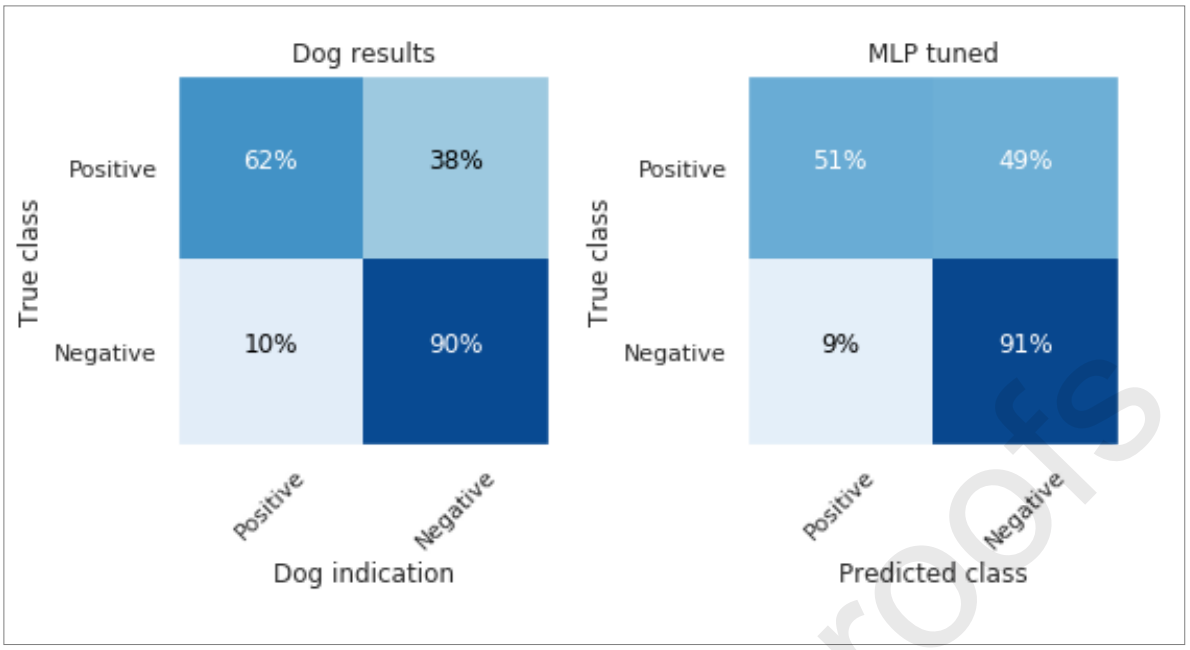

Figure 15 Confusion matrices showing the accuracy of the dogs (left) and the tuned MLP (right) on the balanced dataset (N_D=154)

Figure 16 shows the model's receiver operating characteristic (ROC). This indicates that an operating point exists where the model could achieve 60% TPR (true positive rate or "recall"), if a FPR (false positive rate) of 20% can be tolerated. Certainly, a false positive in a test for cancer is less critical than a false negative so this compromise may be acceptable.

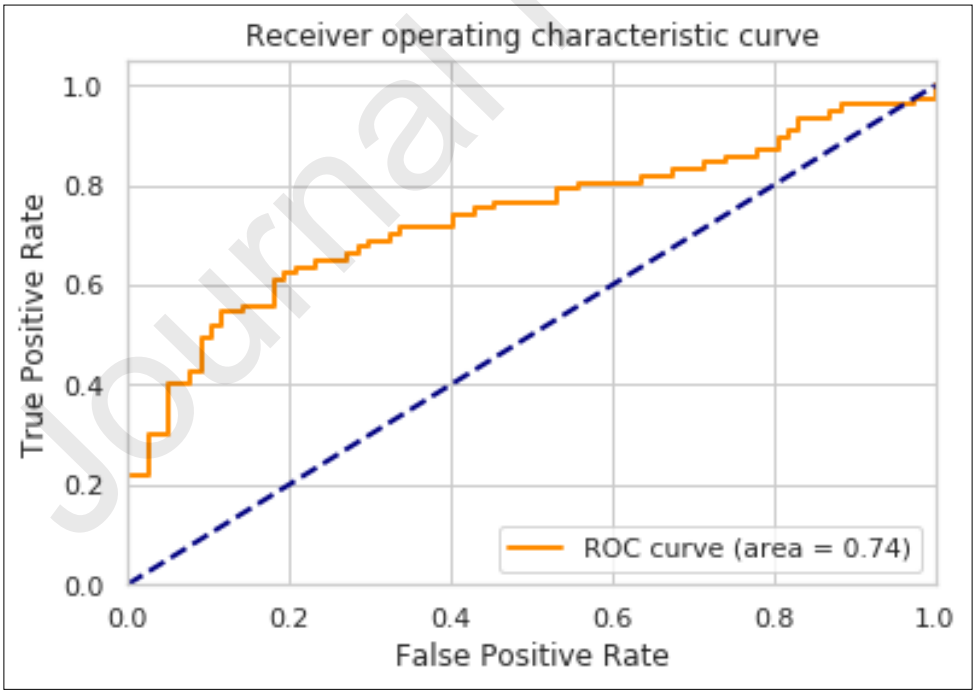

Figure 16 ROC of the tuned MLP. Tested on the balanced dataset (N_D=154)

Figure 17 demonstrates the effect of changing the model's threshold probability from 0.5 to 0.3. That is, if the model calculates that a sample belongs to class 1 (positive) with a

probability of 0.3 or higher, then its prediction is set to class 1. As expected, this results in a model TPR of 61%, similar to that of the dogs. Consequently, the model's TNR (true negative rate) declined to 81%, inferior to the dogs. Overall model accuracy remains unchanged, at 71%.

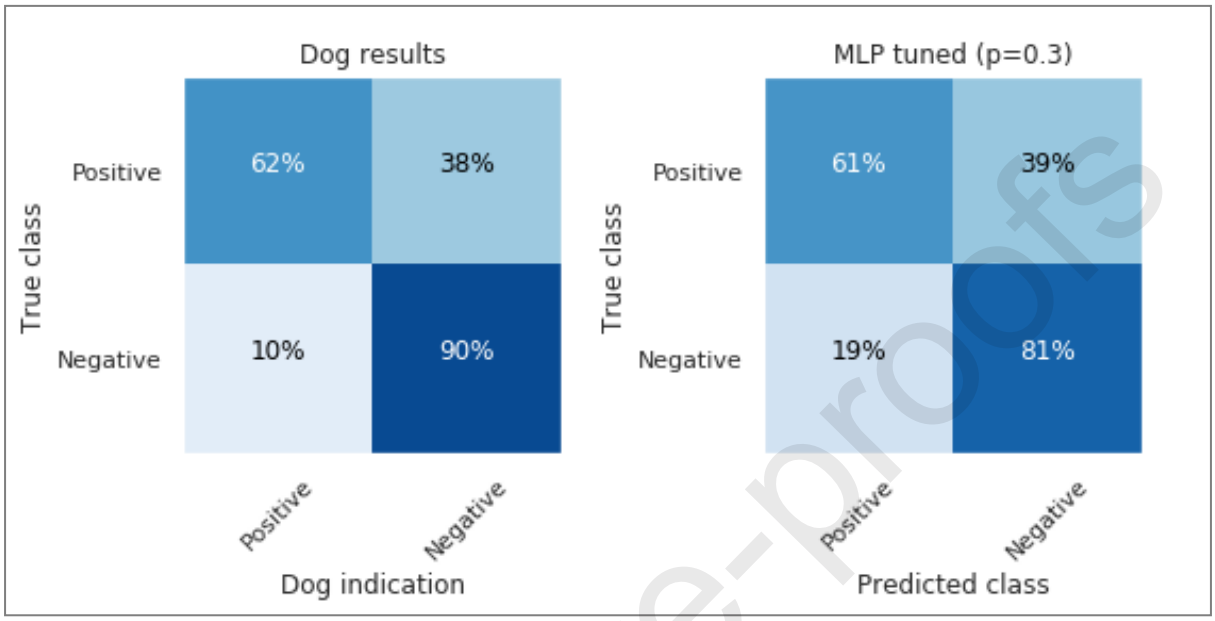

Figure 17 Confusion matrices showing the accuracy of the dogs (left) and the tuned MLP (right) with classification threshold 0.3. Tested on the balanced dataset (N_D=154)

5.4. Final model

To generate a final model, the tuned MLP was trained using all available dog-correct balanced data (N_R = 380 samples). It was then tested on the final test set, which was imbalanced, containing 142 negative samples and 81 positive samples (i.e. 64% are class 0). The results are shown in Figure 18. The model's TPR of 59% matches that of the dogs. As expected from our development results, the model is less performant than the dogs on the negative samples. The model TNR is 79% while the dogs achieved 91%.

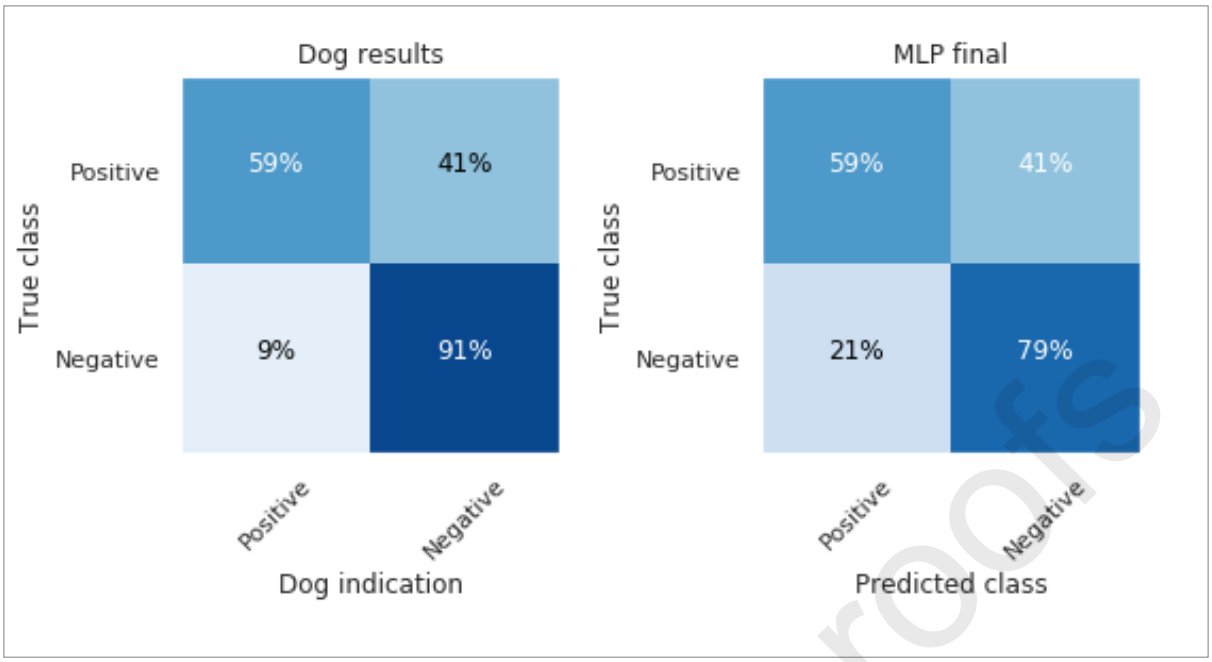

Figure 18 Confusion matrices showing the accuracy of the dogs (left) and the final MLP (right) on the final test set (N_D=223)

Table 3Tables 3 and 4 summarise the results from the final test set.

Table 3 Final test summary results

Class 1 samples (positive)	81
Class 0 samples (negative)	142
Class balance (class 0 / total)	64%
Dogs: number of true positives	48
Dogs: number of true negatives	129
MLP final: number of true positives	48
MLP final: number of true negatives	112
MLP final: Area under the ROC Curve (AUC)	0.70

Table 4 Final test summary metrics

	Dogs	MLP final
accuracy	79%	72%
TPR (recall)	59%	59%
TNR	91%	79%
precision	79%	62%
F1-measure	0.68	0.60

6. Discussion

6.1. Differences between model and dog

The following analysis relates to the final model of Section 5.4: the tuned MLP trained on the dog-correct balanced dataset and tested on the final test dataset.

Table 5 provides further information regarding the confusion matrices of Figure 18. While the model's and dogs' TPR matched, at 59% (correctly identifying 48 of the 81 positive samples), the model and dog classified fourteen of these positive samples differently.

Table 5 Differences in model and dog results on the final test dataset

True class	**Dog result**	**Final model result**	**Number of samples**
Class 1 (positive)	False Negative	True Positive	7
Class 1 (positive)	True Positive	False Negative	7
Class 0 (negative)	False Positive	True Negative	2
Class 0 (negative)	True Negative	False Positive	19

Figure 19 shows the seven positive samples in the final test set where the model was correct and the dog gave a false negative. In the plots we see the active searching that is usually associated with a positive sample. The dog, however, did not perform a sit at the sample, resulting in a false negative, which is more critical than a false positive in a test for a disease. For these seven samples, the model provides valuable information that could potentially be used to identify the dog's result as a false negative.

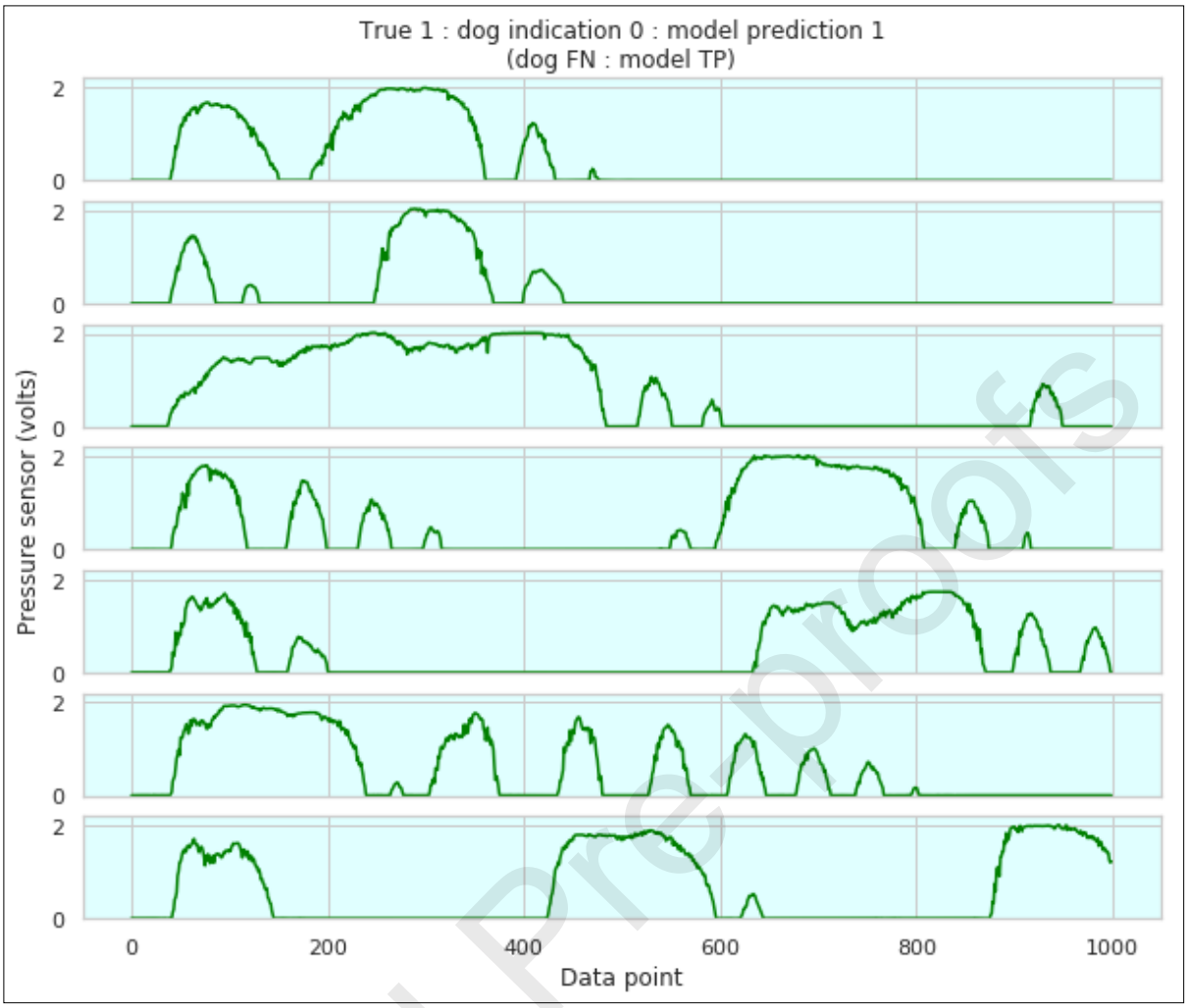

Figure 19 Class 1 data samples where the model was correct and the dog was incorrect

Figure 20 shows the two negative samples where the model was correct and the dog gave a false positive. These pressure samples do not appear to have the active searching associated with a positive sample; however, the dog performed the sit indication and was incorrect. The model, on the other hand, was correct and could potentially provide the dog trainer with additional information about the contents of the sample pot.

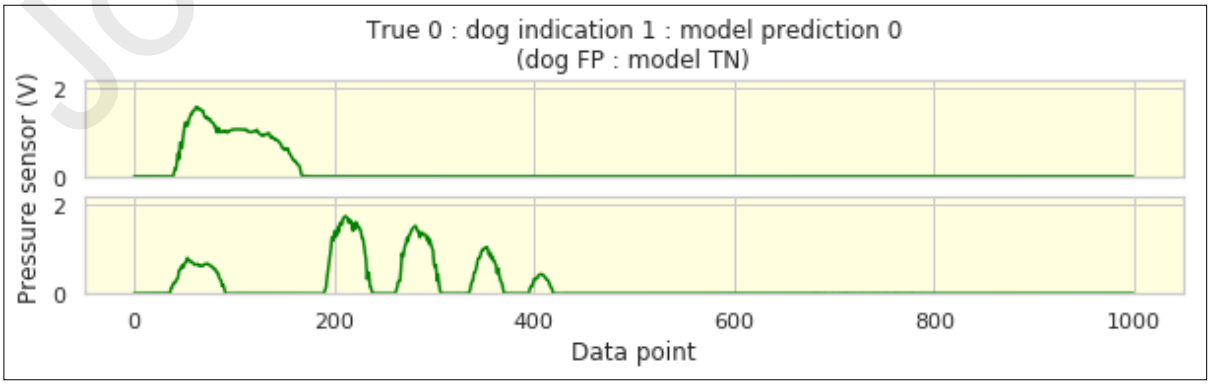

Figure 20 Class 0 data samples where the model was correct and the dog was incorrect

Table 5 shows that there were 35 samples where the model and dog disagreed (16% of all samples). Of these, there were 26 where the dog was correct. Nine were where the model was correct and had the potential to add useful information to the dog's indication. In a non-experimental setting, whether the dog or the model is correct would not be known but their disagreement could be used to flag the sample for further testing, perhaps by obtaining a new sample from the patient or repeating the test with another detection dog.

6.2. Pressure sensor data as an indicator of dogs' stimulus response

For an artificial neural network to successfully classify the scent samples from the pressure sensor data, there needs to be a correlation between the data and the detection dog's stimulus response to a positive sample. When testing the final model on previously unseen data, accuracy was 72% (compared to 64% by chance, Table 3 Table 3), which suggests that there is a correlation.

Dog accuracy on the final test set was 79%. The model can only be as good as the dog's ability to recognise the absence or presence of the target odour. On samples where the dog was incorrect there is a possibility that the dog did not recognise the scent and therefore his search behaviour would not have reflected whether or not the target odour was present. However, when the dog does recognise the scent but does not indicate, a model could be better than the dog at indicating that recognition and therefore the presence of the target odour. Thus, to further evaluate the model, the final model was tested on only the dog-correct samples in the final test set (imbalanced; 73% negative samples). Figure 21 shows that the model's accuracy on these samples was 85%. This suggests that, when the dog is presented with a sample that he recognises, the pressure sensor data that he generates is an indicator of the target scent being present.

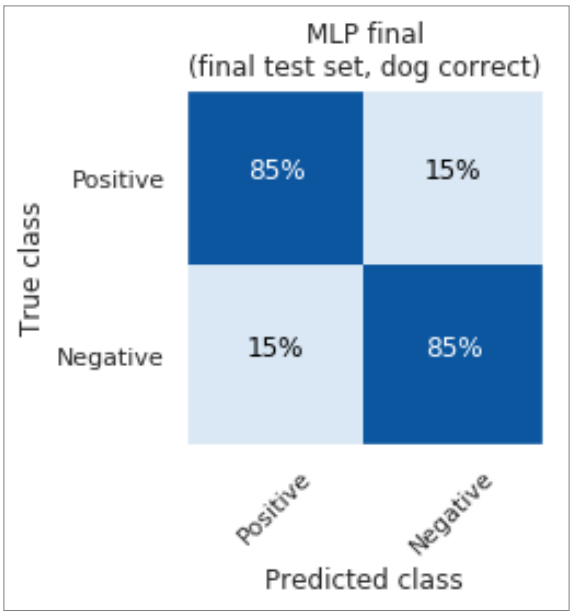

Figure 21 Confusion matrix showing the accuracy of the final model tested on data samples where the dog was correct (N_D=177, 73% negative samples)

7. Conclusions

The most accurate neural network model was found to be MLP, with hyperparameters tuned for optimal performance on the balanced dataset. The tuned MLP has three hidden layers and sixteen nodes per layer plus dropout and L^2 norm regularisation. A novel step was introduced into the neural network training, which was to train the model using samples where the dog was able to correctly identify the sample as positive or negative. This increased the likelihood that there was a correlation between the presence of the target odour and the dog's searching behaviour. This was corroborated by the resulting improvement in model accuracy.

The final model was evaluated using a previously unseen dataset, where the dog was not always correct. The accuracy of the final model was compared, using a confusion matrix, to that of the dogs (Figure 18). The model's true positive rate of 59% reflected the dogs' TPR of 59%. The model's true negative rate, 79%, was lower than the dogs' which was 91%. The model's AUC is 0.70, which is categorised as "acceptable" for a discriminator (Hosmer et al., 2013, Chapter 5). The neural network's accuracy was significantly greater than chance, demonstrating that there is discriminatory information in the pressure sensor data and that the model can discern it.

The neural network outperformed the baseline non-neural-network model, the 1-NN classifier. This study demonstrated that neural networks have potential as an automatic system for determining whether a sample is positive or negative from dog-generated pressure sensor data alone, without requiring the dog to sit. This is the first step towards an

automated system that could supplement a dog's indication gesture by providing additional information in those cases where the trainer reports that the dog's indication is unclear or hesitant. Ultimately, the automated system could supplant the dog's indication gesture, lessening the cognitive load on the dog, reducing dog-training time and enabling them to work faster, screening more samples for disease.

MLP was found to be more accurate than CNN, FCN or ResNet. This was unexpected because MLP disregards temporal information. In contrast, convolutional neural networks (CNN, FCN and ResNet) exhibit translational equivariance. It had been expected that the salient features of the data might occur at different times in each sample and therefore would be discernible to a CNN but not to an MLP. MLP's accuracy may have been aided by the data processing step that located an event window and truncated the time series, effectively aligning all the samples. MLP was presented by Wang et al. (2017b) as merely a benchmark ANN. However, this research suggests that MLP should be considered as a contender for any new time series classification application, particularly if the data samples can be aligned easily.

7.1. Further work

This research has demonstrated that it is possible to classify scent samples using an artificial neural network. To verify the practicality of this approach, the study will need to be repeated using real target samples: for example, urine from people with and without prostate cancer.

For this study, the training samples were selected by identifying those where the dog was correct, which was 74%. It would be preferable to design the data collection procedure such that there is greater confidence that the dog is operating at their peak accuracy. This could include a short training session prior to data collection to focus the dog and to assess their performance on that day.

Further work is needed to determine if the dog's action of performing the sit is recorded by the pressure sensor. For example, if the dog knocks the sample pot, the pressure sensor data for this time period should be excluded. A proximity sensor could be used to ascertain the time at which the dog moves his nose away from the sample and truncate the data at this point.

The results of this research might not hold for other medical detection dogs. Adding data generated by a fourth dog to the model could result in poor accuracy if their search behaviour is very different. Furthermore, some medical detection dogs, particularly smaller breeds, do not touch the plate when searching a sample and the pressure sensor would

record nothing. It might be useful to repeat this study with a different sensor such as a microphone to record the sniffing sound.

A valuable extension to this research would be to use an artificial neural network to classify samples according to the concentration of the target compound, as the concentration of a biological indicator of a disease can be used to guide treatment decisions (Madu & Lu, 2010). A starting point would be to use two concentration strengths and three classes: negative, positive weak and positive strong.

8. Acknowledgements

The authors would like to acknowledge the invaluable contribution of the three detection dogs that took part in this study and thank their trainers at Medical Detection Dogs.

The authors would also like to thank Joe Mills for developing the instrumentation used to record the dogs' stimulus response, and for his part in designing the testing procedure and interpretation of the data.

9. References

Ackermann, N. Introduction to 1D Convolutional Neural Networks in Keras for Time Sequences (2018). https://blog.goodaudience.com/introduction-to-1d-convolutional-neural-networks-in-keras-for-time-sequences-3a7ff801a2cf Accessed 3 May 2019.

Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *31*(3), 606–660. https://doi.org/10.1007/s10618-016-0483-9

Brewster, L. R., Dale, J. J., Guttridge, T. L., Gruber, S. H., Hansell, A. C., Elliott, M., … Gleiss, A. C. (2018). Development and application of a machine learning algorithm for classification of elasmobranch behaviour from accelerometry data. *Marine Biology*, *165*(4), 62. https://doi.org/10.1007/s00227-018-3318-y

Brugarolas, R., Yuschak, S., Adin, D., Roberts, D. L., Sherman, B. L., & Bozkurt, A. (2019). Simultaneous Monitoring of Canine Heart Rate and Respiratory Patterns During Scent Detection Tasks. *IEEE Sensors Journal*, *19*(4), 1454–1462. https://doi.org/10.1109/JSEN.2018.2883066

Chollet, F. (2017). *Deep learning with Python* (1st ed.). Retrieved from https://pmt-eu.hosted.exlibrisgroup.com/permalink/f/gvehrt/TN_sbo_s1_9781617294433

Concha, A., Mills, D. S., Feugier, A., Zulch, H., Guest, C., Harris, R., & Pike, T. W. (2014). Using Sniffing Behavior to Differentiate True Negative from False Negative Responses in Trained Scent-Detection Dogs. *Chemical Senses*, *39*(9), 749–754. https://doi.org/10.1093/chemse/bju045

Craven, B. A., Paterson, E. G., & Settles, G. S. (2010). The fluid dynamics of canine olfaction: Unique nasal airflow patterns as an explanation of macrosmia. *Journal of The Royal Society Interface*, *7*(47), 933–943. https://doi.org/10.1098/rsif.2009.0490

Dau, H. A., Bagnall, A., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., … Keogh, E. (2018). *The UCR Time Series Archive*. Retrieved from http://arxiv.org/abs/1810.07758

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2018). Deep learning for time series classification: A review. *Accepted at Data Mining and Knowledge Discovery*. https://doi.org/10.1007/s10618-019-00619-1

Figo, D., Diniz, P., Ferreira, D., & Cardoso, J. (2010). Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing*, *14*(7), 645–662. https://doi.org/10.1007/s00779-010-0293-9

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Retrieved from http://www.deeplearningbook.org

Guest, C., Pinder, M., Doggett, M., Squires, C., Affara, M., Kandeh, B., Dewhirst, S., Morant, S. V., D'Alessandro, U., Logan, J. G., & Lindsay, S. W. (2019). Trained dogs identify people with malaria parasites by their odour. *The Lancet Infectious Diseases*, *19*(6), 578–580. https://doi.org/10.1016/S1473-3099(19)30220-8

Guralnik, V., & Srivastava, J. (1999). Event detection from time series data. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '99*, 33–42. https://doi.org/10.1145/312129.312190

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. Retrieved from https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

Hosmer, J., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons, Inc. https://pmt-eu.hosted.exlibrisgroup.com/permalink/f/gvehrt/TN_wilbookl10.1002/9781118548387

Japkowicz, N., & Shah, M. (2010). *Evaluating Learning Algorithms: A Classification Perspective*. Retrieved from https://doi-org.libezproxy.open.ac.uk/10.1017/CBO9780511921803.007

Jia, H., Pustovyy, O. M., Waggoner, P., Beyers, R. J., Schumacher, J., Wildey, C., Barrett, J., Morrison, E., Salibi, N., Denney, T. S., Vodyanoy, V. J., & Deshpande, G. (2014). Functional MRI of the Olfactory System in Conscious Dogs. *PLOS ONE*, *9*(1), e86362. https://doi.org/10.1371/journal.pone.0086362

Johnston-Wilder, O., Mancini, C., Aengenheister, B., Mills, J., Harris, R., & Guest, C. (2015). Sensing the shape of canine responses to cancer. *Proceedings of the 12th International Conference on Advances in Computer Entertainment Technology*, 63–66. https://doi.org/10.1145/2832932.2837017

Kingma, D., & Ba, J. (2015, July 23). *Adam: A Method for Stochastic Optimization*. Presented at the 3rd International Conference for Learning Representations, San Diego. Retrieved from http://arxiv.org/abs/1412.6980v8

Ladha, C., Hammerla, N., Hughes, E., Olivier, P., & Ploetz, T. (2013). Dog's Life: Wearable Activity Recognition for Dogs. *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 415–418. https://doi.org/10.1145/2493432.2493519

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

London School of Hygiene & Tropical Medicine. (2020, November). *Using dogs to detect COVID-19*. LSHTM. https://www.lshtm.ac.uk/research/centres-projects-groups/using-dogs-to-detect-covid-19

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440. Retrieved from https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Long_Fully_Convolutional_Networks_2015_CVPR_paper.pdf

Madu, C. O., & Lu, Y. (2010). Novel diagnostic biomarkers for prostate cancer. *Journal of Cancer*, *1*, 150–177.

Mancini, C. (2017). Towards an animal-centred ethics for Animal–Computer Interaction. *International Journal of Human-Computer Studies*, *98*, 221–233. https://doi.org/10.1016/j.ijhcs.2016.04.008

Mancini, C., Harris, R., Aengenheister, B., & Guest, C. (2015). Re-Centering Multispecies Practices: A Canine Interface for Cancer Detection Dogs. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2673–2682. https://doi.org/10.1145/2702123.2702562

Nathan, R., Spiegel, O., Fortmann-Roe, S., Harel, R., Wikelski, M., & Getz, W. M. (2012). Using tri-axial acceleration data to identify behavioral modes of free-ranging animals: General concepts and tools illustrated for griffon vultures. *Journal of Experimental Biology*, *215*(6), 986–996. https://doi.org/10.1242/jeb.058602

Nweke, H. F., Teh, Y. W., Al-Garadi, M. A., & Alo, U. R. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems With Applications*, *105*, 233–261. https://doi.org/10.1016/j.eswa.2018.03.056

Ronao, C. A., & Cho, S.-B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, *59*, 235–244. https://doi.org/10.1016/j.eswa.2016.04.032

Straiton, J. (2020). COVID-19: How has the scientific community risen to the challenge? *BioTechniques*, *68*(5), 232–234. https://doi.org/10.2144/btn-2020-0041

van Rijn, J. N., & Hutter, F. (2017). Hyperparameter Importance Across Datasets. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2367–2376. https://doi.org/10.1145/3219819.3220058

Wang, Z., Yan, W., & Oates, T. (2017a). *Fully Convolutional Neural Networks for state-of-the-art time series classification*. Retrieved from https://github.com/cauchyturing/UCR_Time_Series_Classification_Deep_Learning_Baseline

Wang, Z., Yan, W., & Oates, T. (2017b). Time series classification from scratch with deep neural networks: A strong baseline. *2017 International Joint Conference on Neural Networks (IJCNN)*, 1578–1585. https://doi.org/10.1109/IJCNN.2017.7966039

Willis, C. M., Britton, L. E., Harris, R., Wallace, J., & Guest, C. M. (2011). Volatile organic compounds as biomarkers of bladder cancer: Sensitivity and specificity using trained sniffer dogs. *Cancer Biomarkers*, *8*(3), 145–153. https://doi.org/10.3233/CBM-2011-0208

Withington. Deepscent. (2019). https://github.com/Withington/deepscent Accessed 22 January 2020.

A novel application of machine learning in the field of Animal-Computer Interaction

Pressure sensors record dogs’ sniffing pattern to generate time series data

Neural network time series classifiers applied to detection dogs' scenting signals

Neural network classifier matches dogs’ true positive rate

Machine learning models show promise to augment dog response interpretation accuracy