Support Vector Machine Learning for ECG Classification

Paul Walsh[1,2]

[1] Cork Institute of Technology, Bishopstown, Cork Ireland
[2] NSilico Life Science, Nova UCD, Ireland
paul.walsh@nsilico.com

Abstract. Connected health has huge potential to enhance the diagnosis, monitoring and treatment of a range of conditions. With advances in wearable technology it is now becoming more feasible to monitor and control a range of conditions. This includes heart conditions, which can now be monitored via wearable devices such as the Apple Watch, which is a propriety device that uses machine learning to predict likelihood of arrhythmia and other heart conditions. This paper investigates a Support Vector Machine Learning approach for ECG monitoring and outlines advantages of such an approach. This paper shows that support vector machines can provide useful classification on ECG signals using the Kaggle ECG Heartbeat Categorization Dataset and is potentially a viable machine learning approach to ECG classification.

Keywords: Machine Learning, ECG, Arrhythmia, Connected Health.

1 Introduction

Connected health is a model for the management and delivery of healthcare that uses technology to provide remote services [1]. It aims to optimize the use of re-sources, provide additional opportunities for patients to engage with clinicians and to allow them to participate more in their own care [2]. It leverages relatively low-cost consumer technologies to deliver patient care outside of the clinical setting to provide support in remote care, chronic care, disease and lifestyle management. It is often deployed over existing technical infrastructure such as 4G mobile networks and plans are already evolving to provide connected health services over 5G networks [3].

Consumer devices such as wearables are accelerating the acceptance of connected health solutions [4]. For example, insurance companies are prompting the use of wearable health and fitness trackers and regulatory bodies such as the U.S. Food and Drug Administration are streamlining the approval process for digital health products from smartwatch companies Fitbit, Apple and Samsung [4].

In terms of recent advances in connected health, Apple has hit the headlines in Q4 2018 by launching a new Apple Watch with built in ECG functionality [5]. This is a big step in wearable eHealth devices as it uses a machine learning algorithm to classify data gathered from the watch's sensors. This algorithm can be used to detect Atrial fibrillation (A-fib), which is an irregular heartbeat that is linked with an increased risk of heart failure, dementia, and stroke. A-fib is often symptomless and contributes to approximately 130,000 deaths annually in the United States.

An electrocardiography (ECG) is a record of the electrical activity of the heart usually gathered using electrodes placed on the skin [6]. To capture ECG signals the user must create a closed circuit across their chest. Apple get users to do this by simply placing their finger on the front of the watch, so that an electrode touching the wearers wrist on the back of the watch can read the signal. Where the real innovation comes in is the use of machine learning to classify these signals.

Apple developed this machine learning AI using deep learning technology known as convolutional neural networks (CNNs), which are inspired by models of how the brain works. CNNs are the basis of many AI applications especially in the field of computer vision. Such neural network technology is now widely available to developers on AI platforms from Microsoft Google, Facebook and many more. However, a major disadvantage of ANNs is convergence to local minima rather than finding a global minimum. Support Vector Machines were chosen for this study as they provide a way to circumvent such issues, as SVMs tend towards an optimal margin separation, as the search space constraints define a convex set. Furthermore, ANNs are prone to overfitting, whereas SVMs provide intrinsic margin control meta-parameters, which can be configured to reduce overfitting.

SVMs deliver a unique solution, since the optimality problem is convex. This is an advantage compared to Neural Networks, which have multiple solutions associated with local minima and for this reason may not be robust over different samples.

Moreover, a highly cited paper from Manuel Fernandez-Delgado et al evaluated 179 classifiers from 17 machine learning classes on 121 data sets from the UCI data base [7]. They found that the classifiers most likely to perform the best are the random forest (RF) and SVM with a non-linear kernel. In this paper, we will explore the performance of SVM's on the ECG data from "ECG Heartbeat Categorization Dataset" hosted on Kaggle [8].

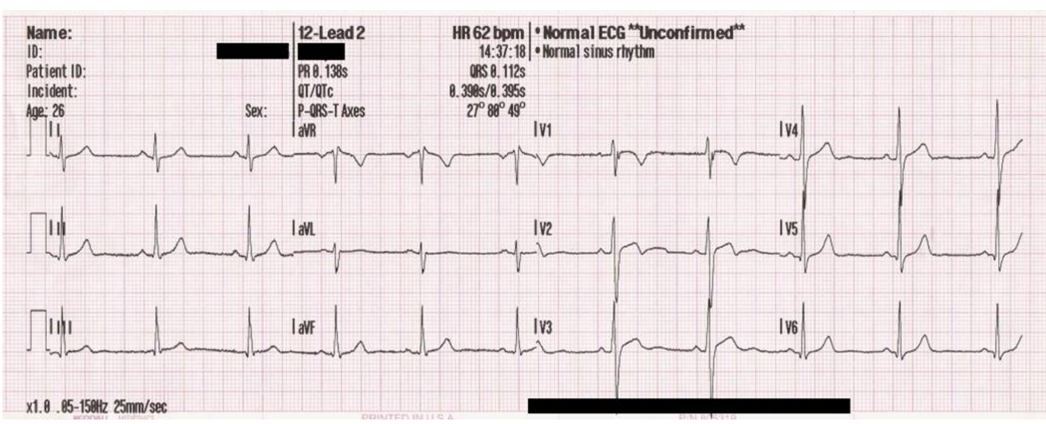

Fig. 1. Example of an ECG electrocardiogram, source: https://bit.ly/2GvS4Ej.

2 Methods

The data used in this study is from the MIT-BIH Arrhythmia Dataset [8] and the signals used in this data set contain a mix of normal heartbeats and heartbeats affected by different forms of arrhythmia. Signals are normally collected and charted in an electro cardiogram, see Figure 1, but in this data set the signals are separated into individual heartbeats.

2.1 ECG Data Sets

The data used in this study is available at https://bit.ly/2XadCLV. This data was used in exploring heartbeat classification using deep neural network architectures [9]. The signals correspond to electrocardiogram (ECG) shapes of heartbeats for the normal case and the cases affected by different arrhythmias and myocardial infarction. These signals are preprocessed and segmented, with each segment corresponding to a heartbeat. The type of heart beat for each sample is stored in the last column of each row, where the beat type is represented by the following integers.

- Normal (N) = 0
- Supraventricular (S) = 1
- Ventricular (V) = 2
- Fusion (F) = 3
- Unclassified (Q) = 4

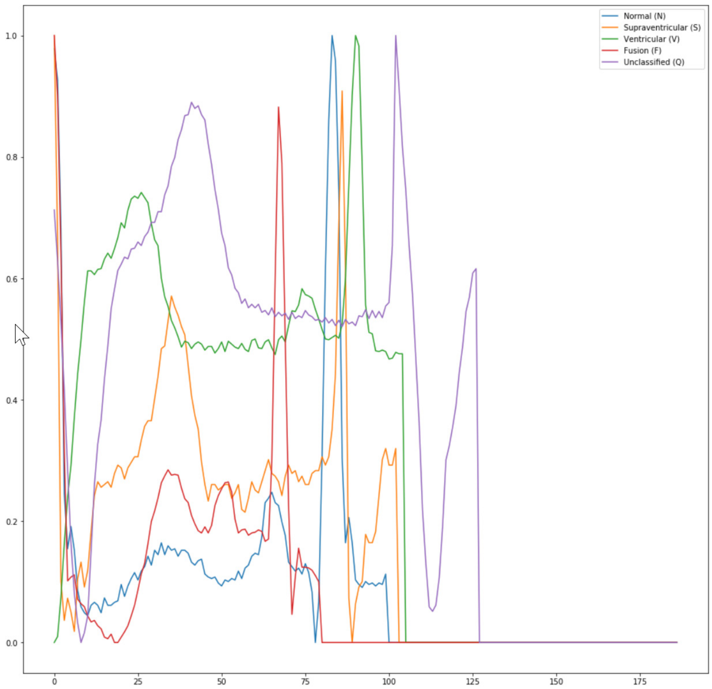

Fig. 2. Samples of the ECG data set.

Abnormal heart beats include supraventricular tachycardia, which is an abnormally fast heart rhythm arising from improper electrical activity in the upper part of the heart. A sample of the various heartbeat types from the data set is shown in Figure 2.

A stacked bar chart is presented in Figure 3, showing that the data set is massively imbalanced towards normal heartbeats. Dataset imbalance is a significant issue in machine learning as over represented exemplars can skew the machine learning model towards classifying input data into these overrepresented classes. This issue is explored in the results section of this research and rectifying approaches are proposed as future work. Figure 3 also shows the split between training data and test data. The test data was held back until the final evaluation phase of the machine learning model built using the training data.

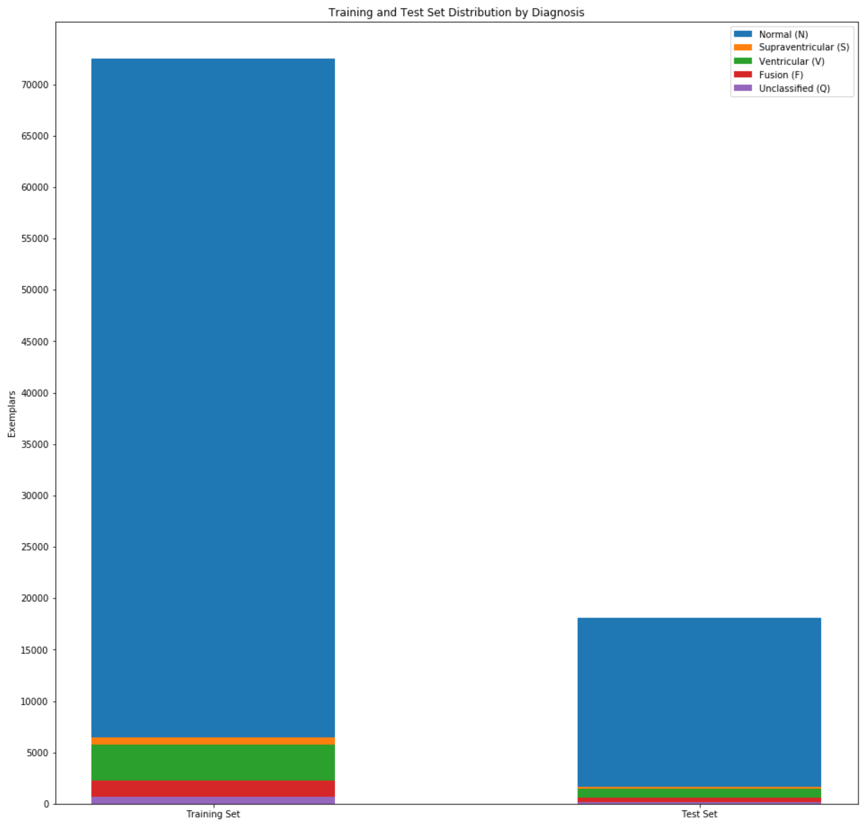

Fig. 3. Stacked bar charts shown data set composition and the split between training data and test data.

2.2 Support Vector Machine

The power of machine learning is in its ability to generalize by correctly classifying unseen data based on models build using training data. Here we use a Support vector machines to build a machine learning model for the ECG dataset, using a portion of the data (80%) for training and the rest for testing the model (20%), reproducing the data split used in the CNN study by Kachuee et al [9].

A Support Vector Machine (SVM) is a supervised learning algorithm that has been shown to have good performance as a classifier [10]. The SVM Algorithm iterates over a set of labeled training samples to find a hyperplane that produces an optimal decision boundary by finding data points, known as support vectors that maximizes the separation between classes.

In order to gauge the performance of the classifier an F1 score is computed, which is a useful measure of the level of precision and recall in a machine learning system

[11]. This can easily be extended to multiclass problems by calculating averages of scores for the classes in question [12]. Precision is the portion of instances among the classified instances that are relevant, while recall or sensitivity is the fraction of correctly classified relevant instances that have been retrieved over the total amount of relevant instances. An algorithm with high precision over a data set will return more relevant results than irrelevant ones. For cardiac diagnosis this is critical as false positive and in particular false negative errors should be avoided. Precision is the ratio of correctly classified true positives t_p, over the sum of true positives t_p and falsely classified positives f_p:

$$Precision = \frac{t_p}{t_p + f_p}$$

An algorithm with high recall will classify most of the relevant data correctly and can be thought of as the ratio of correctly classified true positives t_p, over the sum of true positives t_p and false negatives f_n (the number of instances falsely classified as negative instances):

$$Recall = \frac{t_p}{t_p + f_n}$$

There is usually a trade-off between precision and recall as it is possible to have an algorithm with high precision but low recall and vice versa. For example, the algorithm may be precise by correctly classifying a subset of arrythmia cases, however if it could achieve this by being stringent in its classification and could exclude many other cases, which would give it a low recall.

The balance between precision and recall can be captured using an F1 score which is the harmonic mean of the precision and recall scores, where a score of 1 indicates perfect precision and recall [13].

$$F_1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}}$$

The ECG dataset is partitioned into training and test sets as shown in Figure 3. The SVM machine learning model is trained using the data set and this should be done in such a way that the model does not overfit the data, which occurs when the algorithm fits a decision boundary tightly to the data, including any errors in the data, so that it performs poorly on any unseen input. To avoid overfitting a test data set is held back and is used as the final unbiased measure of the algorithm's performance. A model that produces a high score on the training set but a low score on the test set will have overfit the data, while a model that produces a high score on the training set and a high score on the test set should provide good classifications. A model that underfits, by failing to find any useful decision boundary will perform poorly on both data sets.

SVMs also use a technique known as the kernel trick, which maps data points to a higher dimensional space where a linear separation may be found [14]. The choice of using a kernel is an important machine learning hyperparameter and practitioners needs to consider if the data set is linearly separable or not. Choosing a non-linear kernel for a linear data set will tend to cause the model to over fit the data, which will reduce its ability to generalize as indicated by a poor performance on the test data set F1 score. In this study we establish the best algorithm hyper-parameters by performing a grid search. The hyper-parameters for the support vector machine implemented in this study include a cost function denoted C, which penalizes the algorithm for points that fall within the separating margin. A small value of C, imposes a low penalty for misclassification, thereby allowing a "soft margin", which promotes better generalization at the risk of lower precision. A large value of C imposes a high cost of misclassification, thereby producing a “hard margin", which promotes higher precision but poorer generalization and recall. The challenge here is to find a balance that maximizes the F1 score.

The SVMs can use a linear kernel or non-linear kernels such as Gaussian radial basis function, which allows the SVM algorithm to fit the maximum margin separating hyperplane in a transformed input feature space. The gamma hyper-parameter controls how far the impact a single training has on the model, with low values having a ‘far’ influence and high values having a ‘close’ influence. High values of gamma narrow the region of influence of the kernel for vectors in the feature space, which can cause the SVM to overfit the data. Low values of gamma widen the region of influence, making the algorithm better at generalizing at the expense of losing precision. To find optimal setting for C and gamma a grid search was performed, see Figure 4.

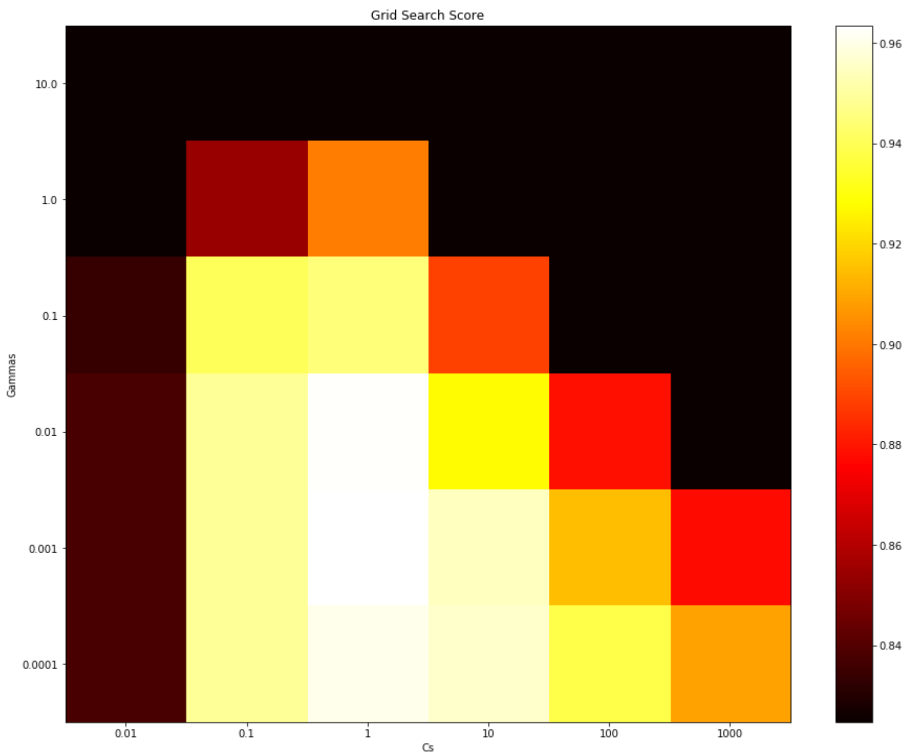

Fig. 4. Grid search scores for the hyper-parameters C and gamma, plotted as a heatmap. Optimal results are in the region C=1, gamma=[0.001:0.01].

3 Results

The `GridSearchCV` method of the `scikit-learn` python machine learning package was used to perform an exhaustive search over the C and gamma support vector machine parameters and both linear and non-linear radial basis function kernels (rbf) were evaluated:

```
tuned_parameters = [{'kernel': ['rbf'], 'gamma': [1, 1e-
1, 1e-2, 1e-3, 1e-4, 1e-5], 'C': [0.1, 1, 10, 100,
1000]}, {'kernel': ['linear'], 'gamma': [1, 1e-1, 1e-2,
1e-3, 1e-4, 1e-5],'C': [0.1, 1, 10, 100, 1000]}]
```

The scores are plotted as a heatmap, showing that optimal results are in the region C=1, gamma=[0.001:0.01], see Figure 4.Using these grid search results it is possible to find good support vector machine configuration settings to produce the results shown in Figure 5.

	precision	recall	f1-score
0	0.97	1.00	0.98
1	0.96	0.58	0.72
2	0.97	0.88	0.92
3	0.77	0.43	0.55
4	1.00	0.89	0.94
micro avg	0.97	0.97	0.97
macro avg	0.93	0.76	0.82
weighted avg	0.97	0.97	0.97

Fig. 5. Precision Recall and average F1 scores on a SVM model built using the full data set, and tested on the unseen test data.

The F1 score is 0.97 for the micro average, which computes global metrics by counting the total true positives, false negatives and false positives. However, this can be misleading for imbalanced data sets, which is the case here. A more pragmatic measure is the macro average, which computes metrics for each label, and finds their unweighted mean, which in this case is 0.82. This metric does not take label imbalance into account and indicates that the model would not perform accurately in its current configuration and it is likely to make classification errors for under-represented instances. Nevertheless, the results are encouraging with a *weighted* average of 0.97. This is calculated by finding the average score weighted by support, which is the number of true instances for each label [12]. This is meaningful as it accounts for label imbalance, as shown in Figure 3.
A confusion matrix for the system evaluated on the full test set is shown in Figure 6, where each row of the matrix represents the instances in the predicted classes, while each column represents the instances in actual classes. While the results are not sufficient accurate classification across all classes, the results are encouraging. These results

are effected by the massive bias towards normal heartbeats in the current data set. Kachuee *et al* [9] have dealt with this issue using data augmentation; by deriving new samples from the existing classes and altering the heartbeat signals amplitude and wavelength and their approach has improved their CNN classification accuracy. Such techniques also work for support vector machines [15] and will be applied to this work in future research.

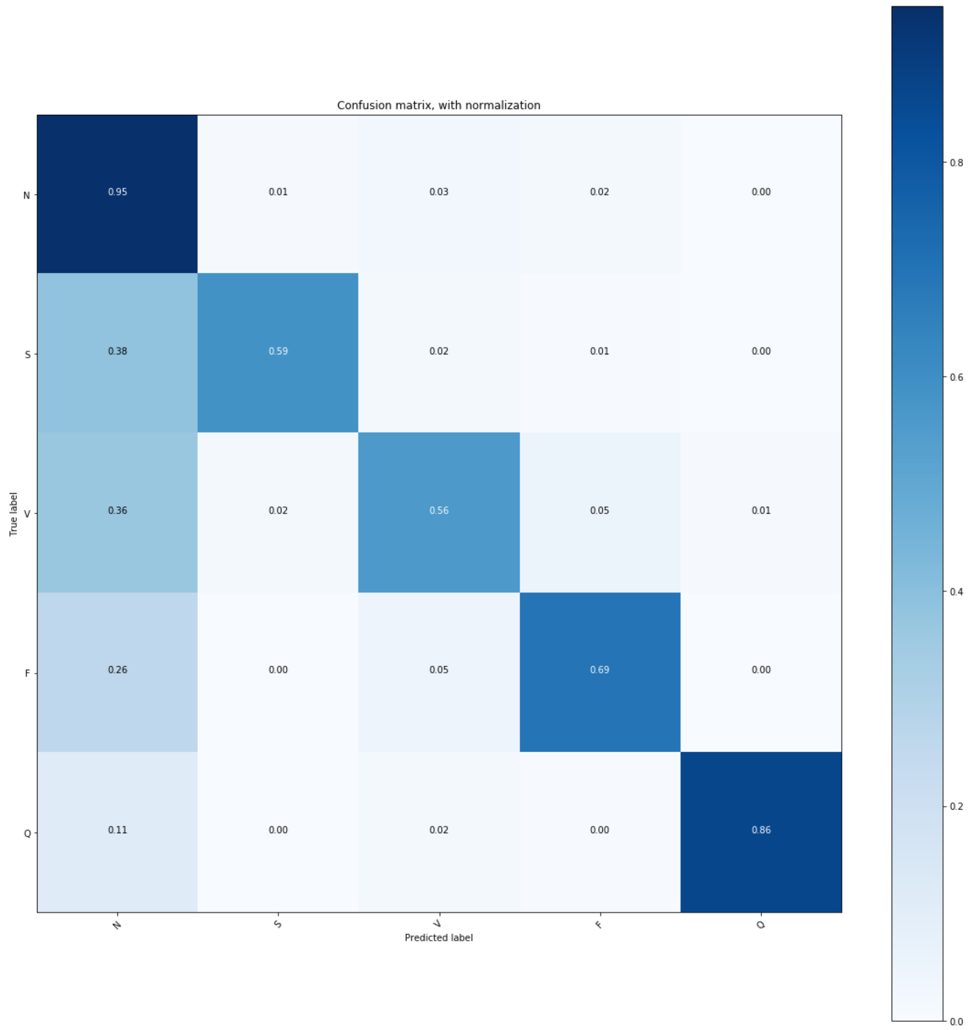

Fig. 6. Confusion matrix of SVM on full ECG data set.

4 Summary

A support vector machine was built to perform analysis on electrocardiogram signals and a grid search was performed to find SVM hyperparameters that balance precision and recall. The resulting SVM produced a weighted average F1 score of 0.97, although the macro-average F1 score was 0.82, due to imbalance in the data set. This compares well with the deep learning approaches such as those used by Kachuee et al [9], where data augmentation resulted in a F1 score of 0.95. These results indicate that support vector machines can provide useful classification on ECG signals with the added benefit of providing a basis for converging to a global minimum and can be configured to avoid over fitting. This SVM approach aligns with results reported by Manuel Fernandez-Delgado et al [7], who evaluated 179 classifiers on 121 data sets from the UCI database. They found that one of the classifiers most likely to perform the best is the SVM with a non-linear kernel and the results presented here provide a basis for similar findings.

Future work will expand on these findings by evaluating data augmentation techniques informed by a time series analysis of the various heartbeat types. A comparison with other machine learning techniques will also be performed including evaluation of random forest, convolutional neural networks and other approaches.

5 Acknowledgements

Paul Walsh is supported by funding through Science Foundation Ireland (SFI) Multi-Gene Assay Cloud Computing Platform - (16/IFA/4342.)

References

[1] N. Carroll, "Key Success Factors for Smart and Connected Health Software Solu-tions," *Computer,* vol. 49, no. 11, pp. 22-28, 2016.

[2] N. Carroll, C. Kennedy and I. Richardson, "Challenges towards a Connected Community Healthcare Ecosystem (CCHE) for managing long-term conditions," *Gerontechnology,* vol. 14, no. 2, pp. 64-77, 2016.

[3] M. a. W. P. Healy, "Detecting demeanor for healthcare with ma-chine learning.," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Kansas, 2017.

[4] C. Pettey, "Wearables Hold the Key to Connected Health Monitoring," 8 3 2018. [Online]. Available: https://www.gartner.com/smarterwithgartner/wearables-hold-the-key-to-connected-health-monitoring/. [Accessed 2 2019].

[5] S. Sanyal, "5 Reasons Why You Should Be Excited About Apple Watch 4's ECG Sensor," 28 10 2018. [Online]. Available: https://www.forbes.com/sites/shourjyasanyal/2018/10/28/5-reasons-why-you-should-be-excited-about-apple-watch-4-ecg-sensor/. [Accessed 2 2019].

[6] H. A. American, "Electrocardiogram (ECG or EKG)," 31 7 2015. [Online]. Available: http://www.heart.org/en/health-topics/heart-attack/diagnosing-a-heart-attack/electrocardiogram-ecg-or-ekg. [Accessed 2 2019].

[7] M. Fern´andez-Delgado, E. Cernadas and S. Barro, "Do we need hundreds of classifiers to solve real world classification problems?," *Journal of Machine Learning Research,* vol. 15, no. 1, pp. 3133-3181, 2014.

[8] S. Fazeli, "ECG Heartbeat Categorization Dataset:Segmented and Preprocessed ECG Signals for Heartbeat Classification," Kaggle, 6 2018.

[Online]. Available: https://www.kaggle.com/shayanfazeli/heartbeat . [Accessed 2 2019].

[9] M. Kachuee, S. Fazeli and M. Sarrafzadeh, "ECG Heartbeat Classification: A Deep Transferable Representation," in *IEEE International Conference on Healthcare Informatics (ICHI)*, New York, 2018.

[10] B. Scholkopf and A. J. Smola, Learning with kernels: support vector machines, regularization, optimization, and beyond, MIT Press Cambridge, 2001.

[11] D. M. Powers, "Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation," *Journal of Machine Learning Technologies,* vol. 2, no. 1, pp. 37-63, 2011.

[12] scikit-learn, "sklearn.metrics.f1_score," scikit learn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html. [Accessed 2 2019].

[13] Y. Sasaki, "The truth of the F-measure.," *Teach Tutor mater 1,* vol. 1, no. 5, pp. 1-5, 2007.

[14] B. Schölkopf, C. Burges and V. Vapnik, "Incorporating invariances in support vector learning machines," in *International Conference on Artificial Neural Networks*, 1996.

[15] N. G. Polson and t. L. Scott, "Data augmentation for support vector machines," *Bayesian Analysis,* vol. 6, no. 1, pp. 1-23, 2011.