Machine learning and the electrocardiogram over two decades: Time series and meta-analysis of the algorithms, evaluation metrics and applications

Machine Learning and the Electrocardiogram over Two Decades: Time Series and Meta-Analysis of the Algorithms, Evaluation Metrics and Applications

Khaled Rjoob, MSc[a], Raymond Bond, PhD[a*], Dewar Finlay, PhD[a], Victoria McGilligan, PhD[b], Stephen J Leslie, FRCP, PhD[c], Ali Rababah, MSc[a], Aleeha Iftikhar, MSc[a], Daniel Guldenring, PhD[a], Charles Knoery, MBChB[c], Anne McShane, MSc[e], Aaron Peace, MB BCh BAO, PhD[f], Peter W. Macfarlane DSc, FRSE[g].

[a] Faculty of Computing, Engineering & Built Environment, Ulster University, Northern Ireland, UK.
[b] Faculty of Life & Health Sciences, Centre for Personalised Medicine, Ulster University, Northern Ireland, UK.
[C] Department of Diabetes & Cardiovascular Science, University of the Highlands and Islands, Centre for Health Science, Inverness, UK.
[d] HTW Berlin, Wilhelminenhofstr. 75A, 12459 Berlin, Germany.
[e] Emergency Department, Letterkenny University Hospital, Donegal, Ireland.
[f] Western Health and Social Care Trust, C-TRIC, Ulster University, Northern Ireland, UK.
[g] Institute of Health and Wellbeing, University of Glasgow, UK.

Abstract:
Background: The application of artificial intelligence to interpret the electrocardiogram (ECG) has predominantly included the use of knowledge engineered rule-based algorithms which have become widely used today in clinical practice. However, over recent decades, there has been a steady increase in the number of research studies that are using machine learning (ML) to read or interrogate ECG data.
Objective: The aim of this study is to review the use of ML with ECG data using a time series approach.
Methods: Papers that address the subject of ML and the ECG were identified by systematically searching databases that archive papers from January 1995 to October 2019. Time series analysis was used to study the changing popularity of the different types of ML algorithms that have been used with ECG data over the past two decades. Finally, a meta-analysis of how various ML techniques performed for various diagnostic classifications was also undertaken.
Results: A total of 757 papers was identified. Based on results, the use of ML with ECG data started to increase sharply ($p<0.001$) from 2012. Healthcare applications, especially in heart abnormality classification, were the most common application of ML when using ECG data ($p<0.001$). However, many new emerging applications include using ML and the ECG for biometrics and driver drowsiness. The support vector machine was the technique of choice for a decade. However, since 2018, deep learning has been trending upwards and is likely to be the leading technique in the coming few years. Despite the accuracy paradox, accuracy was the most frequently used metric in the studies reviewed, followed by sensitivity, specificity, F1 score and then AUC.
Conclusion: Applying ML using ECG data has shown promise. Data scientists and physicians should collaborate to ensure that clinical knowledge is being applied appropriately and is informing the design of ML algorithms. Data scientists also need to consider knowledge guided feature engineering and the explicability of the ML algorithm as well as being transparent in the algorithm's performance to appropriately calibrate human-AI trust. Future work is required to enhance ML performance in ECG classification.

Index terms: Machine learning, deep learning, electrocardiogram, artificial intelligence.

1. Introduction

Electrocardiogram (ECG) signals represent the electrical activity of the heart muscle as sensed by electrodes that are placed on the skin. The ECG plays an important role in identifying normal and abnormal heart rhythms, acute coronary syndrome as well as other cardiac and non-cardiac abnormalities (e.g. pericarditis, channelopathies). The ECG has also been used by other researchers to study sleep, emotions and stress. The ECG is one of the most widely used tools in medicine for patient monitoring and to assist in patient diagnosis. The ECG is commonly used because it is cost-effective, non-invasive and is an efficient diagnostic tool [1]. It is also the gold standard for identifying arrhythmias. Key components of the ECG signal are routinely analysed which include distinct morphological features such as the P wave (corresponding to excitation of the atria), QRS complex (corresponding to excitation of the ventricles) and the T wave (corresponding to repolarisation of the ventricular cardiomyocytes). Assessment of these components is used to interpret and discriminate between different ECG signals to help in patient diagnosis [2-3]. However, in some cases these morphological features are difficult for a human to interpret [4]. Furthermore, a number of studies have shown that physicians are often poor at reading ECGs in clinical practice [2][4]. Hence, machine learning (ML) has been used by a number of researchers to investigate if artificial intelligence (AI) can improve ECG interpretation and clinical decision-making. A basic rationale for using ML for ECG interpretation is that ML can potentially consider subtleties in the ECG signals that are beyond the ECG features that are routinely considered by humans. AI can be defined as the ability of machines to undertake tasks that are normally carried out by humans. ML is a branch of AI and is divided into three main categories, including: 1) supervised ML (for building algorithms to automatically interpret ECGs by using a dataset that include labels such as a gold standard disease classification), 2) semi-supervised ML (similar to supervised ML except that some cases in the dataset have labels and other cases do not - using such a partially labelled dataset can be used to refine the decision boundary in the algorithm), and 3) unsupervised ML (when the dataset is unlabelled and techniques such as clustering and association rule mining are used to discover labels or new knowledge and associations). Deep learning (DL) has become a popular ML technique and is most commonly used in supervised ML. DL is somewhat distinct from traditional supervised ML techniques, in that 1) DL algorithms can outperform other techniques when using big datasets [5], and 2) DL does not require handcrafted feature engineering which is an approach whereby a data scientist selects a set of variables that have predictive power, whereas DL can do this filtering automatically. In medicine, ML algorithms can be also used to augment or assist clinicians to improve their decision making, which can reduce unnecessary costs or delays - thus saving time and improving patient outcomes [6]. The aim of this work is to analyse and summarise as many as possible of the published studies that used ML algorithms and ECG data for the exemplary application areas such as ECG classification [7-8], ECG signal quality analysis [9-10], ECG lead misplacement detection [11-14], emotion detection [15-16], activity classification [17-18], heart disease diagnostics [19], driver drowsiness detection [20], false alarm reduction [21-22], and biometric authentication systems [23-24]. This paper addresses the following research questions:

1. What is the general growth in ML and ECG research according to the number of pertinent papers published?
2. What are the trends and frequencies in the use of different ML techniques?
3. What are the trends and frequencies in the use of various metrics that have been used by researchers to evaluate ML algorithms?
4. What are the most common applications of ML when using ECG data?
5. How has ML performed when being used to classify ECGs in order to assist in patient diagnoses?

2. Methods

Online searches were performed using SCOPUS, PubMed and IEEE databases to seek relevant items published between 1995 and October 2019. The search terms: "ECG", "Machine Learning", "Artificial Intelligence", "Deep Learning" were combined as keywords in different sequences and combinations in order to achieve maximal search sensitivity. Inclusion criteria were used with the following conditions: 1) original studies related to machine learning involving ECG data that are written in the English, 2) a clearly defined ECG dataset and a clear specification of ML techniques. Studies were excluded if they did not use ML with ECG data. The type of ML technique and the results such as accuracy, sensitivity and specificity were extracted from each article. The five tribes taxonomy developed by Domingos [25] was applied for further analysis to categorise ML classifiers into five different types (tribes). These include 1) symbolists (e.g. decision trees), 2) connectionists (e.g. artificial neural networks), 3) evolutionaries (e.g. genetic algorithms), 4) Bayesian (e.g. Bayesian networks) and 5) analogizers (e.g. k-nearest neighbour).

2.1 Data analysis

For data analysis, we used R programming (R Studio version 3.5.1) and R libraries such as 'forestplot', 'ggplot' and 'metafor' for meta-analysis. Chen and Liu's method [26] was used for time series analysis to detect change, trend and frequency in time series data. Statistical tests were applied using the Chi-square test where alpha=0.05.

3. Results

The database searches identified a total of 2116 articles. A total of 761duplicate articles were removed and the remaining articles (n=1355) were subject to screening based on title and abstract. After the screening, 598 articles were removed, and 757 articles were included as shown in figure 1.

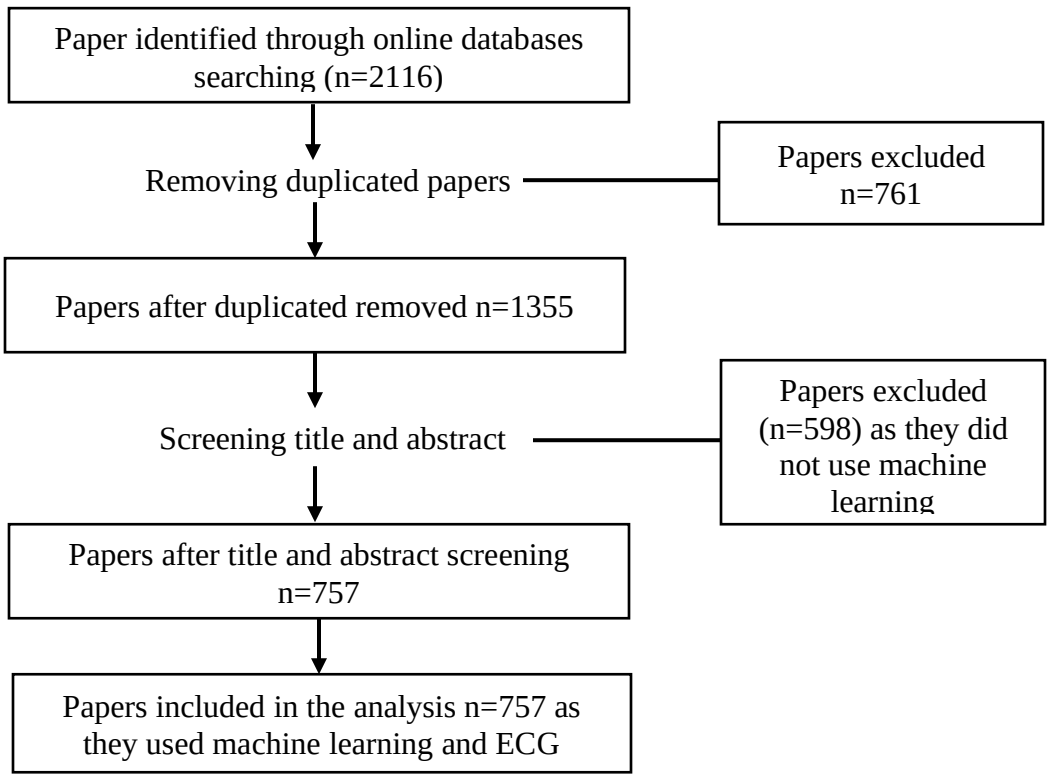

Figure 1: Literature search strategy and selection.

All articles which passed primary screening based on the title and abstract were considered as shown in figure 1. Figure 2 clearly shows the significant increase in the amount of research that uses ML with ECG data.

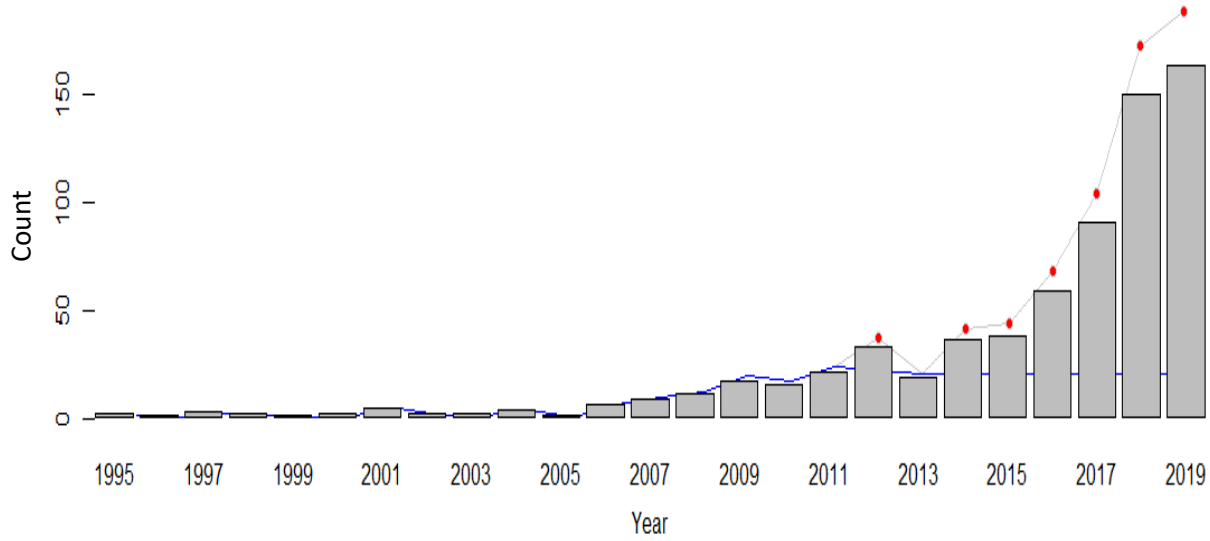

Figure 2: Number of papers per year that address the use of ML with ECGs as identified in IEEE, PubMed and SCOPUS databases. The blue line represents the predicted number of published papers per year, while the grey line shows the actual number of published papers per year. Red dots represent the years where the number of published papers is different from the number of predicted published papers and the predicted number significantly increased. As this search ended at the beginning of October 2019, the last three months in 2019 were extrapolated by taking the average number of published papers per month in the previous nine months in the same year.

According to Chen and Liu's analysis, the years 2012, 2014, 2016, 2017, 2018 and 2019 are the years where the number of ML papers have a significant step change. After 2012, the number of publications that used ML with ECG data increased significantly, i.e. there were 640 after 2012 out of 757 from 1995 until Oct 2019 ($p<0.001$).

3.1 Machine learning techniques

A large number of different ML techniques were used with ECG data (number of techniques=65). Figure 3 shows the time series of these frequencies over the past two decades. This shows that support vector machines (SVM) have been the most dominant technique. All 65 ML techniques and their frequency of use can be seen in Table 1.

When collapsing the ML techniques down into just five categories using the five tribes taxonomy, the connectionist algorithms were trending upwards in the years 1995, 1996, 2006 and then re-emerged as a popular trend in 2019. Analogizers which include KNN and SVM techniques set the trend in the years 2000 to 2004 and again from 2007 to 2018. Bayesian techniques were trending upwards in 1997 and 1998 as shown in Figure 4.

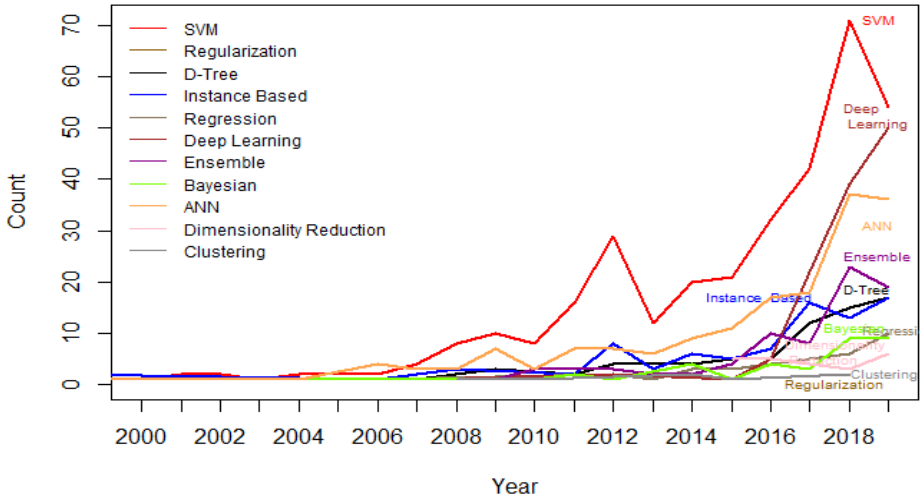

Figure 3: Time series of the frequency of each ML technique over the last two decades. The support vector machine (SVM) has been the most dominant algorithm. Year 2005 was not included because either it did not have published works that used ML or full articles were not available. In 2019, the last three months were extrapolated by taking the average number of published papers per month in the previous nine months in the same year.

Table 1: ML techniques and their frequency of use from 1995 to 2019.

ML	#	%	ML	#	%
SVM: Support Vector Machine	332	43.9%	**ENL: Elastic Net Logistic**	1	0.1%
ANN: Artificial Neural Network	125	16.5%	**SM: Statistical Model**	1	0.1%
KNN: K-Nearest Neighbour	73	9.6%	**KLR: Kernel Logistic Regression**	1	0.1%
DT: Decision Tree	69	9.1%	**SPDR: Sample Percentage in the Dynamic Range**	1	0.1%
RF: Random Forest	63	8.3%	**ZCR: Zero-Crossing Rate**	1	0.1%
CNN: Convolutional Neural Network	62	8.2%	**SDSM: Smart Decision Support Module**	1	0.1%
ELM: Extreme Learning Machine	46	6.1%	**TDEBOOST**	1	0.1%
NB: Naive Bayes	31	4.1%	**SL: Supper Learner**	1	0.1%
DL: Deep Learning	29	3.8%	**TREEBOOST**	1	0.1%
LOG: Logistic Regression	23	3.0%	**TASOM: Time-Adaptive Self-Organizing Map**	1	0.1%
LDA: Linear Discriminant Analysis	22	2.9%	**BICO: Online Clustering Algorithm**	1	0.1%
LSTM: Long Short-Term Memory	16	2.1%	**RB: Rule-Based**	1	0.1%
HMM: Hidden Markov Model	10	1.3%	**ESS: Ensemble Based Score System**	1	0.1%
DBN: Deep Belief Network	8	1.1%	**AMGLVQ: Adaptive Multilayer Generalized Learning Vector Quantization**	1	0.1%
AdaBOOST: Adaptive Boosting	8	1.1%	**CFM: C-F model**	1	0.1%
GA: Genetic Algorithm	7	0.9%	**VF15: Voting Feature Intervals**	1	0.1%
LR: Linear Regression	7	0.9%	**DTW: Dynamic Time Warping**	1	0.1%
K-means	6	0.8%	**SKF: Switching Kalman Filter**	1	0.1%
BPN: Back Propagation Network	5	0.7%	**D-Logic: Decision Logic**	1	0.1%
GMM: Gaussian Mixture Model	5	0.7%	**ADMM: Alternating Direction Method of Multipliers**	1	0.1%
DL-SVD: Dictionary Learning Algorithm Based on Singular Value Decomposition	4	0.5%	**HDC-MER: HD Computing-based Multimodality Emotion Recognition**	1	0.1%
SVR: Support Vector Regression	3	0.4%	**ZC: Zero Crossing**	1	0.1%
ESN: Echo State Networks	3	0.4%	**LTMIL: Latent Topic Multiple Instance Learning**	1	0.1%
ANFIS: Adaptive Neuro-Fuzzy Inference System	3	0.4%	**AIRS: Artificial Immune Recognition System**	1	0.1%
SOM: Self-Organizing Map	2	0.3%	**XGBOOST**	1	0.1%
GBM: Gradient Boosting Machines	2	0.3%	**FIA: Fuzzy Immune Approach**	1	0.1%
MLC: Maximum-Likelihood Classifier	2	0.3%	**FCM: Fuzzy C-Means**	1	0.1%
CRF: Conditional Random Fields	2	0.3%	**SRC: Sparse Representation Classifier**	1	0.1%
LVQ: Learning Vector Quantization	2	0.3%	**BOOSTSTRAP**	1	0.1%
J48	2	0.3%	**DFA: Discriminant Function Analysis**	1	0.1%
SMO: Sequential Minimal Optimization	2	0.3%	**NCA: Neighbourhood Components Analysis**	1	0.1%
BEAT: Beat-to-Beat Estimation by Adaptive Training	2	0.3%	**RVM: Relevance Vector Machine**	1	0.1%
EMD: Empirical Mode Decomposition	2	0.3%			

%= #/757 (where 757 is the total number of papers)

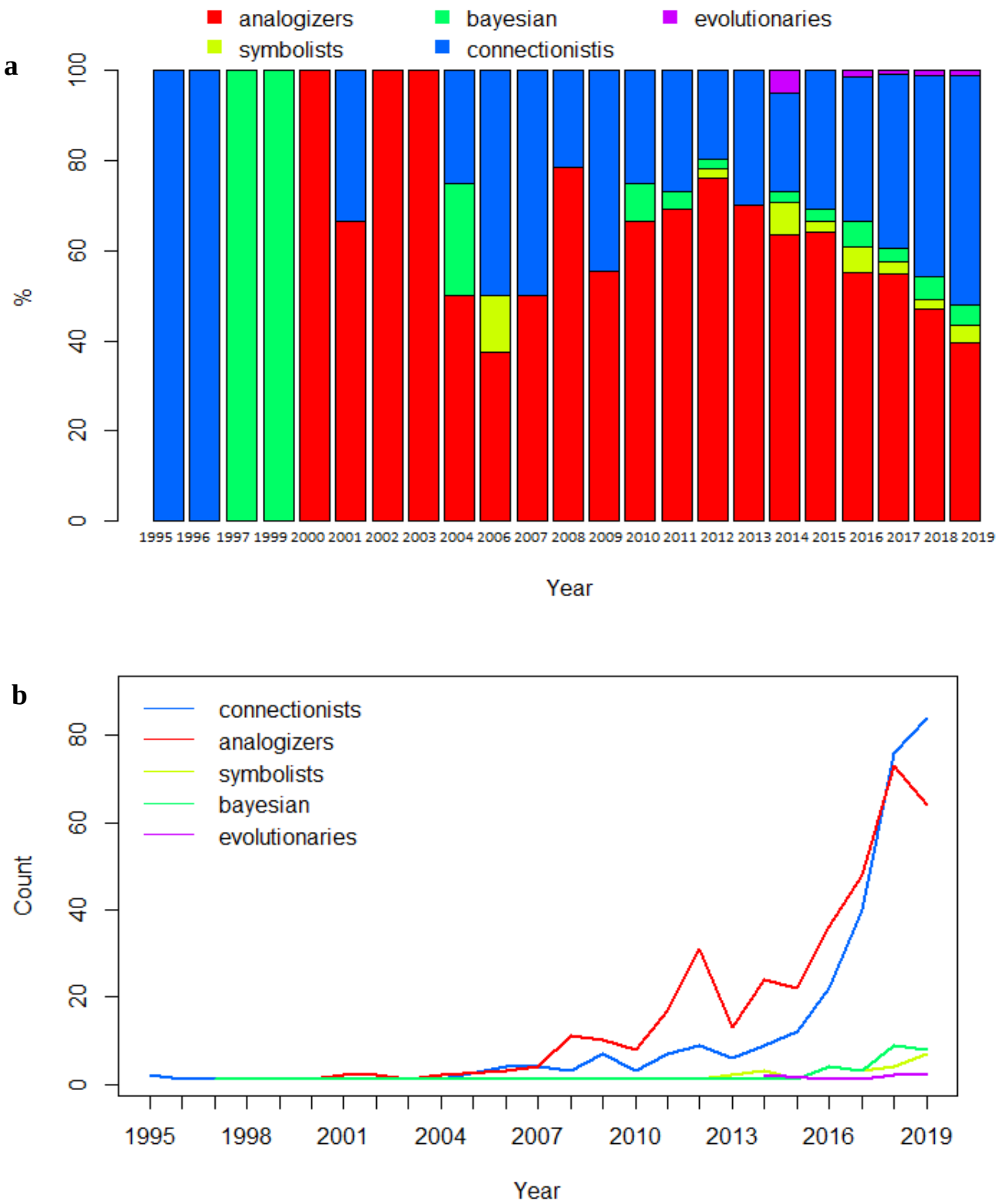

Figure 4: Five tribes analysis, **a:** represents the percentage of studies using algorithms in each tribe in each year and **b:** represents the total number of times an algorithm from each tribe has been used. The last three months in 20119 were extrapolated by taking the average number of published papers per month in the previous nine months in the same year.

3.2 Machine learning evaluation metrics

Figure 5 shows the time series of metrics that have been used over the past number of decades to evaluate ML models. Accuracy was the most used metric (49.2% of studies) followed by sensitivity (20.3% of studies) and specificity (17.1% of studies) from 1995 to 2019. The use of sensitivity and specificity might be expected to be identical, but this was not the case. Different metrics were used to complement sensitivity such as positive detection rate and precision instead of specificity. The significant increase in the use of accuracy is surprising given the obvious problem of the 'accuracy paradox' [27] and the no-information rate (NIR), which represents the largest proportion of the observed classes especially in the case of an unbalanced dataset [27]. A total of 99 articles (13.01%) out of 757 articles used only accuracy in their work ($p<0.001$). Of these 99, all author backgrounds were from computing and engineering (computing=56/99, electrical and electronic engineering=32/99 and biomedical engineering=11/99). The significant increase of using accuracy only started from 2008 (95/99, $p<0.001$). The significant use of the accuracy metric occurred because it was used to compare the relative accuracy achieved between different ML classifiers that were trained and tested using the same data. Since 2016 there has been an increase in the use of AUC, precision (or positive predictive value (PPV)), sensitivity, specificity and F1 scores (F1 is the harmonic mean between sensitivity and specificity).

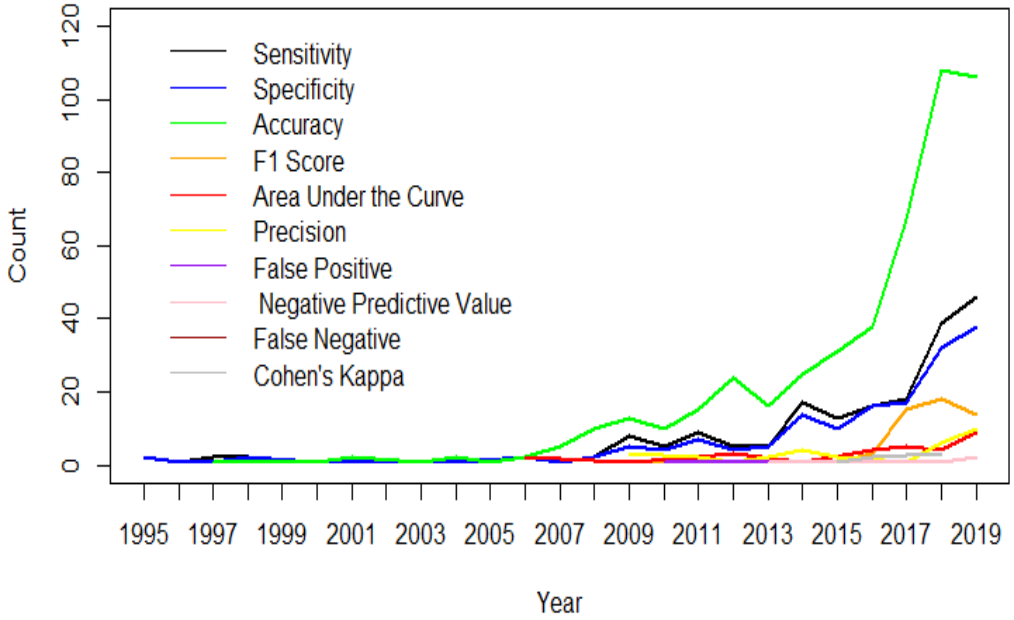

Figure 5: Time series of the number of studies that use each of the respective ML evaluation metrics. Regression metrics such as root mean square error (RMSE) were excluded, because few papers ($n<4$) used them. In 2019, the last three months were extrapolated by taking the average number of published papers per month in the previous nine months in the same year.

3.3 ECG applications

The studies analysed in this paper used and evaluated ML algorithms for different purposes or applications as shown in Figure 6. The majority of these studies (n=400/757) focused on classifying different cardiac abnormalities as shown in Figure 6a and Figure 6b. Cardiac abnormality classification algorithms have been categorised into two different groups: 1) arrhythmias (AR) (n=202/400) and 2) non-arrhythmias (Non AR) (n=62/400) ($p<0.001$) as shown in Figure 6b. A large proportion of the AR group (57.53%) focused on detecting atrial fibrillation followed by premature ventricular contraction (17.8%), ventricular fibrillation (9.5%), bradycardia (5.5%), tachycardia (4.1%), ventricular tachycardia (4.1%) and supraventricular tachycardia (1.4%) as shown in Figure 6c.

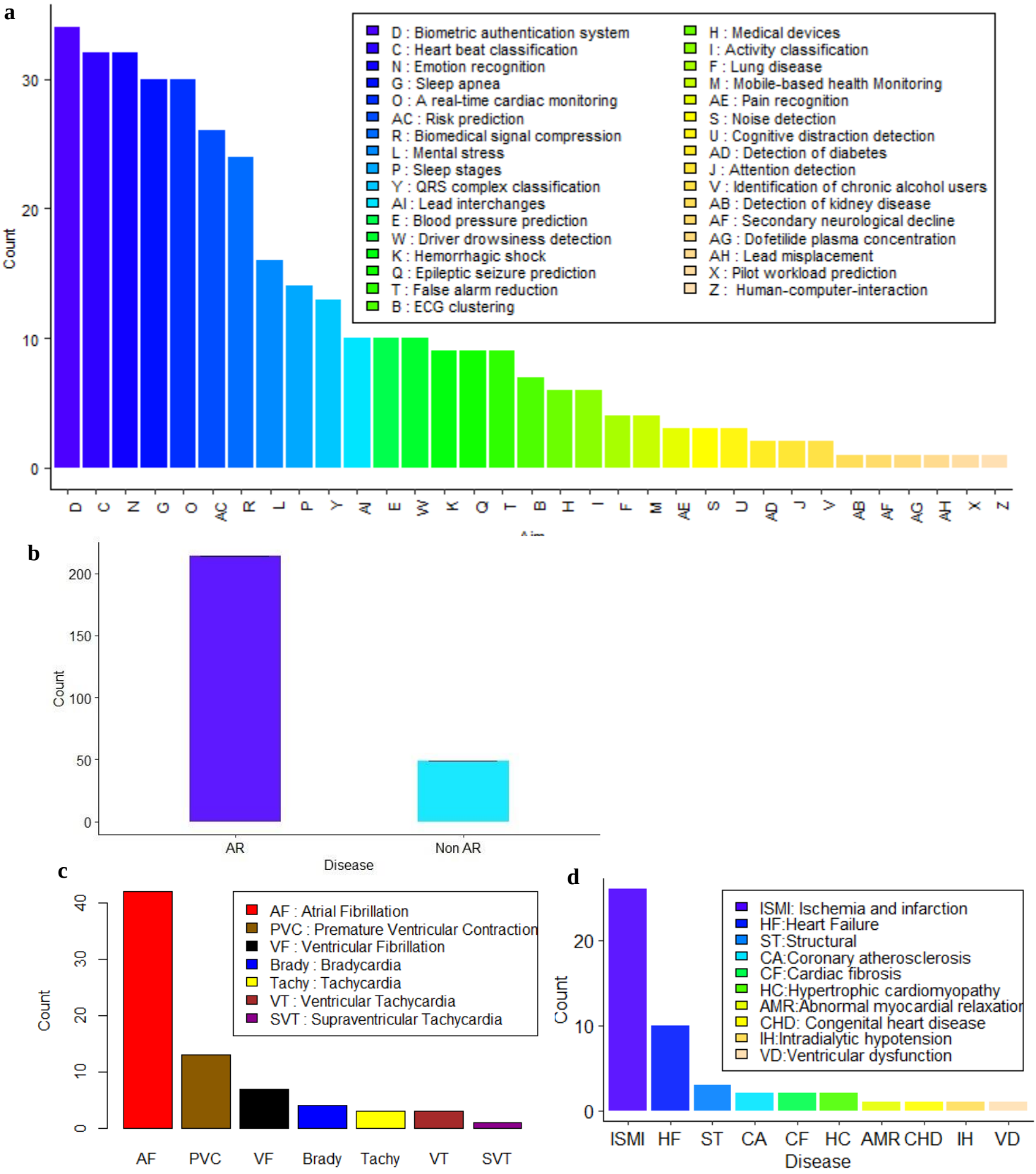

Figure 6: ECG application topics. **a** represents frequency of each application topic using ML. **b** is the frequency of each arrhythmia and non-arrhythmia group. **c** shows the frequency of each type in the arrhythmia group. **d** represents the frequency of each type in the non-arrhythmia group.

In the non-arrhythmia group, most studies focused on detecting ischemia and infarction (44.79%) followed by heart failure, structural abnormalities, coronary atherosclerosis, cardiac fibrosis, hypertrophic cardiomyopathy, abnormal myocardial relaxation, congenital heart disease, intradialytic hypotension and ventricular dysfunction as shown in Figure 6d.

3.4 Meta-analysis for ECG classification

Meta-analysis was applied to the arrhythmia and non-arrhythmia group to show the performance of ML algorithms for classifying each arrhythmia and non-arrhythmia based on the mean sensitivity and the mean specificity. In the arrhythmia group, as shown in Figure 7, using ML to detect tachycardia (it can be detected from simple threshold crossing and in some situations with relevant age related definitions) achieved the highest scores of sensitivity and specificity with 94.5% and 97.3% respectively, while algorithms to detect other arrhythmias achieved lower performance scores. As shown in Figure 7, KNN models outperformed the other classifiers for AF recognition (sensitivity 99% and specificity 95%), while SVM achieved the best sensitivity (92.5%) and specificity (98.5%) scores for premature ventricular contraction detection. These results demonstrate the 'no free lunch theorem' which postulates that no single ML technique can be the winning classifier for all problems. Some results being reported also seem likely to be dependent on a clean test set and are unlikely to transfer or generalise to other hospital ECG data sets that perhaps exhibit noise and artefacts with greater variations.

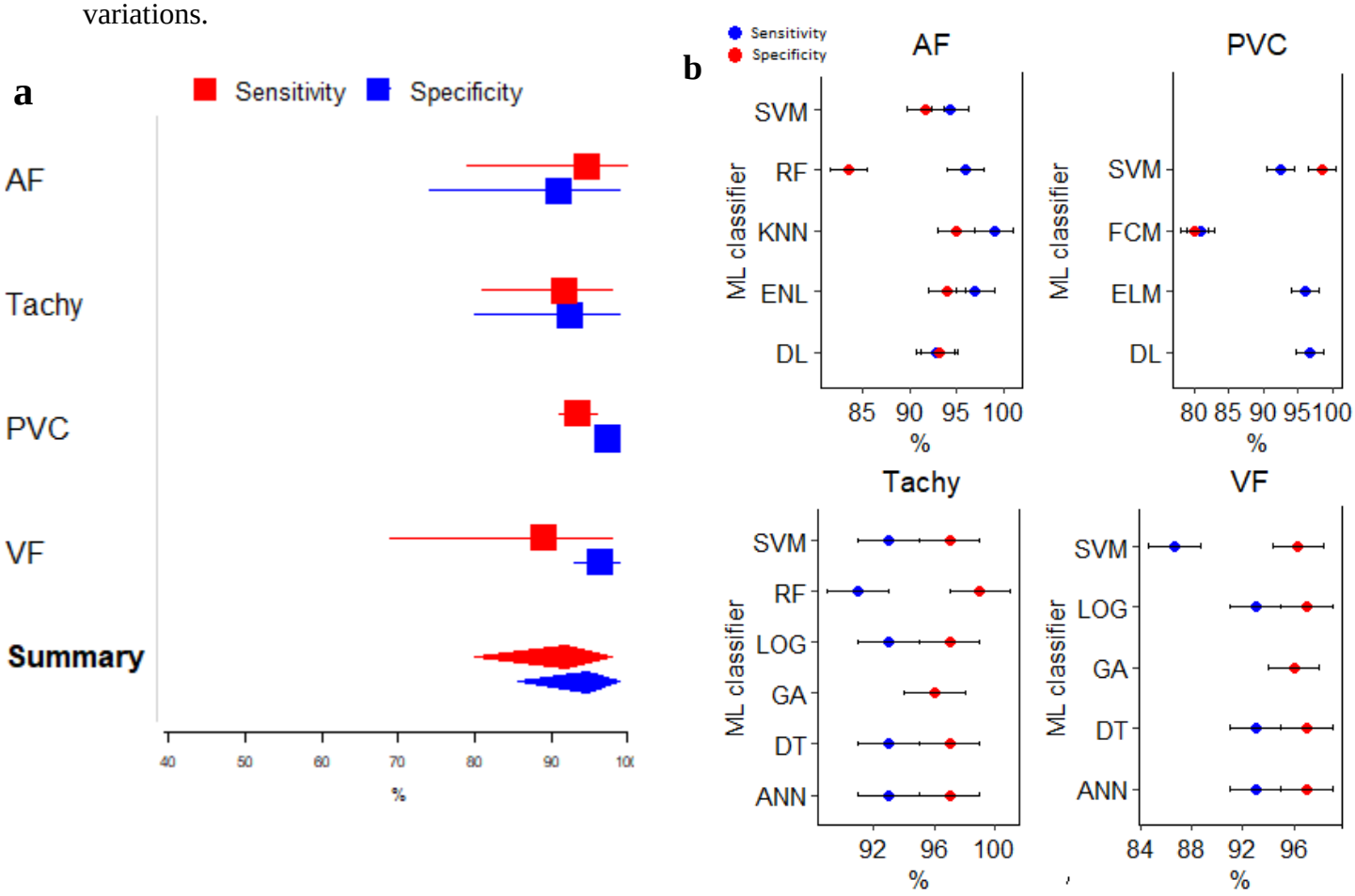

Figure 7: A meta-analysis showing the performance of ML for cardiac arrhythmia classification. **a** summarised mean sensitivity and specificity of ML methods (using forest plot) for detecting each arrhythmia. **b** shows mean sensitivity and specificity of each ML algorithm for detecting each arrhythmia.

For tachycardia, all ML classifiers obtained a high similar performance showing that the obvious ECG morphological features of tachycardia are prominent in most cases which enable almost any ML technique to detect them easily. However, these studies are very small in number and good test sets of ventricular tachycardia, for example, are rare.

In the non-arrhythmia group, some types were not included in the meta-analysis because they used different metrics such as accuracy or F1 score instead of sensitivity and specificity. SVM outperformed the other classifiers for detecting coronary atherosclerosis (sensitivity 92% and specificity 94%). KNN and the decision tree achieved the best performance with a sensitivity of 99% and specificity 99% for detection of heart failure which strictly is not an ECG diagnosis. Such results might suggest that the control population consisted of healthy individuals and the heart failure population was selected from those in the most severe category. In ischemia and infarction detection [28], latent topic multiple instance learning obtained the best sensitivity (95%) and specificity (82%) as shown in Figure 8.

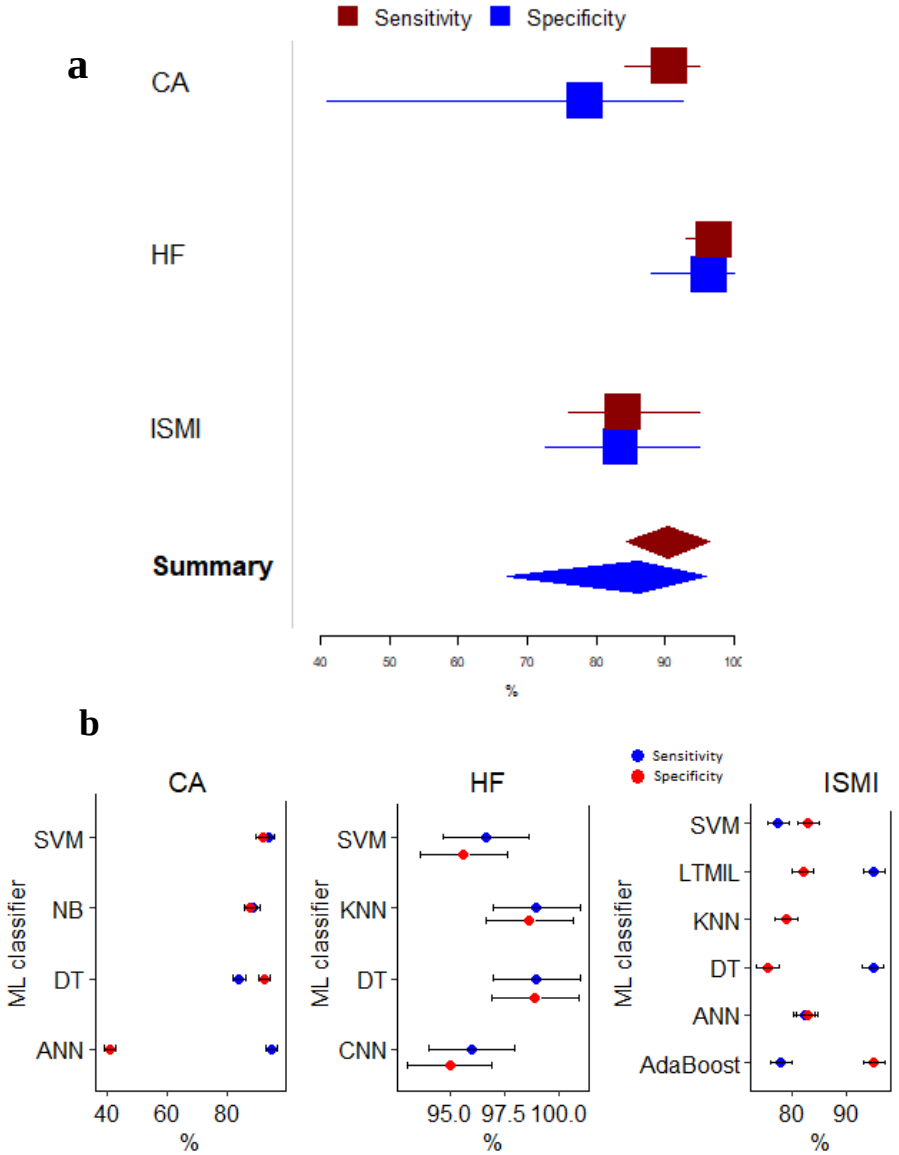

Figure 8: A meta-analysis showing the performance of ML for cardiac abnormality classification in the non-arrhythmia group. **a** summarise mean sensitivity and specificity of ML generally (using forest plot) to detect some abnormality in the non-arrhythmia group. **b** shows mean sensitivity and specificity of each ML algorithm particularly to detect coronary atherosclerosis (CA), heart failure (HF) as well as myocardial ischemia and infarction (ISMI).

As shown in Figures 6 and 7 and according to the five tribes taxonomy, analogizer algorithms such as KNN and SVM achieved the highest sensitivity and specificity in three classification problems related to cardiac arrhythmias (AF, premature ventricular contraction and tachycardia) and in two out of three classification problems in the non-arrhythmia group (coronary atherosclerosis and heart failure).

4. Discussion

According to the articles reviewed, different ML algorithms were applied using ECG data for different purposes, from disease classification to detection of driver drowsiness. The meta-analysis focused more on the use of ML and ECG data for detecting cardiac problems, while other topics such as emotion detection and activity classification were not meta-analysed due to a lack of multiple independent studies. In addition, there was significant variability in the use of different metrics (e.g. accuracy, sensitivity, specificity and AUC), ML classifiers and databases in the studies that were reviewed. This kind of variability made it difficult to compare or combine the results from different studies and different ML algorithms. According to the meta-analysis results, ECG analysis using ML showed promising results, especially in detecting cardiac abnormalities. However, ML generally and DL especially do not perform well when using small datasets. Another key issue is the lack of transparency when using some of the ML algorithms that are being used. DL algorithms are considered a 'black box' which means that there is no transparency and as a consequence, there is no explanation available to the user to provide some rationale as to what is going on inside the black box or why the DL algorithm provides a specific output or disease classification. Experts suggest that there is an inverse relationship between ML performance/accuracy and explainabilty, where the higher performing techniques such as DL are less explainable, and hence Occam's razer does not seem to hold true for ML algorithms (a ML algorithm can be considered a hypothesis, hence the reference to Occam's razer) [29]. Computer scientists have suggested new methods such as attention maps to show which feature is considered important to the DL network before it makes a prediction. Most AI interventions, particularly diagnostic algorithms, were evaluated only in the context of diagnostic accuracy. Hence, the CONSORT (Consolidated Standards of Reporting Trials) and SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) groups are trying to prepare an international consensus to address challenges such as transparency and the reporting of the new AI intervention or ML results. CONSORT and SPIRIT plan to publish a standard reporting guideline very shortly. Hence, this review paper could be used to help them in their first steps which include "a systematic literature search for any existing recommendations and current practice for clinical trials of AI interventions" [30]. The US Food and Drug Administration (FDA) needs to ensure the safety and effectiveness of AI algorithms which might take considerable time, because the FDA requires a premarket submission to approve new AI algorithms [30][31]. Since 2019, the FDA has approved more than 30 artificial intelligence (AI) algorithms [30].

Limitations

The full text of some papers was not freely accessible using the databases available to the research team. Hence, they were only included in the frequency analysis, but were excluded in the meta-analysis. Most papers used different datasets and applied different ML techniques which may explain the variability in results that can be partly explained by the different types of datasets and the size of the dataset. Some papers that were published by authors who are not from computer science backgrounds did not specify which DL algorithm was applied, and hence most of those papers were combined into one category called DL which might include convolutional neural networks, deep neural networks or long short-term memory networks.

This emphasises the importance of engineers/computer scientists co-authoring and collaborating with medical experts to ensure that the right terminology is being used and that clinical knowledge is being applied appropriately. In addition, it is critically important that engineers and data scientists work with medical experts so that clinically useful contributions can be presented.

5. Conclusion

This study was conducted to evaluate the performance of ML on ECG data in the last ~24 years regardless of the application area. Table 2 shows the general research questions presented mapped to the answers according to our findings which highlight the fact that ML is a rapidly evolving area in electrocardiology.

Table 2: Key research questions that have been answered by this work.

Research question	Answers
1. What is the general growth in ML and ECG research according to the number of pertinent papers published?	According to this study, the total number of publications that used ML with ECG data before 2012 was 117, while the total number after 2012 was 640 ($P<0.001$).
2. What are the trends and frequencies in the use of different ML techniques?	ANN was the leading technique from 1995 to 1997. From 1997 to 2000, NB and HMM were trending upwards. SVM has since been the trend and the most frequently used ML algorithm for almost two decades from 2000-2019 with two significant peaks in 2012 and 2018. However, from 2018, DL has appeared and used for ECG classification with promising results.
3. What are the trends and frequencies in the use of various metrics that have been used by researchers to evaluate ML algorithms?	Accuracy has been trending upwards significantly from 2006 to 2019 for evaluating ML algorithms. Before 2006, there was no significant trend towards using a specific metric. From 2008, sensitivity and specificity started increasing significantly, while from 2016, the use of F1 scores started increasing in use and has become the fourth most popular metric. However, accuracy still remains the dominant metric in terms of reporting ML performance [30].
4. What are the most common applications of ML when using ECG data?	The most common applications are healthcare applications/cardiac abnormality classification ($p<0.001$) as shown in Figure 6. There is a specific dominance for AF detection followed by myocardial ischemia and infarction and premature ventricular contraction. New emerging areas were included such as detection of ECG electrode misplacement and real time cardiac monitoring.
5. How has ML performed when using it to classify ECGs to assist in patient diagnosis?	ML outperformed physicians in detecting some arrhythmias such as AF. However, ML algorithms still require physician over-reading because ML is sub-optimal for many classifications according to the meta-analyses presented in Figures 6 and 7.

It shows promising results in medical applications, which emphasises the importance of applying ML in healthcare. However, DL requires future investigation in terms of performance when being trained using a large amount of ECG data. Hence, availability of large digital health records that include ECGs could motivate researchers to investigate the performance of DL models to detect different cardiac problems. However, DL has a number of challenges such as a lack of transparency and explicability, so new techniques and methods should be developed to provide explainable DL approaches to show which features are responsible for influencing the algorithmic decision. One implication of this work is a need for standardised reporting of data provenance and a standard protocol for presenting ML results with the use of universal metrics that would allow for unambiguous meta-analysis. Presenting a sample of the ECGs that are being used to test the algorithm is important. For example, an algorithm might perform very well when trained and tested using archetypical textbook quality ECGs. However, the real environment can involve ECGs that exhibit mains noise and baseline wonder along with other artefacts caused for example by poor electrode connections. The algorithm may become useless when applied to these datasets. In forthcoming years, around 50% of hospitals plan to invest in ML because they realise that ML could improve patient outcomes and reduce costs. However, only 50% of decision makers in healthcare are familiar with the concept of AI and ML, and hence there remain a number of challenges before ML can be used in routine clinical practice [32].

Acknowledgments:
This work is supported by the European Union's INTERREG VA programme, managed by the Special EU Programmes Body (SEUPB). The work is associated with the project – 'Centre for Personalised Medicine – Clinical Decision Making and Patient Safety'. The views and opinions expressed in this study do not necessarily reflect those of the European Commission or the Special EU Programmes Body (SEUPB).

References:

[1] Suh W. et al. Importance of electrocardiography in the assessment and management of heart failure: a case report. Journal of Electrocardiology. Vol.41, pp.44-48, 2008.

[2] Bond R, Finlay D, et al. The effects of electrode misplacement on clinicians' interpretation of the standard 12-lead electrocardiogram. Eur J Intern Med. Vol 23(7), pp:610-5, 2012. https://doi.org/10.1016/j.ejim.2012.03.011.

[3] Kania M. & Maniewski R. Optimal ECG Lead System for Exercise Assessment of Ischemic Heart Disease. Journal of Cardiovascular Translational Research,(2019). https://doi.org/10.1007/s12265-019-09949-3.

[4] Maršánová, L., Ronzhina, M., Smíšek, R. et al. ECG features and methods for automatic classification of ventricular premature and ischemic heartbeats: A comprehensive experimental study. Sci Rep 7, 11239 (2017). https://doi.org/10.1038/s41598-017-10942-6.

[5] Sanjay Purushotham, et al. Benchmarking deep learning models on large healthcare datasets. Journal of Biomedical Informatics, Volume 83, pp.112-134, 2018. https://doi.org/10.1016/j.jbi.2018.04.007

[6] Miyazawa AA. Artificial intelligence: the future for cardiology. Heart 2019;105:1214.http://dx.doi.org/10.1136/heartjnl-2018-314464.

[7] Irfan, Muhammad, Hameed, Ibrahim A. Deep learning based classification for healthcare data analysis system . 2007 IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2007. IEEE Transactions on Instrumentation and Measurement. 2018. https://doi.org/10.1109/BESC.2017.8256396.

[8] Walinjkar, Amit, "Woods, John" . ECG classification and prognostic approach towards personalized healthcare . Proceedings International Electronics Symposium on Knowledge Creation and Intelligent Computing, IES-KCIC 2017 . IEEE Access . 2017. https://doi.org/10.22489/CinC.2017.168-168.

[9] Taji, Bahareh, Chan, Adrian D.C et al. Classifying measured electrocardiogram signal quality using deep belief networks. 1st International Conference on Robotics, Electrical and Signal Processing Techniques, ICREST 2019. IEEE Access. 2018 . https://doi.org/10.1109/TBME.2013.2279998.

[10] Li, Qiao, Rajagopalan, Cadathur et al. A machine learning approach to multi-level ECG signal quality classification. Advances in Intelligent Systems and Computing. Physiological Measurement. 2014. https://doi.org/10.1587/transcom.2018hmi0001.

[11] Rjoob K. , Bond R. et al. Data driven feature selection and machine learning to detect misplaced V1 and V2 chest electrodes when recording the 12-lead electrocardiogram. Journal of Electrocardiology, Vol.57, Pages 39-43, 2019 .https://doi.org/10.1016/j.jelectrocard.2019.08.017.

[12] Rjoob K., Bond R. Machine Learning Improves the Detection of Misplaced v1 and v2 Electrodes During 12-Lead Electrocardiogram Acquisition. CinC 2019. https://doi.org/10.22489/CinC.2019.035.

[13] Finlay D. . et al. Effects of electrode placement errors in the EASI-derived 12-lead electrocardiogram. Journal of Electrocardiology. Vol.43, pp.606–611, 2010. https://doi.org/10.1016/j.jelectrocard.2010.07.004.

[14] Bond R., Dewar D. et al. Human factors analysis of the CardioQuick Patch®: A novel engineering solution to the problem of electrode misplacement during 12-lead electrocardiogram acquisition. Journal of Electrocardiology. Vol.49, pp.911–918, 2016. https://doi.org/10.1016/j.jelectrocard.2016.08.009.

[15] Bong, Siao Zheng, Murugappan, M et al. Analysis of electrocardiogram (ECG) signals for human emotional stress classification. 2019 8th International Conference on Modern Circuits and Systems Technologies, MOCAST 2019. IEIE Transactions on Smart Processing and Computing 2017. https://doi.org/10.1186/1472-6947-15-S3-S5.

[16] Ferdinando, H., lasaarela, E. et al. Enhancing emotion recognition from ECG signals using supervised dimensionality reduction. Communications in Computer and Information Science. Neurocomputing. 2018. https://doi.org/10.1016/j.bspc.2018.03.014.

[17] Xia, Vitoria, Jaques, Natasha et al. Active learning for electrodermal activity classification . Computers in Cardiology 1997. IEEE Sensors Journal. 2016. https://doi.org/10.1109/SPMB.2015.7405467.

[18] Uddin, Md Zia & Hassan, Mohammad Mehedi . Activity Recognition for Cognitive Assistance Using Body Sensors Data and Deep Convolutional Neural Network . 1st International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2019. Healthcare Technology Letters. 2019. https://doi.org/10.1109/BSN.2018.8329658.

[19] Raihan, M., Mandal, Parihay Kumar et al. Risk Prediction of Ischemic Heart Disease Using Artificial Neural Network. Proceedings 21st Euromicro Conference on Digital System Design, DSD 2018. IEEE Journal of Biomedical and Health Informatics. 2019. https://doi.org/10.1109/HNICEM.2018.8666331.

[20] Babaeian, Mohsen & Mozumdar, Mohammad. Driver drowsiness detection algorithms using electrocardiogram data analysis. IEEE Sensors Journal. 2007. https://doi.org/10.1109/IEMBS.2011.6090220.

[21] Manna, Tishya, Swetapadma, leena et al. Decision Tree Predictive Learner-Based Approach for False Alarm Detection in ICU. Advances in Intelligent Systems and Computing. Sensors (Basel, Switzerland). 2019. https://doi.org/10.1155/2013/158056.

[22] Taji, Bahareh, Chan, Adrian D.C. et al. False Alarm Reduction in Atrial Fibrillation Detection Using Deep Belief Networks. Proceedings of 2010 IEEE EMBS Conference on Biomedical Engineering and Sciences, IECBES 2010. IEEE Journal of Biomedical and Health Informatics. 2009. https://doi.org/10.1109/ECACE.2019.8679362.

[23] Hejazi, Maryamsadat, Al-Haddad, S. A.R. et al. Non-fiducial based ECG biometric authentication using one-class Support Vector Machine. 2018 IEEE Biomedical Circuits and Systems Conference, BioCAS 2018 Proceedings. IEEE Access. 2017. https://doi.org/10.1109/ICCKE.2018.8566583.

[24] Derawi, Mohammad. Wireless chest-based ECG biometrics. Advances in Intelligent Systems and Computing . ISA Transactions. 2015. https://doi.org/10.1186/1472-6947-12-116.

[25] Domingos, P. (2015). The master algorithm: How the quest for the ultimate learning machine will remake our world. Basic Books. [Requires publisher etc]

[26] Chung Chen and Lon-Mu Liu. Forecasting time series with outliers. Journal of forecasting. Vol.12, pp.13-35, 1993. https://doi.org/10.1002/for.3980120103

[27] Valverde-Albacete FJ. 100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox. PLOS ONE 9(1): e84217, 2014, https://doi.org/10.1371/journal.pone.0084217

[28] Sudarshan, Vidya K. et al. Automated Diagnosis of Congestive Heart Failure Using Dual Tree Complex Wavelet Transform and Statistical Features Extracted from 2 s of ECG Signals. Computers in Biology and Medicine, Vol.83, pp.48–58, 2017. https://doi.org/10.1016/j.compbiomed.2017.01.019

[29] Pintelas, E., et al. Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability. Algorithms, Vol.17, pp.1-17, 2020. https://doi.org/10.3390/a13010017

[30] Liu, X., Rivera, S.C., Faes, L. et al. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nat Med 25, 1467–1468 (2019). https://doi.org/10.1038/s41591-019-0603-3.

[31] US Food and Drug Administration (FDA). Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback. FDA, 2020. https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf

[31] DAIC(AUGUST 08, 2019). Retrieved from https: // www .dicardiology. com/ content / half- hospital -decision-makers-plan-invest-ai-2021.