THE USE OF ARTIFICIAL INTELLIGENCE AND DECISION SUPPORT SYSTEMS IN CLINICAL DIAGNOSIS: A SYSTEMATIC REVIEW

Azzam F. G. Taktak

Department of Clinical Engineering, Royal Liverpool University Hospital, Duncan Building, Liverpool L7 8XP, UK
email: afgt@liv.ac.uk

Abstract

In this paper, a systematic review was conducted to assess the benefit of Artificial Intelligence (AI) algorithms as diagnostic tools, the majority of which involved the use of Artificial Neural Networks (ANN). The number of clinical trials involving the use of ANN in diagnosis increased from 1 to 38 in the last decade. However, the proportion of clinical trials in all studies published using ANN in diagnosis was only 1:16. In the area of cancer, there were 396 studies involving the use of ANN in diagnosis, only 24 of which were clinical trials. In artefact detection, there were 269 studies using ANN and other AI techniques, only 9 of which were clinical trials. Out of these trials, 20 showed an increase in benefit to healthcare provision, 7 did not and 6 were not clear. None however showed decrease in benefit.

Introduction

In the last decade, the use of artificial intelligence (AI) has become widely accepted in medical applications. There is an increasing number of medical devices currently available on the market with embedded AI algorithms. These algorithms offer a number of advantages:

• They can analyse a vast amount of data in relatively short periods of time and can therefore be more cost effective
• They can be designed to mimic the expert’s knowledge thereby facilitating distribution of expertise
• They make knowledge dissemination easier by providing explanation
Modern medicine generates huge amounts of data. Much of this data is contaminated with artefact and information is usually hidden within the data. Clinicians often rely on knowledge and past experience to separate good data from artefact to make diagnosis. This is very subjective, prone to human error and can be rather time consuming. AI algorithms have been shown to be valuable tools in reducing the workload on the clinicians by detecting artefact, providing decision support and the ability to learn and adapt on-line. The number of clinical trials conducted in this area however remains very limited. In cancer, the use of AI techniques offer many advantages due to problems associated with the number of variables, uncertainties and missing values in the data. AI techniques are better suited than traditional linear methods in dealing with these problems because they do not assume any underlying mathematical formula in the data. Although the number of clinical trials in this area has increased in the last decade, there are many drawbacks associated with such techniques which will also be highlighted and discussed.

Scientific literature is overwhelmed with papers describing the theory and applications of AI systems. The most commonly used systems include artificial neural networks (ANN), fuzzy logic, machine learning, clustering analysis and rule based systems. All these techniques have one thing in common; they are all data driven and some more than others (viz. ANN) are particularly data hungry. Whilst the science behind these techniques has improved significantly in the last decade, evidence of their clinical utility remains under scrutiny. A common method in assessing the efficacy of a particular technology or a group of technologies with a common theme is to look at the number of clinical trials it has undergone (Lisboa 2002).

This paper reviews the use of AI in general and more specifically ANN in diagnostic applications in two areas; cancer and artefact detection. The paper will then demonstrate the combination of a number of these techniques to create an artefact-free diagnostic system in neonatal ventilation management.

**Literature Search**

The literature search was conducted using Pubmed with the keywords (artificial intelligence) or (artificial neural networks) and (diagnosis). In the case of ANN, there were 2697 hits but when these were filtered to look at
clinical trials, the number dropped to 173 giving a ratio of approximately 1:16. Further filtering was carried out using the keywords cancer and artefact (separately) and the abstracts of the resulting hits were analysed. The effectiveness of this technique has been shown to have 50-60% sensitivity (Gant at al 2001).

The use of ANN in cancer has been on the increase in the last decade (fig. 1). This is mainly due to advances in data acquisition systems making it easier to generate electronic databases thereby storing large amount of data, which is what ANN systems thrive on.

Some trials showed clear added benefit in using ANN or other AI techniques in the two areas reviewed. Others, were only able to show that these techniques performed as well as traditional methods. A third group showed that there are advantages and disadvantages in using such techniques. The trials were therefore classified into two categories: those that did show an added benefit containing the first group and those that did not containing the second and third groups. There was a fourth group where it was unclear from the abstract into which category they belonged. None of the trials examined showed conclusive decrease in benefit in using these techniques.

Cancer

Cancer is one of the most common diseases in Europe and the main cause of mortality amongst its citizens. The three most common pathologies are: breast, prostate and lungs. Figures provided by the International Agency for Research on Cancer (IARC) and International Association for Cancer Registries (IACR) (Parkin et al. 1997) seem to suggest that the first two are more common in Western Europe and the last in Eastern Europe.

Many clinicians have realised the potential in the use of ANN as an aid tool in the analysis. The main concern over the ‘black box’ issue has been addressed by a number of researchers by providing a statistical framework for ANN (Biganzoli et al. 1998, Lisboa et al. 2003, Ripley & Ripley 2001). ANN have many advantages in providing much wider (but not infinite) flexibility in fitting models to data where patterns is not so obvious (Ripley 1996). The use of a large number of input variables, different types of variables and the ability to cope with heterogeneity in the data provide some of the attractions in the use of ANN with such databases.
The number of clinical trials using ANN in clinical diagnosis in cancer are shown in fig. (2). These studies were categorised according to the type of cancer and this was plotted alongside the number of incidence (per 100,000) in Europe.

The majority of clinical trial studies compared the ANN performance with traditional screening methods. In prostate cancer, this involves the use of prostate specific antigen (PSA) serum marker, digital rectal examination, Gleason sum, age and race (Gamito et al. 2000, Tewari et al. 2001, Stephan et al. 2003, Remzi et al. 2003). Some studies have compared ANN with statistical methods (Remzi, et al 2003, Chan et al. 2003, Finne et al. 2000, Matsui et al. 2002). Remzi demonstrated that ANN is more accurate than multivariate logistic regression (LR) using ROC analysis and therefore reduced the number of unnecessary repeat biopsies. They went on to conclude that their system would allow individual counselling of patients with an initial negative biopsy. Finne on the other hand showed that ANN and LR are both accurate than PSA alone and also reached the conclusion that they reduced the number of unnecessary repeat biopsies. Cervical cancer applications concentrated mainly on evaluating the benefits of the widely known PAPNET system (Doornewaard et al. 1999, Kok & Boon 1996, Mango & Valente 1998, Nieminen et al. 2003, Sherman et al. 1997), one of very few ANN systems to gain FDA approval for clinical use. The system uses ANN to extract abnormal cell appearance from vaginal smear slides and describe them in histological terms (Boon & Kok 2001). The alternative more conventional way is to re-screen the slides under the microscope. Mango and Valente have shown that the PAPNET system has uncovered a higher proportion of false negatives than conventional microscopic re-screening as confirmed by cytologists. Sherman looked at the results of PAPNET in 200 specific cases where initial screening was inconclusive and compared them with conventional microscopy, DNA analysis and biopsy. The study showed that for these cases, PAPNET would have reduced unnecessary biopsies but at the expense of increasing false positives. Parekattil (Parekattil et al 2003) showed in a clinical trial on bladder cancer that their ANN model was more accurate in identifying patients who required cystoscopy thereby providing possible savings. Whilst the majority of ANN algorithms used in the trials concentrated on multilayer
perceptrons (MLP), a rare combination of a hybrid system combining non-supervised Kohonen self-organising map (SOM) with MLP was used by Glass and Reddick (Glass & Reddick 1998) to study the response of paediatric osteosarcoma to chemotherapy using MRI images. The technique showed a high correlation with histopathologic analysis. One study used ANN as a validation method for another technique, namely electrical impedance spectroscopy to separate basal cell carcinoma (BCC) from benign skin lesion (Dua et al. 2004). A complete list of the trials with the clinical application, number of subjects, methods used and major findings is presented in table (1). As explained earlier, some studies showed a clear added benefit in the use of ANN techniques in cancer diagnosis whereas others did not. Fig. (3) separates these two groups and shows them plotted against the number of subjects in the trial as a method of assessing the statistical power of such trial.

**Artefact Detection**

ANN and other AI systems have also been used widely in Physiological Measurement as a diagnostic tool. Such measurements are riddled with artefact which may have a physiological or a technical origin. Diagnostic procedures should be highly sensitive so as to reduce the risk of missed positives and at the same time not be sensitive to noise. Separating artefact from good data in large recordings is very time consuming and requires expertise. An automated system that carries out this task can provide major impact on the provision of clinical service with added benefit to the patient.

A variety of techniques were used in artefact detection. These included fuzzy logic, ANN, clustering analysis, principal component analysis (PCA) and independent component analysis (ICA). As before, a list of all the applications is shown in table (2) including the number of subjects, methods used and major findings. Fig. (4) separates the studies reviewed into two groups and plots against the number of subjects in the trial.

**Example in Artefact Detection**

We have previously carried out some work related to neonatal and paediatric oxygenation and ventilation management. We developed various algorithms based on a number of techniques to remove artefact, diagnose
and recommend therapy. The work started in 2000 when we developed an ANN system which uses oxygen saturation (SaO2) and transcutaneous oxygen values to detect and remove artefact (Taktak et al. 2000). We extended the technique to pulse plethysmography signals using fuzzy logic (Belal et al. 2001) and adaptive Neuro-fuzzy system (ANFIS) (Belal et al. 2002). We developed a ventilator management system which uses the artefact-free signal values and trends to provide diagnosis and make recommendations ((Taktak et al. 2003)). The motivation behind this work is to model the clinician’s thinking process in interpreting the information presented to them when they make their diagnosis. In neonatal intensive care, if we consider a single task such as altering the ventilator settings, the clinician is presented with the following data:

- data from the multi-parameter bedside monitor (e.g. SaO2, TcPCO2, Heart Rate, etc.) typically at a rate of 1 sample/second
- ventilator measured parameters (e.g. Peak Inspiratory Pressure, Positive End-Expiratory Pressure, Tidal Volume, etc.) at a rate of 1 sample/second
- invasive blood-gas measurement (e.g. pO2, pCO2, Total Haemoglobin) at a variable rate of 1 sample/1-4 hours
- other variables that can affect ventilation management of the neonate such as suctioning, handling, etc.

The system incorporated a knowledgebase acquired from 4 clinicians and was validated on data from 7 neonates. The sensitivity, specificity and accuracy figures of the advisor were 93%, 94% and 93% respectively.

**Ethical and Legal Issues**

Before embarking on using ANN or any other AI techniques clinically, there are a number of legal and ethical issues that must be addressed. The first question is: are such systems considered as a medical device and therefore subject to the Medical Device Directive (MDD) and CE marking? Article 1 in the Directive defines a medical device as “any instrument, apparatus, appliance, material or other article whether used alone or in combination, including the software necessary for its proper application intended by the manufacturer to be used for human beings for the purpose of: diagnostic,
prevention, monitoring, treatment or alleviation of disease, …”. This suggests that software whether standalone or part of a device can be a medical device. The view of the UK regulatory body, The Medicines and Healthcare Products Regulatory Agency (MHRA), on software is summarized in table (3). Taking this into account therefore, the type of ANN and other AI systems covered in this review would probably fall under the definition of a medical device. Suppliers of medical devices are under legal obligation to demonstrate that the device meets the Essential Requirements detailed in Annex I in the MDD. The most reliable way of ensuring that a device complies with the Essential Requirements is to ensure its compliance with the appropriate harmonized standards such as the IEC 60601-1-4 in the case of software. Unfortunately none of these or any other current European standards make a special case for the incorporation of AI in software so it is difficult to know how to meet the Essential Requirements for such systems. However, the Food and Drug Agency in the USA have issued a guidance document for software in medical devices which is based on the European IEC 60601-1-4 standards (FDA 1998). The document has a section on expert systems and ANN software which provides useful tips for manufacturers and assessors. Some of the points highlighted by this document are that ANN can behave in a non-deterministic manner. The designer should therefore justify and explain the choices made for the artificial neural network model, topology, and training sets, as well as explain and justify the data set class that the ANN is intended to analyze or process. The designer should also describe how overfitting was avoided and should demonstrate how the relevant features were extracted (such as a specific pattern to be detected) and not a peculiarity contained only in the training set. The document sets out a requirement for additional data sets to be processed through the network to ensure that the performance remains as expected.

Discussion

A review of the literature revealed that although the development of AI and ANN systems has been on the increase, their use in clinical practice remains very limited. It is evident that apart from the PAPNET system, there are no other such systems that have been incorporated seamlessly into everyday practice. The number of clinical trials in this area remains limited.
and in the area of cancer diagnosis these trials are not targeting the most common types of cancer. This seems to indicate a the ‘thin-spread’ of expertise in this area and the lack of integration between different disciplines in the scientific community. On the positive side, the evidence from these trials seems to suggest a clear bias towards added benefit in the use of such systems. Before such systems can be fully integrated into clinical practice however, there is a number of ethical and legal issues to be addressed. Gant presented a thorough argument for the use of ANN and suggested a framework for their applications in medical practice (Gant et al 2001). The framework recommended the establishment of a statutory regulatory body to ensure continued quality performance.

Fig. (1)
Number of Clinical Trials involving the use of ANN in clinical diagnosis in the last decade

Fig. (2)
Number of incidence of different types of cancer in Europe on the left axis and number of Clinical Trials for each type on the right axis
Table (1) A list of clinical trials involving the use of AI in cancer diagnosis.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Application</th>
<th>Subjects</th>
<th>Methods</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chan et al (2003)</td>
<td>Prostate</td>
<td>N/A</td>
<td>MLC, SVM, FLD</td>
<td>F LD and SVM have superior performance compared to MLC</td>
</tr>
<tr>
<td>Doornewaard et al (1999)</td>
<td>Cervical</td>
<td>846 patients and 52 normals</td>
<td>PAPNET</td>
<td>PAPNET has similar diagnostic value as conventional screening</td>
</tr>
<tr>
<td>Dua et al (2004)</td>
<td>Basal Cell Carcinoma</td>
<td>18 BCC and 16 benign</td>
<td>ANN and electrical impedance spectroscopy</td>
<td>Electrical impedance may be a promising clinical diagnostic tool</td>
</tr>
<tr>
<td>Finne et al (2000)</td>
<td>Prostate</td>
<td>656 (23% malignant, 77% benign)</td>
<td>MLP, LR</td>
<td>MLP and LR could reduce number of FP</td>
</tr>
<tr>
<td>Gamito et al (2000)</td>
<td>Prostate</td>
<td>5099</td>
<td>ANN</td>
<td>ANN has the potential to increase accuracy in clinical staging and improving treatment</td>
</tr>
<tr>
<td>Glass &amp; Reddick (1998)</td>
<td>Paediatric Osteosarcoma</td>
<td>43</td>
<td>SOM followed by MLP</td>
<td>The technique had a high correlation with histopathological analysis</td>
</tr>
<tr>
<td>Gletsos et al. (2003)</td>
<td>Liver</td>
<td>N/A</td>
<td>MLP, GA</td>
<td>GAs improved classification performance</td>
</tr>
<tr>
<td>Kok &amp; Boom (1996)</td>
<td>Cervical</td>
<td>91,294</td>
<td>PAPNET</td>
<td>PAPNET enhanced screening efficacy</td>
</tr>
<tr>
<td>Kothari et al (1996)</td>
<td>Leukemia</td>
<td>170</td>
<td>MLP</td>
<td>Objective classification and further categorisation can be achieved successfully</td>
</tr>
<tr>
<td>Mango &amp; Valente (1998)</td>
<td>Cervical</td>
<td>10,000</td>
<td>PAPNET</td>
<td>PAPNET uncovered a significant higher proportion of FN than conventional screening</td>
</tr>
<tr>
<td>Matsui et al (2002)</td>
<td>Prostate</td>
<td>178</td>
<td>ANN, LR</td>
<td>ANN might possibly evolve in clinical staging</td>
</tr>
<tr>
<td>Ravdin et al. (1992)</td>
<td>Breast</td>
<td>N/A</td>
<td>ANN</td>
<td></td>
</tr>
<tr>
<td>Naguib et al. (1996)</td>
<td>Breast</td>
<td>81</td>
<td>ANN</td>
<td>ANN can provide strong indicators as to lymph node status</td>
</tr>
<tr>
<td>Ng et al. (2001)</td>
<td>Breast</td>
<td>N/A</td>
<td>ANN</td>
<td></td>
</tr>
<tr>
<td>Nieminen et al (2003)</td>
<td>Cervical</td>
<td>108,686</td>
<td>PAPNET</td>
<td>PAPNET was feasible as part of routine screening and performed as well as conventional methods</td>
</tr>
<tr>
<td>Parekattil et al (2003)</td>
<td>Bladder</td>
<td>253</td>
<td>ANN</td>
<td>ANN model is superior to conventional screening tests and can provide possible savings</td>
</tr>
<tr>
<td>Remzi et al (2003)</td>
<td>Prostate</td>
<td>820</td>
<td>Hybrid (MLP and GA)</td>
<td>ANN reduced unnecessary repeat biopsies significantly</td>
</tr>
<tr>
<td>Sherman et al (1997)</td>
<td>Cervical</td>
<td>200</td>
<td>PAPNET</td>
<td>Referrals would have been reduced but at the cost of lower sensitivity</td>
</tr>
<tr>
<td>Simpson et al. (1995)</td>
<td>Breast</td>
<td>41 precancerous and 50 normals</td>
<td>FLD, ANN</td>
<td>High sensitivity and specificity achieved</td>
</tr>
<tr>
<td>Tewari et al (2001)</td>
<td>Prostate</td>
<td>1,400</td>
<td>ANN</td>
<td>Recurrence could be predicted with high accuracy</td>
</tr>
<tr>
<td>Tomatis et al. (2003)</td>
<td>Skin</td>
<td>534</td>
<td>MVDA, ANN</td>
<td>ANN performed better than MDA</td>
</tr>
<tr>
<td>Vomweg et al. (2003)</td>
<td>Breast</td>
<td>604</td>
<td>ANN, expert</td>
<td>ANN outperformed expert</td>
</tr>
</tbody>
</table>

Abbreviations:
- MLC = Maximum likelihood classifiers
- SVM = Support vector machines
- FLD = Fisher linear discriminators
- FP = False positives
- GA = Genetic algorithms
- FN = False negatives
- CART = Classification and regression tree
- MVDA = Multivariate discriminate analysis
### Table (2) A list of clinical trials using AI in artefact detection

<table>
<thead>
<tr>
<th>Reference</th>
<th>Application</th>
<th>Subjects</th>
<th>Methods</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Belal et al 2001)</td>
<td>Oxygen saturation measurements in paediatrics</td>
<td>17</td>
<td>Fuzzy Logic</td>
<td>The system can be implemented in real time monitoring</td>
</tr>
<tr>
<td>(Chen et al 2000)</td>
<td>Gastric Emptying</td>
<td>152</td>
<td>ANN</td>
<td>High specificity and sensitivity achieved</td>
</tr>
<tr>
<td>(Jackson &amp; Sherratt 2004)</td>
<td>EEG Analysis</td>
<td>4 patients and 40 normals</td>
<td>Clustering</td>
<td>Artefact due to eye movement was successfully detected</td>
</tr>
<tr>
<td>(Jung et al. 2000)</td>
<td>EEG Analysis</td>
<td>N/A</td>
<td>ICA, PCA, Regression</td>
<td>ICA was superior in detecting artefact than PCA and regression</td>
</tr>
<tr>
<td>(Kemeny et al. 1999)</td>
<td>Embolus detection in Cerebrovascular diseases</td>
<td>11 patients and 11 normals</td>
<td>MLP</td>
<td>The technique is a promising tool but not in signals with heavy contamination</td>
</tr>
<tr>
<td>(Lin et al. 2003)</td>
<td>Arterial blood pressure measurement</td>
<td>47</td>
<td>Fuzzy logic and recursive regression</td>
<td>This method is more robust than traditional curve fitting algorithms</td>
</tr>
<tr>
<td>(Manke et al 2003)</td>
<td>Respiratory motion</td>
<td>10</td>
<td>PCA</td>
<td>This model was superior over more conventional models</td>
</tr>
<tr>
<td>(McKeown et al. 1998)</td>
<td>fMRI analysis</td>
<td>4</td>
<td>ICA</td>
<td>Highly promising in detecting transients in brain activity</td>
</tr>
<tr>
<td>(Taktak et al 2000)</td>
<td>Sleep monitoring in paediatrics</td>
<td>10</td>
<td>MLP</td>
<td>Intra-subjects errors were reduced but inter-subject errors remained significant</td>
</tr>
</tbody>
</table>

### Table (3) MHRA view on the use of software in medical applications

<table>
<thead>
<tr>
<th>Software which is a Medical Device</th>
<th>Software which is not a Medical Device</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Control or influence the functioning of a medical device</td>
<td>• Administrative handling</td>
</tr>
<tr>
<td>• Use for/by patients to diagnose or treat a physical or mental condition or disease</td>
<td>• Education software</td>
</tr>
<tr>
<td>• Analysis of patient data generated by a medical device with a view to diagnosis and monitoring</td>
<td>• Maintenance of medical devices or components of medical devices.</td>
</tr>
<tr>
<td></td>
<td>• Design and manufacturing processes of the medical device. (e.g. compilers, CM systems, MRP, production control, inventory control, SPC etc.)</td>
</tr>
<tr>
<td></td>
<td>• ‘Operating System’, support or system software</td>
</tr>
</tbody>
</table>
Fig. (3) Clinical Trials that did (black circles) and did not (white circles) show increased clinical benefit in cancer diagnosis using ANN.

Fig. (4) Clinical Trials that did (black circles) and did not (white circles) show increased clinical benefit in artefact detection using ANN and other AI techniques.
References


Manke, D., Nehrke, K., & Bornert, P. 2003, “Novel prospective respiratory motion


