

Reliable diagnosis of acute abdominal pain with conformal prediction

Harris Papadopoulos^{1,2}, Alex Gammerman², Volodya Vovk²

¹Computer Science and Engineering Department, Frederick University, 7 Y. Frederickou St., Palouriotisa, Nicosia 1036, Cyprus. E-mail: h.papadopoulos@frederick.ac.cy

²Computer Learning Research Centre, Royal Holloway, University of London, Egham Hill, Egham, Surrey TW20 0EX, England. E-mail: {alex,vovk}@cs.rhul.ac.uk

Medical decision support is an area in which a lot of machine learning research has been conducted and several diagnostic and prognostic systems have been developed. The majority of these systems only produce bare predictions, without any indication of how reliable each of these predictions is. An indication of this kind however, is highly desirable especially in the medical field. In this paper we address this problem with the use of a recently developed technique, called *conformal prediction*, for accompanying the predictions of traditional machine learning algorithms with measures of their accuracy and reliability. We apply conformal prediction based on a Neural Network classifier to the problem of acute abdominal pain diagnosis and obtain predictions which have a high level of accuracy and are complemented with well-calibrated and practically useful confidence measures.

1. INTRODUCTION

Machine learning techniques have been applied successfully to many medical decision support problems [1, 2] and many good results have been achieved. The resulting systems learn to predict the diagnosis of a new patient based on past history of patients with known diagnoses. Most such systems produce as their prediction only the most likely diagnosis of the new patient, without giving any confidence information in this prediction. This is a major disadvantage, as measures of confidence are of paramount importance in a medical setting [3]. Confidence measures are an indication of how likely each prediction is of being correct. In the ideal case, the percentage of predictions which have a confidence of 99% or higher and are wrongly classified do not exceed 1%; when this is true we say that the confidence measures are well calibrated.

Conformal prediction (CP) [4] is a recently developed technique, which can be used for obtaining confidence measures. Conformal predictors are built on top of traditional machine learning algorithms, called *underlying algorithms*, and complement the predictions of these algorithms with measures of confidence. Different variants of CPs are described in [5, 6, 7, 8, 9, 10, 11, 12]. The results reported in these papers show that not only the confidence values output by CPs are useful in practice, but also their accuracy is comparable to, and sometimes even better than, that of traditional machine learning algorithms.

In this paper we apply CP to the problem of acute abdominal pain diagnosis. This is a relatively popular problem in medical decision support due to the poor discrimination between the diseases that cause acute abdominal pain, which results in high diagnostic error rates [13]. Wrong diagnoses usu-

ally result in unnecessary emergency abdominal operations, which place the patients in unwarranted high risk, discomfort and subsequent loss of working days. In the opposite case, when due to the wrong diagnoses necessary operations are not performed, they result in complications, such as perforation of the appendix, which is even worse. Therefore decreasing the number of wrong diagnoses by the use of a medical decision support tool is an important goal in this area. Although many studies have been performed on this problem, none of the approaches followed in these studies provided any kind of confidence measures in its predictions.

The CP we use is based on Neural Networks (NNs). NNs have not only been successfully applied to many medical problems [2], but they are also one of the most popular machine learning techniques for almost any type of application. Some indicative fields in which NNs have been used with success are image processing, environmental modelling, communications, robotics and the industry; see e.g. [14, 15, 16, 17, 18]. In medicine they have been widely used in a variety of areas such as oncology, cardiology, urology, rheumatology, immunology, and medical imaging; some examples of their use in these areas can be found in [19, 20, 21, 22, 23, 24]. In order to use NNs as the underlying algorithm of a CP, we follow a modified version of the original CP approach called Inductive Conformal Prediction (ICP) [25]. ICP is based on the same general idea as CP but, as its name suggests, it replaces the transductive inference used in the original approach with inductive inference. ICP was first proposed in [7, 8] in an effort to overcome the computational inefficiency problem of CPs. As demonstrated in [25] this computational inefficiency problem renders the original CP approach highly unsuitable for use with NNs; and in general any other method that requires long training times.

The rest of this paper is structured as follows. The next section gives an overview of related work on Conformal Prediction, alternative approaches to obtaining confidence information and acute abdominal pain diagnosis. This is followed by an analysis of the data used in this study in section 3. In section 4 we summarise the general idea behind CP and its inductive version ICP, while in section 5 we detail the Neural Network ICP method. Section 6 describes our experiments and lists and discusses their results. Finally, section 7 gives our conclusions and the future directions of this work.

2. RELATED WORK

This section gives a synopsis of the work carried out on Conformal Prediction since it was first proposed, mentions the other machine learning approaches that can be used for obtaining confidence information and discusses their drawbacks, and summarises related work on acute abdominal pain diagnosis.

2.1 Conformal Prediction

Conformal Prediction was initially proposed in [26] and later greatly improved in [11] and [27]. In these papers CP was applied to support vector machines for classification. Soon it started being applied to other popular classification algo-

rithms, such as k -nearest neighbours [10], and decision trees. In the case of regression algorithms, where its application becomes more complicated, an initial attempt to apply it to ridge regression was made in [28], while soon after a much better version was introduced in [5]. Slightly later it was also applied to k -nearest neighbours for regression [6].

At the same time work was being carried out on improving the computational efficiency of CP, which was its main drawback. This was a result of its transductive nature, since for every test example all computations had to start from scratch. At first some ways of improving the efficiency of transductive CP were studied, such as “competitive transduction” and “transduction with hashing” [12]. Then a much more radical modification was made, by replacing the transductive inference of the original approach with inductive inference. This modified version of CP, called Inductive Conformal Prediction (ICP), was first introduced in [7], where it was applied to ridge regression, and in [8], where it was applied to k -nearest neighbours. This is also the approach we follow in this paper in conjunction with Neural Networks [9].

In terms of applications, to date CPs have been applied successfully to a variety of problems, such as the early detection of ovarian cancer [29], the classification of leukaemia subtypes [30], the prediction of plant promoters [31], the recognition of hypoxia electroencephalograms (EEGs) [32], the prediction of network traffic demand [33] and the estimation of effort for software projects [34].

2.2 Alternative Approaches to Obtaining Confidence Information

There are two other machine learning theories that can be used for obtaining some kind of confidence information. One can apply the theory of Probably Approximately Correct learning (PAC theory) [35] to an algorithm in order to obtain upper bounds on the probability of its error with respect to some confidence level. On the other hand, the Bayesian framework can be used for producing algorithms that complement individual predictions with probabilistic measures of their quality, such as gaussian processes [36]. Both these approaches however, have important drawbacks.

While there are some PAC methods that are capable of establishing non-trivial bounds that might be interesting in practice, in order for them to do so the dataset should be particularly clean. If this is not the case, which is not for the majority of datasets, the bounds obtained from these methods are very weak and as a result not very useful in practice. A demonstration of the crudeness of PAC bounds can be found in [37]. In addition, PAC theory has two other drawbacks: (a) the majority of relevant results either involve large explicit constants or do not specify the relevant constants at all; (b) the bounds obtained by PAC theory are for the overall error and not for individual predictions.

On the contrary, Bayesian methods can give strong confidence bounds for individual predictions. These confidence bounds however, are based on some a priori assumptions about the distribution generating the data. When these assumptions are violated, which is typically the case since for real world datasets such information is not available, the outputs

of Bayesian methods can become quite misleading; for example the predictive regions output for the 95% confidence level may contain the true label in much less than 95% of the cases. This signifies a major failure, as we would expect confidence levels to bound the percentage of expected errors. A comparison of the Bayesian framework with CP and an experimental demonstration of this negative aspect of Bayesian methods can be found in [38].

2.3 Acute Abdominal Pain Diagnosis

The application of machine learning methods to the problem of acute abdominal pain (AAP) diagnosis has been the subject of quite a few studies. In [39] two Bayesian methods (Naive and Proper Bayes) were applied to a relatively large dataset consisting of 6387 patients; this is the same dataset used in this paper. The results of the two Bayesian methods were compared with those of the classification tree algorithm CART and the preliminary diagnoses of hospital physicians. The Naive Bayes classifier gave 74% correct diagnoses outperforming all other techniques and coming relatively close to the 76% correct diagnoses of the hospital physicians.

Ohmann et al. [40] evaluated the performance of 7 machine learning techniques on a clinical database of AAP patients. Similarly Blazadonakis et al. [41] assessed the performance of 6 machine learning techniques on the problem of diagnosing acute abdominal pain in children. In both these studies the Naive Bayes classifier was shown to be superior to all other techniques used.

Pesonen et al. [42] compared the performance of 4 neural network algorithms on a dataset of AAP patients consisting of 911 cases. The experiments performed included evaluation of the algorithms on different groups of parameters as inputs. Overall the two best performing algorithms were backpropagation and learning vector quantization. In [43] the same authors examined the use of 4 substitution methods for the replacement of missing data values on the same problem. The full dataset was used in this study consisting of 1333 patients. In their experiments the difference in the results obtained when using each of the 4 methods was very small.

Mantzaris et al. [44] studied the application of backpropagation and probabilistic neural networks to a dataset of children with AAP. The experimental results showed that the backpropagation neural networks had a very satisfactory performance which was better than that of the probabilistic neural networks. The same dataset was also used by Anastassopoulos and Iliadis [45] who evaluated the performance of various neural network architectures using different learning rules, transfer functions and optimisation algorithms.

Kuo et al. [46] examined the application of a fuzzy-neural system designed through symbiotic evolution to a dataset of male AAP patients. The performance of the approach described in the paper was compared to that of 3 other methods and performed quite well.

Zorman et al. [13] studied the use of decision trees for separating acute appendicitis from other diseases that cause acute abdominal pain on three datasets consisting of 1254, 2286, and 4020 patients respectively. Tsymbal et al. [47] used the same three datasets to explore the use of 4 different search strategies for ensemble feature selection, two of which were

proposed in the same study. The ensembles used consisted of simple Bayesian classifiers. The experimental results showed that the best search strategy was one of the two proposed in the study, called Ensemble Forward Sequential Selection.

3. ACUTE ABDOMINAL PAIN DATA

The acute abdominal pain database used in this study was originally used in [39], where a more detailed description of the data can be found. The data consist of 6387 records of patients who were admitted to hospital suffering from acute abdominal pain. During the examination of each patient 33 symptoms were recorded, each of which had a number of different discrete values. For example, one of the symptoms is “Progress of Pain” which has the possible values: “Getting Better”, “No Change”, “Getting Worse”. In total there are 135 values describing the 33 symptoms. These values compose the attribute vector for each patient in the form of 135 binary attributes that indicate the absence (0) or presence (1) of the corresponding value. It is worth to mention that there are symptoms which have more than one value or no value at all in many of the records. A list of the 33 symptoms along with the possible values of each one is given in the Appendix.

There are nine diseases or diagnostic groups in which the patients were allocated according to all information after their initial examination, including the results of surgical operations. These are: Appendicitis (APP), Diverticulitis (DIV), Perforated Peptic Ulcer (PPU), Non-specific Abdominal Pain (NAP), Cholecystitis (CHO), Intestinal Obstruction (INO), Pancreatitis (PAN), Renal Colic (RCO) and Dyspepsia (DYS). NAP is not actually a diagnostic group, it is a residual group in which all patients that did not belong to one of the other groups were placed. It is also worth to mention that PAN is a poorly characterised group as the relevant symptoms for pancreatitis, such as a blood test for levels of alcohol, were not recorded in the database.

The data are divided into a training set consisting of 4387 examples and a test set consisting of 2000 examples. These are the same training and test sets as in [39]. Table 1 reports the number of examples that belong to each diagnostic group.

4. CONFORMAL PREDICTION

In this section we give a brief description of the idea behind CP, for more details see [4]. We are given a training set $\{z_1, \dots, z_l\}$ of examples, where each $z_i \in Z$ is a pair (x_i, y_i) ; $x_i \in \mathbb{R}^d$ is the vector of attributes for example i and $y_i \in \{Y_1, \dots, Y_c\}$ is the classification of that example. We are also given a new unclassified example x_{l+1} and our task is to state something about our confidence in each possible classification Y_1, \dots, Y_c of x_{l+1} . Our only assumption in doing this is the general i.i.d. model (z_1, z_2, \dots are independent and identically distributed).

CP is based on assigning each one of the possible labels $Y_j \in Y_1, \dots, Y_c$ to the new example x_{l+1} one by one and measuring how likely it is for the extended set of examples

$$\{(x_1, y_1), \dots, (x_l, y_l), (x_{l+1}, Y_j)\}, \quad (1)$$

Table 1 Data distribution.

	APP	DIV	PPU	NAP	CHO	INO	PAN	RCO	DYS	Total
Training Set	585	108	88	1941	372	290	65	326	612	4387
Test Set	259	35	42	894	200	127	31	147	265	2000
Total	844	143	130	2835	572	417	96	473	877	6387

to have been generated independently from the same probability distribution. This as a result corresponds to the likelihood of Y_j being the true label of x_{l+1} as this is the only part in (1) for which we are not sure.

First we measure how strange, or non-conforming, each example in (1) is for the rest of the examples in the same set. We use what is called a *non-conformity measure* which is based on a traditional machine learning algorithm, called the *underlying algorithm* of the CP. This measure assigns a numerical score $\alpha_i^{(Y_j)}$ to each example (x_i, y_i) indicating how different it is from all other examples in (1). In effect we train the underlying algorithm using (1) as training set to generate a prediction rule

$$D_{\{(x_1, y_1), \dots, (x_l, y_l), (x_{l+1}, Y_j)\}}, \quad (2)$$

which maps any unclassified example x to a predicted label \hat{y} and for each pair (x_i, y_i) in (1) we measure the disagreement between the prediction

$$\hat{y}_i = D_{\{(x_1, y_1), \dots, (x_l, y_l), (x_{l+1}, Y_j)\}}(x_i)$$

and the actual label y_i ; in the case of x_{l+1} we use the assumed label Y_j in the place of y_{l+1} . Alternatively, we create the prediction rule

$$D_{\{(x_1, y_1), \dots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \dots, (x_l, y_l), (x_{l+1}, Y_j)\}}, \quad (3)$$

using all the the examples in (1) except (x_i, y_i) and measure the degree of disagreement between

$$\hat{y}_i = D_{\{(x_1, y_1), \dots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \dots, (x_l, y_l), (x_{l+1}, Y_j)\}}(x_i)$$

and y_i . This degree of disagreement, which is measured in a different way in each CP, is called the nonconformity score of example (x_i, y_i) for the assumed label Y_j and is denoted as $\alpha_i^{(Y_j)}$.

The non-conformity score $\alpha_{l+1}^{(Y_j)}$ of (x_{l+1}, Y_j) on its own does not really give us any information, it is just a numeric value. However, we can find out how unusual (x_{l+1}, Y_j) is according to our non-conformity measure by comparing $\alpha_{l+1}^{(Y_j)}$ with all other non-conformity scores. This comparison can be performed with the function

$$p((x_1, y_1), \dots, (x_l, y_l), (x_{l+1}, Y_j)) = \frac{\#\{i = 1, \dots, l+1 : \alpha_i^{(Y_j)} \geq \alpha_{l+1}^{(Y_j)}\}}{l+1}. \quad (4)$$

We call the output of this function, which lies between $\frac{1}{l+1}$ and 1, the p-value of Y_j , also denoted as $p(Y_j)$. An important property of (4) is that $\forall \delta \in [0, 1]$ and for all probability

distributions P on Z ,

$$P^{l+1}\{((x_1, y_1), \dots, (x_{l+1}, y_{l+1})) : p(y_{l+1}) \leq \delta\} \leq \delta; \quad (5)$$

for a proof see [37]. As a result, if the p-value of a given label is under some very low threshold, say 0.05, this would mean that this label is highly unlikely as such sequences will only be generated at most 5% of the time by any i.i.d. process.

After calculating the p-value of every possible label Y_j , as described above, we are able to exclude all labels that have a p-value under some very low threshold (or *significance level*) δ and have at most δ chance of being wrong. Consequently, given a confidence level $1 - \delta$ a CP outputs the set

$$\{Y_j : p(Y_j) > \delta\}. \quad (6)$$

If we prefer to obtain single predictions rather than prediction sets the CP can predict the most likely classification together with a confidence and a credibility measure in this prediction. In this case it predicts the classification with the largest p-value, outputs one minus the second largest p-value as confidence to this prediction and as credibility it outputs the p-value of the predicted classification, i.e. the largest p-value. Confidence tells us how likely the predicted classification is compared to all other possible classifications, while credibility tells us how suitable the CP is for classifying the given example; in effect only very low credibility values are important, which indicate that the example does not really seem to belong to any of the possible classifications.

In order to demonstrate the calculation of the two kinds of outputs that can be produced by a CP and the different cases of these outputs, consider the three selected test cases given in table 2. The table reports the true label and the p-values calculated for each example by the CP for every one of the 9 diagnostic groups of our problem. In the case of the first example at the confidence level of 95% the prediction set of the CP will contain two labels {APP, NAP}, since the p-values of these two labels are above the corresponding significance level of 5%. If we now decrease our confidence level to 90% the prediction set will contain only one label {APP}, since no other p-value exceeds 10%. On the contrary if we increase our confidence level to 99% the prediction set widens and it now contains all the possible labels.

For the second example all p-values are relatively high, which is actually typical for misclassified examples. In this case, at all confidence levels above 90% the prediction set of the CP will contain all the possible labels. If we reduce the required confidence to less than 85% the prediction set will be narrowed down to the two most likely labels {PPU, DYS}. The third example is a much more clear cut case. All but one

Table 2 The p-values produced by the Neural Networks Inductive Conformal Predictor for three selected test examples along with their true labels.

True Label	P-values								
	APP	DIV	PPU	NAP	CHO	INO	PAN	RCO	DYS
APP	60.66%	1.67%	1.67%	6.33%	1.67%	1.67%	1.67%	1.67%	1.67%
PPU	12.00%	12.00%	25.33%	12.00%	14.00%	12.00%	14.00%	11.67%	26.00%
RCO	0.33%	0.33%	0.33%	0.33%	0.33%	0.33%	0.33%	91.67%	0.33%

of its p-values are below 1% and therefore even at the 99% confidence level the prediction set of the CP will only consist of the true label {RCO}.

In terms of the single label predictions, for the first example the CP will predict the true label, as it has the highest p-value, with a confidence of 93.67% (1 minus 0.0633, which is the second largest p-value) and a credibility of 60.66%. The second example will be misclassified as DYS with a confidence of 74.67% and a credibility of 26%, which is an indication of the example being strange. The third example will be correctly classified as RCO with a confidence of 99.67% and a credibility of 91.67%.

4.1 Inductive Conformal Prediction

The original CP technique requires training the underlying algorithm once for each possible classification of every new test example. This means that if our problem has 9 possible classifications and we have to classify 2000 test examples, as is the case in this study, the training process will be repeated $9 \times 2000 = 18000$ times. This makes it very computationally inefficient especially for algorithms that require long training times such as Neural Networks.

Inductive Conformal Predictors (ICPs) [25] are based on the same general idea described above, but follow a different approach which allows them to train their underlying algorithm just once. This is achieved by splitting the training set (of size l) into two smaller sets, the *proper training set* with $m < l$ examples and the *calibration set* with $q := l - m$ examples. The proper training set is used for training the underlying algorithm to generate the prediction rule

$$D_{\{(x_1, y_1), \dots, (x_m, y_m)\}}, \quad (7)$$

and only the examples in the calibration set are used for calculating the p-value of each possible classification of the new test example. More specifically, we calculate the nonconformity score of each pair (x_{m+i}, y_{m+i}) , $i = 1, \dots, q$ in the calibration set as the degree of disagreement between the prediction

$$\hat{y}_{m+i} = D_{\{(x_1, y_1), \dots, (x_m, y_m)\}}(x_{m+i})$$

and y_{m+i} . For any new test example x_{l+g} we can then assume each possible classification $Y_j \in \{Y_1, \dots, Y_c\}$ and calculate the nonconformity score $\alpha_{l+g}^{(Y_j)}$ of the pair (x_{l+g}, Y_j) as the degree of disagreement between

$$L + g = D_{\{(x_1, y_1), \dots, (x_m, y_m)\}}(x_{l+g})$$

and Y_j . After calculating the nonconformity score for each possible classification Y_j we calculate its p-value as

$$p(Y_j) = \frac{\#\{i = m + 1, \dots, m + q, l + g : \alpha_i \geq \alpha_{l+g}^{(Y_j)}\}}{q + 1}. \quad (8)$$

Notice that now the training of the underlining algorithm has to be performed only once.

The number of training examples q that are allocated to the calibration set should take up a small portion of the training set, so that their removal will not dramatically reduce the predictive ability of the prediction rule (7). As we are typically interested in the confidence levels of 99% and 95%, the calibration sizes we use are of the form $q = 100n - 1$, where $n \in \mathbb{N}$; so that the denominator of (8) will be $100n$.

5. NEURAL NETWORKS INDUCTIVE CONFORMAL PREDICTOR

In this section we analyse the Neural Networks ICP (NN-ICP) algorithm [9]. We first describe the typical output encoding for Neural Networks (NNs) and then, based on this description, we define two non-conformity measures for NNs. Finally, we detail the complete NN-ICP algorithm.

5.1 Non-conformity Measures

Typically the output layer of a classification NN consists of c units, each representing one of the c possible classifications of the problem at hand; thus each label is encoded into c target outputs. To explicitly describe this encoding consider the label, $y_i = Y_u$ of a training example i , where $Y_u \in \{Y_1, \dots, Y_c\}$. The resulting target outputs for y_i will be

$$t_1^i, \dots, t_c^i,$$

where

$$t_j^i = \begin{cases} 1, & \text{if } j = u, \\ 0, & \text{otherwise,} \end{cases}$$

for $j = 1, 2, \dots, c$. In the same manner we will denote the actual outputs of the NN for an example i as

$$o_1^i, \dots, o_c^i.$$

According to this encoding the higher the output o_u^i (which corresponds to the example's true classification) the more conforming the example, and the higher the other outputs the less conforming the example. In fact, the most important of all other outputs is the one with the maximum value

$\max_{j=1,\dots,c:j\neq u} o_j^i$, since that is the one which might be very near or even higher than o_u^i . So a natural non-conformity measure for an example $z_i = (x_i, y_i)$ where $y_i = Y_u$ would be defined as

$$\alpha_i = \max_{j=1,\dots,c:j\neq u} o_j^i - o_u^i, \quad (9)$$

or as

$$\alpha_i = \frac{\max_{j=1,\dots,c:j\neq u} o_j^i}{o_u^i + \gamma}, \quad (10)$$

where the parameter $\gamma \geq 0$ in the second definition enables us to adjust the sensitivity of our measure to small changes of o_u^i depending on the data in question. We added this parameter in order to gain control over which category of outputs will be more important in determining the resulting non-conformity scores; by increasing γ one reduces the importance of o_u^i and consequently increases the importance of all other outputs.

5.2 The Algorithm

We can now use the non-conformity measure (9) or (10) to compute the non-conformity score of each example in the calibration set and each test set pair (x_{l+g}, Y_u) . These can then be fed into the p-value function (8), giving us the p-value for each classification Y_u . The exact steps the Neural Networks ICP follows for a training set $\{z_1, \dots, z_l\}$ and a test set $\{x_{l+1}, \dots, x_{l+r}\}$ are:

- Split the training set into the *proper training set* with $m < l$ examples and the *calibration set* with $q := l - m$ examples.
- Use the proper training set to train the Neural Network.
- For each example $z_{m+t} = (x_{m+t}, y_{m+t})$, $t = 1, \dots, q$ in the calibration set:
 - supply the input pattern x_{m+t} to the trained network to obtain the output values $o_1^{m+t}, \dots, o_c^{m+t}$ and
 - calculate the non-conformity score α_{m+t} of the pair (x_{m+t}, y_{m+t}) by applying (9) or (10) to these values.
- For each test pattern x_{l+g} , $g = 1, \dots, r$:
 - supply the input pattern x_{l+g} to the trained network to obtain the output values $o_1^{l+g}, \dots, o_c^{l+g}$,
 - consider each possible classification Y_u , $u = 1, \dots, c$ and:
 - * compute the non-conformity score $\alpha_{l+g} = \alpha_{l+g}^{(Y_u)}$ of the pair (x_{l+g}, Y_u) by applying (9) or (10) to the outputs of the network,
 - * calculate the p-value $p(Y_u)$ of the pair (x_{l+g}, Y_u) by applying (8) to the non-conformity scores of the calibration examples and $\alpha_{l+g}^{(Y_u)}$:

$$p(Y_u) = \frac{\#\{i = m+1, \dots, m+q, l+g : \alpha_i \geq \alpha_{l+g}^{(Y_u)}\}}{q+1},$$

- predict the classification with the largest p-value (in case of a tie choose the one with the smallest non-conformity score) and output one minus the second largest p-value as confidence to this prediction and the p-value of the output classification as its credibility,
- or given a confidence level $1 - \delta$ output the prediction set (6).

6. EXPERIMENTS AND RESULTS

The NN used in our experiments was a 2-layer fully connected feed-forward network, with sigmoid hidden units and softmax output units. It consisted of 135 input, 35 hidden and 9 output units. The number of hidden units was selected by following a cross validation scheme on the training set and trying out the values: 20, 25, 30, 35, 40, 45, 50, 55, 60. More specifically, the training set was split into five parts of almost equal size and five sets of experiments were performed, each time using one of these parts for evaluating the NNs trained on the examples in the other four parts. For each of the five test parts, a further 10-fold cross validation process was performed to divide the examples into training and validation sets, so as to use the validation examples for determining when to stop the training process. Training was performed with the backpropagation algorithm minimizing a cross-entropy loss function.

The results reported here were obtained by following a 10-fold cross validation procedure on the training set in order to divide it into training and validation examples. To create the calibration set of the ICP, 299 examples were removed from the training set before generating the 10 splits. This experiment was repeated 10 times with random permutations of the training examples. Here we report the mean values of all 100 runs.

Table 3 reports the accuracy of the NN-ICP and original NN methods and compares them to that of the Simple Bayes, Proper Bayes and CART methods as reported in [39]. Additionally it compares them to the accuracy of the preliminary diagnoses of the hospital physicians, also reported in [39]. Both the original NN and NN-ICP outperform the other three methods and are almost as accurate as the hospital physicians. As was expected the original NN performs slightly better than the ICP due to the removal of the calibration examples from the training set, however the difference between the two is negligible. This is a very small price to pay considering the advantage of obtaining a confidence measure for each prediction.

Table 4 lists the results of the NN-ICP when producing set predictions for the 99%, 95%, 90% and 80% confidence levels. More specifically it reports the percentage of examples for which the set output by the ICP consisted of only one label, two labels, more than two labels or was empty. It also reports in the last column the percentage of errors made by the ICP, i.e. the percentage of sets that did not include the true classification of the example. A very important thing to notice in this table is that the percentage of errors made by the ICP is always below the corresponding significance level (1 minus the required confidence level). This demonstrates empirically that the confidence measures produced by ICP are well-calibrated and highly reliable.

Table 3 Predictive Accuracy of NN-ICP Compared to Other Methods.

Method	Correct Diagnoses (%)
Neural Networks ICP	75.74
Original Neural Networks	75.87
Simple Bayes	74
Proper Bayes	65
Classification Tree (CART)	65
Physicians (preliminary diagnoses)	76

Table 4 NN-ICP Set Prediction Results.

NC Measure	Confidence Level	Only one label (%)	Two labels (%)	More than two labels (%)	No label (%)	Errors (%)
(9)	99%	23.76	6.52	69.71	0.00	0.95
	95%	46.62	12.25	41.14	0.00	4.10
	90%	62.38	16.87	20.76	0.00	8.59
	80%	82.22	13.04	4.75	0.00	16.94
(10)	99%	25.80	1.55	72.66	0.00	0.95
	95%	47.58	2.42	50.00	0.00	3.75
	90%	65.32	2.86	31.82	0.00	8.11
	80%	87.32	2.17	10.22	0.30	17.23

The percentage of examples for which more than one label is included in the set output by the ICP for the different confidence levels reflects the difficulty in discriminating between the 9 diseases. Nevertheless, the set predictions output by the NN-ICP can be very useful in practice since they pinpoint the cases where more attention must be given and the diagnostic groups that should be considered for each one. Bearing in mind the difficulty of the task and the 76% accuracy of the preliminary diagnoses of physicians, achieving a 95% of accuracy by considering more than one possible diagnosis for only about half the patients is arguably a good result.

Tables 5 and 6 analyse further the set prediction results produced by NN-ICP. They focus on the set predictions produced by the two non-conformity measures for the 90% confidence level and list the percentage of the examples belonging to each diagnostic group which were assigned a prediction set of each size. The percentages reported here show that some diagnostic groups are much more difficult to distinguish than others. Namely the groups DIV, PPU and PAN have the lowest percentage of 1 label sets and the highest percentage of 9 label sets. One reason for this might be that they have the smallest number of training examples, while the other reason is their difficulty to be diagnosed by the recorded symptoms. One indication for this is that although DIV has more training examples than the other two groups, it has the smallest percentage of examples with 1 label and the largest percentage with 9 labels. Furthermore, the lack relevant symptoms for pancreatitis (PAN) was already known.

Finally, tables 7 and 8 show which diagnostic groups are confused with which other groups. These tables give the percentage of p-values of the examples belonging to each diagnostic group which exceeded the 0.2 significance level (cor-

responding to the 80% confidence level) produced with non-conformity measures (9) and (10) respectively. More specifically, each row of the tables is related to the examples belonging to a diagnostic group and each of its cells gives the percentage of these examples which had a p-value higher than 0.2 for the label of the corresponding column. For example, the DIV row and PPU column cell in table 7 means that 11% of the examples with DIV as their true diagnosis had a p-value for the PPU diagnosis that was above 0.2 and therefore for 11% of the DIV examples PPU was included in the set produced by the NN-ICP at the 80% confidence level. The diagonals correspond to the percentage of p-values for the correct diagnostic group that were above 0.2, and therefore they should be relatively high. The off diagonals correspond to wrong diagnostic groups and consequently if they are high, this means that the examples belonging to the diagnostic group of that row can easily be misdiagnosed with the diagnosis of the corresponding column. By examining these tables we observe that the NAP column of both has many relatively high values meaning that the examples of those diagnostic groups (rows) can be easily diagnosed as NAP cases. This is partly due to the fact that NAP has by far the biggest number of examples in the training set, and partly due to the much more diverse nature of this group, since being a residual group it includes many different types of cases. Another observation one can make from these tables is that from the examples belonging to the PAN group only a very small percentage (18% and 29%) has a p-value above 0.2 for the correct group, while a much larger percentage (70% and 72%) of these examples has p-values above 0.2 for the DYS diagnosis. This can definitely be attributed to the inadequacy of the recorded symptoms for the diagnosis of PAN.

Table 5 NN-ICP Set Prediction Results by Diagnostic Group with Non-conformity Measure (9) for a Confidence Level of 90%.

Diagnostic Group	Total Test Examples	Percentage of Examples with Set Predictions of each Size (%)								
		1	2	3	4	5	6	7	8	9
APP	259	47.5	37.3	1.6	0.7	0.4	0.2	0.2	0.1	12.0
DIV	35	9.4	13.3	11.0	3.7	1.2	0.7	0.7	0.5	59.5
PPU	42	17.4	12.7	4.0	5.9	4.1	1.5	1.2	1.4	51.7
NAP	894	72.2	15.6	2.0	0.7	0.3	0.2	0.1	0.1	8.8
CHO	200	69.9	8.0	2.5	1.4	0.8	0.6	0.3	0.4	16.2
INO	127	45.0	12.1	6.1	2.4	1.0	0.8	0.6	0.5	31.6
PAN	31	31.9	10.9	4.5	2.0	1.6	1.6	0.8	0.6	46.0
RCO	147	68.0	13.2	3.0	1.0	0.5	0.2	0.2	0.2	13.7
DYS	265	61.2	14.1	3.3	1.4	1.0	0.6	0.5	0.3	17.5

Table 6 NN-ICP Set Prediction Results by Diagnostic Group with Non-conformity Measure (10) for a Confidence Level of 90%.

Diagnostic Group	Total Test Examples	Percentage of Examples with Set Predictions of each Size (%)								
		1	2	3	4	5	6	7	8	9
APP	259	56.3	6.5	0.2	0.1	0.1	0.0	0.0	0.0	36.8
DIV	35	10.1	1.4	0.9	0.3	0.1	0.1	0.0	0.1	87.2
PPU	42	12.5	1.5	0.3	0.4	0.3	0.1	0.1	0.0	84.7
NAP	894	75.6	2.7	0.3	0.1	0.0	0.0	0.0	0.0	21.3
CHO	200	69.9	1.5	0.4	0.2	0.1	0.1	0.0	0.0	27.8
INO	127	45.2	1.9	0.8	0.3	0.1	0.1	0.0	0.0	51.5
PAN	31	31.1	1.5	0.6	0.3	0.1	0.1	0.1	0.1	66.1
RCO	147	70.5	2.3	0.4	0.1	0.1	0.0	0.0	0.0	26.5
DYS	265	62.5	2.3	0.4	0.2	0.1	0.1	0.0	0.0	34.3

Table 7 Percentage of Examples of each Diagnostic Group with P-values Above 0.2 Produced with Non-conformity Measure (9).

Diagnostic Group	Percentage of P-values Above 0.2 (%)								
	APP	DIV	PPU	NAP	CHO	INO	PAN	RCO	DYS
APP	75	1	3	42	1	3	1	1	4
DIV	11	51	11	75	8	36	4	3	9
PPU	19	7	75	17	18	16	11	9	30
NAP	10	2	1	91	3	3	0	3	5
CHO	3	2	3	10	82	9	3	3	18
INO	7	12	5	33	8	79	3	3	11
PAN	11	8	23	16	42	22	18	13	70
RCO	3	4	1	26	3	3	1	81	4
DYS	1	1	3	18	13	6	2	2	81

7. CONCLUSIONS AND FUTURE WORK

We have presented the application of a recently developed technique, called Conformal Prediction, to the problem of acute abdominal pain diagnosis. Unlike most conventional algorithms, our approach produces confidence measures in its predictions which are provably valid under the general i.i.d. assumption. Our experiments demonstrate that the Neural Networks ICP is very successful at this very difficult task, since its predictions are almost as accurate as the preliminary diagnoses of hospital physicians and its confidence mea-

asures are well calibrated and practically useful. The set predictions produced by NN-ICP identify the cases that require more attention as well as the most likely diagnoses of these cases.

One undesirable aspect of the data used in this study is the huge difference in the number of examples that belong to each class. For this reason, in the future we plan to repeat our experiments with an artificially balanced version of the training set created by performing random resampling of the training examples. Additionally, our directions for future research include further experimentation with other datasets

Table 8 Percentage of Examples of each Diagnostic Group with P-values Above 0.2 Produced with Non-conformity Measure (10).

Diagnostic Group	Percentage of P-values Above 0.2 (%)								
	APP	DIV	PPU	NAP	CHO	INO	PAN	RCO	DYS
APP	72	2	3	37	2	4	2	2	5
DIV	26	54	25	79	24	46	22	21	26
PPU	31	22	78	29	30	27	26	24	41
NAP	10	3	2	90	4	4	2	4	5
CHO	7	7	8	13	83	12	8	7	20
INO	16	19	13	36	16	80	13	13	19
PAN	24	24	33	31	49	30	29	30	72
RCO	6	6	5	25	7	5	5	81	6
DYS	6	5	7	19	15	9	6	6	82

for acute abdominal pain and with more non-conformity measures based on other popular algorithms such as support vector machines, decision trees and evolutionary techniques.

Acknowledgements

This work was supported by the Cyprus Research Promotion Foundation through research contract PLHRO/0506/22 (“Development of New Conformal Prediction Methods with Applications in Medical Diagnosis”).

REFERENCES

- Kononenko I. Machine Learning for Medical Diagnosis: History, State of the Art and Perspective. *Artificial Intelligence in Medicine*. 2001;23(1):89–109.
- Lisboa PJG. A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Networks*. 2002;15(1):11–39.
- Holst H, Ohlsson M, Peterson C, Edenbrandt L. Intelligent computer reporting ‘lack of experience’: a confidence measure for decision support systems. *Clinical Physiology*. 1998;18(2):139–147.
- Vovk V, Gammerman A, Shafer G. *Algorithmic Learning in a Random World*. New York: Springer; 2005.
- Nouretdinov I, Melluish T, Vovk V. Ridge Regression Confidence Machine. In: *Proceedings of the 18th International Conference on Machine Learning (ICML’01)*. San Francisco, CA: Morgan Kaufmann; 2001. p. 385–392.
- Papadopoulos H, Gammerman A, Vovk V. Normalized Non-conformity Measures for Regression Conformal Prediction. In: *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*. ACTA Press; 2008. p. 64–69.
- Papadopoulos H, Proedrou K, Vovk V, Gammerman A. Inductive Confidence Machines for Regression. In: *Proceedings of the 13th European Conference on Machine Learning (ECML’02)*. vol. 2430 of Lecture Notes in Computer Science. Springer; 2002. p. 345–356.
- Papadopoulos H, Vovk V, Gammerman A. Qualified Predictions for Large Data Sets in the Case of Pattern Recognition. In: *Proceedings of the 2002 International Conference on Machine Learning and Applications (ICMLA’02)*. CSREA Press; 2002. p. 159–163.
- Papadopoulos H, Vovk V, Gammerman A. Conformal Prediction with Neural Networks. In: *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI’07)*. vol. 2. IEEE Computer Society; 2007. p. 388–395.
- Proedrou K, Nouretdinov I, Vovk V, Gammerman A. Transductive Confidence Machines for Pattern Recognition. In: *Proceedings of the 13th European Conference on Machine Learning (ECML’02)*. vol. 2430 of Lecture Notes in Computer Science. Springer; 2002. p. 381–390.
- Saunders C, Gammerman A, Vovk V. Transduction with Confidence and Credibility. In: *Proceedings of the 16th International Joint Conference on Artificial Intelligence*. vol. 2. Los Altos, CA: Morgan Kaufmann; 1999. p. 722–726.
- Saunders C, Gammerman A, Vovk V. Computationally Efficient Transductive Machines. In: *Proceedings of the Eleventh International Conference on Algorithmic Learning Theory (ALT’00)*. vol. 1968 of Lecture Notes in Artificial Intelligence. Berlin: Springer; 2000. p. 325–333.
- Zorman M, Eich HP, Kokol P, Ohmann C. Comparison of Three Databases with a Decision Tree Approach in the Medical Field of Acute Appendicitis. *Studies in Health Technology and Informatics*. 2001;84(2):1414–1418.
- Rowley HA, Baluja S, Kanade T. Neural Network-Based Face Detection. *IEEE Transactions On Pattern Analysis and Machine Intelligence*. 1998;20(1):23–38.
- Iliadis LS, Maris F. An Artificial Neural Network model for mountainous water-resources management: The case of Cyprus mountainous watersheds. *Environmental Modelling and Software*. 2007;22(7):1066–1072.
- Haralambous H, Papadopoulos H. 24-Hour Neural Network Congestion Models for High-Frequency Broadcast Users. *IEEE Transactions on Broadcasting*. 2009;55(1):145–154.
- Yang S, Wang M, Jiao L. Radar Target Recognition using Contourlet Packet Transform and Neural Network Approach. *Signal Processing*. 2009;89(4):394–409.
- Iliadis LS, Spartalis S, Tachos S. Application of fuzzy T-norms towards a new Artificial Neural Networks’ evaluation framework: A case from wood industry. *Information Sciences*. 2008;178(20):3828–3839.
- Christoyianni I, Koutras A, Dermatas E, Kokkinakis G. Computer aided diagnosis of breast cancer in digitized mammograms. *Computerized Medical Imaging and Graphics*. 2002;26(5):309–319.
- Haraldsson H, Edenbrandt L, Ohlsson M. Detecting Acute Myocardial Infarction in the 12-lead ECG Using Hermite Expansions and Neural Networks. *Artificial Intelligence in Medicine*. 2004;32(2):127–136.

21. Anagnostou T, Remzi M, Djavan B. Artificial Neural Networks for Decision-Making in Urologic Oncology. Review in Urology. 2003;5(1):15–21.
22. Mantzaris DH, Anastassopoulos GC, Lymberopoulos DK. Medical Disease Prediction using Artificial Neural Networks. In: Proceedings of the 8th IEEE International Conference on Bioinformatics and Bioengineering (BIBE 2008). IEEE; 2008. p. 1–6.
23. Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, et al. Reliable Prediction of T-cell Epitopes using Neural Networks with Novel Sequence Representations. Protein Sci. 2003;12(5):1007–1017.
24. Pattichis CS, Christodoulou C, Kyriacou E, Pattichis MS. Artificial Neural Networks in Medical Imaging Systems. In: Proceedings of the 1st MEDINF International Conference on Medical Informatics and Engineering; 2003. p. 83–91.
25. Papadopoulos H. Inductive Conformal Prediction: Theory and Application to Neural Networks. In: Fritzsche P, editor. Tools in Artificial Intelligence. Vienna, Austria: I-Tech; 2008. p. 315–330. Available from: <http://intechweb.org/downloadpdf.php?id=5294>.
26. Gammerman A, Vapnik V, Vovk V. Learning by Transduction. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann; 1998. p. 148–156.
27. Vovk V, Gammerman A, Saunders C. Machine-learning applications of algorithmic randomness. In: Proceedings of the 16th International Conference on Machine Learning (ICML'99). San Francisco, CA: Morgan Kaufmann; 1999. p. 444–453.
28. Melluish T, Vovk V, Gammerman A. Transduction for Regression Estimation with Confidence. In: Neural Information Processing Systems (NIPS'99); 1999.
29. Gammerman A, Vovk V, Burford B, Nourtdinov I, Luo Z, Chervonenkis A, et al. Serum Proteomic Abnormality Predating Screen Detection of Ovarian Cancer. The Computer Journal, doi:10.1093/comjnl/bxn021. 2008.
30. Bellotti T, Luo Z, Gammerman A, Delft FWV, Saha V. Qualified predictions for microarray and proteomics pattern diagnostics with confidence machines. International Journal of Neural Systems. 2005;15(4):247–258.
31. Shahmuradov IA, Solovyev VV, Gammerman AJ. Plant Promoter Prediction with Confidence Estimation. Nucleic Acids Research. 2005;33(3):1069–1076.
32. Zhang J, Li G, Hu M, Li J, Luo Z. Recognition of Hypoxia EEG with a Preset Confidence Level Based on EEG Analysis. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN 2008), part of the IEEE World Congress on Computational Intelligence (WCCI 2008). IEEE; 2008. p. 3005–3008.
33. Dashevskiy M, Luo Z. Network Traffic Demand Prediction with Confidence. In: Proceedings of the IEEE Global Telecommunications Conference 2008 (GLOBECOM 2008). IEEE; 2008. p. 1453–1457.
34. Papadopoulos H, Papatheocharous E, Andreou AS. Reliable Confidence Intervals for Software Effort Estimation. In: Proceedings of the 2nd Workshop on Artificial Intelligence Techniques in Software Engineering (AISEW 2009). vol. 475 of CEUR Workshop Proceedings. CEUR-WS.org; 2009. Available from: ceur-ws.org/Vol-475/AISEW2009/22-pp-211-220-208.pdf.
35. Valiant LG. A Theory of the Learnable. Communications of the ACM. 1984;27(11):1134–1142.
36. Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. MIT Press; 2006.
37. Nourtdinov I, Vovk V, Vyugin MV, Gammerman A. Pattern Recognition and Density Estimation under the general i.i.d. assumption. In: Proceedings of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory. vol. 2111 of Lecture Notes in Computer Science. Springer; 2001. p. 337–353.
38. Melluish T, Saunders C, Nourtdinov I, Vovk V. Comparing the Bayes and Typicalness Frameworks. In: Proceedings of the 12th European Conference on Machine Learning (ECML'01). vol. 2167 of Lecture Notes in Computer Science. Springer; 2001. p. 360–371.
39. Gammerman A, Thatcher AR. Bayesian Diagnostic Probabilities without Assuming Independence of Symptoms. Methods of Information in Medicine. 1991;30(1):15–22.
40. Ohmann C, Moustakis V, Yang Q, Lang K. Evaluation of automatic knowledge acquisition techniques in the diagnosis of acute abdominal pain. Artificial Intelligence in Medicine. 1996;8(1):23–36.
41. Blazadonakis M, Moustakis V, Charissis G. Deep assessment of machine learning techniques using patient treatment in acute abdominal pain in children. Artificial Intelligence in Medicine. 1996;8(6):527–542.
42. Pesonen E, Eskelinen M, Juhola M. Comparison of different neural network algorithms in the diagnosis of acute appendicitis. International Journal of Bio-Medical Computing. 1996;40(3):227–233.
43. Pesonen E, Eskelinen M, Juhola M. Treatment of missing data values in a neural network based decision support system for acute abdominal pain. Artificial Intelligence in Medicine. 1998;13(3):139–146.
44. Mantzaris D, Anastassopoulos G, Adamopoulos A, Gardikis S. A Non-Symbolic Implementation of Abdominal Pain Estimation in Childhood. Information Sciences. 2008;178(20):3860–3866.
45. Anastassopoulos GC, Iliadis LS. ANN for Prognosis of Abdominal Pain in Childhood: Use of Fuzzy Modelling for Convergence Estimation. In: Proceedings of the 1st International Workshop on Combinations of Intelligent Methods and Applications; 2008. p. 1–5.
46. Kuo HC, Chang HK, Wang YZ. Symbiotic evolution-based design of fuzzy-neural diagnostic system for common acute abdominal pain. Expert Systems with Applications. 2004;27(3):391–401.
47. Tsymbal A, Cunningham P, Pechenizkiy M, Puuronen S. Search Strategies for Ensemble Feature Selection in Medical Diagnostics. In: Proceedings of the 16th IEEE Symposium on Computer-Based Medical Systems (CBMS'2003), The Mount Sinai School of Medicine. IEEE Computer Society; 2003. p. 124–129.

Appendix

Table 9 lists the 33 symptoms recorded in the dataset used in this study and the possible values of each one.

Table 9 Symptoms Contained in the Abdominal Pain Data.

	Symptom	Values
1	Sex	male, female
2	Age	0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70+
3	Pain-site Onset	right upper quadrant, left upper quadrant, right lower quadrant, left lower quadrant, upper half, lower half, right half, left half, central, general, right loin, left loin, epigastric
4	Pain-site Present	right upper quadrant, left upper quadrant, right lower quadrant, left lower quadrant, upper half, lower half, right half, left half, central, general, right loin, left loin, epigastric, pain settled
5	Aggravating Factors	movement, coughing, inspiration, food, other, nil
6	Relieving Factors	lying still, vomiting, antacids, milk/food, other, nil
7	Progress of Pain	getting better, no change, getting worse
8	Duration of Pain	under 12 hours, 12-24 hours, 24-48 hours, over 48 hours
9	Type of Pain	steady, intermittent, colicky, sharp
10	Severity of Pain	moderate, severe
11	Nausea	nausea present, no nausea
12	Vomiting	present, no vomiting
13	Anorexia	present, normal appetite
14	Indigestion	history of dyspepsia, no history of dyspepsia
15	Jaundice	history jaundice, no history of jaundice
16	Bowel Habit	no change, constipated, diarrhoea, blood, mucous
17	Micturation	normal, frequent, dysuria, haematuria, dark urine
18	Previous Pain	similar pain before, no similar pain before
19	Previous Surgery	yes, none
20	Drugs	being taken, not being taken
21	Mood	normal, distressed, anxious
22	Colour	normal, pale, flushed, jaundiced, cyanosed
23	Abdominal Movements	normal, poor/nil, visible peristalsis
24	Abdominal Scar	present, absent
25	Abdominal Distension	present, absent
26	Site of Tenderness	right upper quadrant, left upper quadrant, right lower quadrant, left lower quadrant, upper half, lower half, right half, left half, central, general, right loin, left loin, epigastric, none
27	Rebound	present, absent
28	Guarding	present, absent
29	Rigidity	present, absent
30	Abdominal Masses	present, absent
31	Murphy's Test	positive, negative
32	Bowel Sounds	normal, decreased, increased
33	Rectal Examination	tender left side, tender right side, generally tender, mass felt, normal

