Ranking Pathology Data in the Absence of a Ground Truth

Jing Qi[1], Girvan Burnside[2], and Frans Coenen[1]

[1] Department of Computer Science,
The University of Liverpool, Liverpool L69 3BX, UK
[2] Department of Biostatistics, Institute of Translational Medicine,
The University of Liverpool, Liverpool L69 3BX, UK

Abstract. Pathology results play a critical role in medical decision making. A particular challenge is the large number of pathology results that doctors are presented with on a daily basis. Some form of pathology result prioritisation is therefore a necessity. However, there is no readily available training data that would support a traditional supervised learning approach. Thus some alternative solutions are needed. There are two approaches presented in this paper, anomaly-based unsupervised pathology prioritisation and proxy ground truth-based supervised pathology prioritisation. Two variations of each were considered. With respect to the first, point and time series based unsupervised anomaly prioritisation; and with respect to the second kNN and RNN proxy ground truth-based supervised prioritisation. To act as a focus, Urea and Electrolytes pathology testing was used. The reported evaluation indicated that the RNN proxy ground truth-based supervised pathology prioritisation method produced the best results.

Keywords: Data Ranking · Time Series · Deep Learning · Pathology Data.

1 Introduction

It has been well documented that, facilitated by advances in IT technology, large quantities of data are produced on a continuous basis. An exemplar application, and the focus for the work presented in this paper, is in the medical domain where large amounts pathology data are produced continually. Some clinicians may have hundreds of pathology results to review on a single shift; a classic *information overload* situation. A potential solution is to adopt the tools and techniques of machine learning to prioritise pathology results. However, a challenge is the absence of ground truth data. Clinicians observe that they “know a priority result when they see one”, and can explain why, however typically there is no resource available to generate appropriate prioritised training data (especially given the current COVID-19 pandemic). This means that traditional, well established, supervised learning techniques are unavailable. This is a very much unexplored domain of application for machine learning.

This paper presents an exploration of two approaches whereby such data can be ranked, or at least categorised. The first approach is founded on the idea of anomaly detection, the second one using supervised learning but with a proxy for the ground truth data. Anomaly, or outlier, detection has a long history within the context of machine learning [27]. One established technique, and that adopted in this paper, is cluster analysis-based outlier detection [6, 12]. Two variations are considered: point-based and time series based. The first assumes all pathology results are independent, and is used as a bench-mark technique with respect to the work presented. The second variation acknowledges that pathology results are typically part of a sequence and/or part of a set of parallel results, and therefore individual pathology results should not be considered in isolation.

The second approach is founded on the observation that although ground truth data is typically not available, information about what happened to patients is available. For example the final destination of patients: Emergency Patient (EP), In-Patient (IP) or Out-Patient (OP). This information can thus be used to construct a proxy ground truth training set from which classification models can be generated. Two variations are considered, a kNN classification model as traditionally used in time series analysis [10, 17] and a Recursive Neural Network (RNN) deep learning model as popularised in more recent work on time series analysis [9].

To act as a focus for the work the domain of Urea and Electrolytes (U&E) pathology testing is considered. The proposed approaches are compared using U&E data provided by Arrowe Park Hospital in Merseyside in the UK.

The remainder of this paper is organised as follows. A review of relevant previous work is presented in Section 2. This is followed by a review of the Urea and Electrolytes pathology testing application domain, used as a focus for the work, in Section 3. The two proposed approaches are considered in Sections 4 and 5. The comparative evaluation of the two approaches is then discussed in Section 6. The paper is concluded in Section 7 with a summary of the main findings and some suggested avenues for future work.

2 Previous Work

Prioritisation is significant with respect to many application domains and fields of study. The most common application domain, and that most frequently referenced in the literature, is the information retrieval domain [20]. For example the ranking of documents as the result of a web search or a document repository search. However, the prioritisation models considered in this paper are not ranking models but classification models. The proposed mechanism are designed to build models to label data according to a a set of class labels C indicating priority, namely $C = \{high, medium, low\}$. The challenge, as noted above, is the absence of training data. This makes the application domain considered in this paper unique. To address this challenge two approaches are considered and compared:

1. Anomaly-based unsupervised prioritisation.
2. Proxy ground truth-based supervised prioritisation.

Anomaly detection is concerned with the detection of points, observations or events within a data collection which do not satisfy the dataset's normal distribution [8]. A common technology, and that adopted with respect to the work presented in this paper, is unsupervised learning. A typical application domain for unsupervised anomaly detection is cyber security, where anomalous network behaviour is considered to be an indicator of an attack. Examples can be found in [1], [11] and [15]. In [1] a k-medoid customized clustering technique was presented for anomaly detection in wireless sensor network to detect misdirection attacks and blackhole attacks. In [11] a network anomaly detection method based on fuzzy clustering was presented. In [15] a mechanism was presented for anomaly detection with respect to traffic patterns in computer networks. In [16] a survey was presented of unsupervised approaches to identify anomalies in system log files for anomalous events detection relevant to cyber security. There are many other applications where unsupervised learning has been applied for anomaly detection. In [18] the use of unsupervised anomaly detection was used to detect the abnormal operation of aircraft and in [2] to detect abnormal behaviour in the financial domain. A range of techniques are available for unsupervised anomaly detection, there has been some interesting recent work using autoencoders [31]. Unsupervised learning has, of course, been more generally applied to pathology data for analysis purposes, see for example [23]. Whatever the case, the broad concepts that feature in the above referenced work underpin the work presented here with respect to anomaly-based unsupervised pathology data prioritisation.

The challenge of creating training data to support supervised learning is well established and has led to the growing research area of *self-supervised learning*. Broadly, self-supervised learning is a means for training computers to do tasks without humans providing labeled data [13]. An alternative, which was adopted with respect to the work presented here, is to identify a proxy for the training data. There has been some work on using proxy data for classification purposes. Examples can be found in [7], [3] and [30]. In [7], in the context of market segmentation, it was observed that ground truth data is often scarce or unavailable; a proxy labeling scheme was proposed for labeling a population according to a postulated set of shopping behaviors. In [3] a proxy data set was created, using a clustering approach, for anomaly detection in enterprise and cloud networks. In [30] machine learning was applied to internet search behaviour as a proxy for human behaviour. In [5], in the context of the domain of Psychophysiology, the authors argue that training data is frequently flawed and that proxy data is more able to produce a quality classification model. They go on to present a review of techniques whereby a proxy ground truth can be created, and conclude that there is no single technique that is sufficient for the accurate generation of ground truth data for classification and suggest a hybrid approach. The above provides support for the second approach presented in this paper, the proxy ground truth-based supervised prioritisation approach.

3 U&E Testing Application Domain

The work presented in this paper is focused on Urea and Electrolytes pathology test data (U&E testing). U&E testing is usually performed to confirm normal kidney function or to exclude a serious imbalance of biochemical salts in the bloodstream. The U&E test data considered in this paper comprised, for each test, measurement of levels of: (i) Sodium (ii) Potassium (iii) Urea (vi) Creatinine and (v) Bicarbonate. The measurement of each is referred to as a "task", thus we have five tasks per test. Thus each U&E test results in five pathology values. Abnormal levels in any of these tasks may indicate that the kidneys are not working properly. However, a one time abnormal result does not necessarily indicate priority. A new task result that is out of range for a patient who has a previous recent history of out of range task results, but the latest result indicates a trend back into the normal range, may not be a priority result either. Conversely, a new task result that is within the normal range for a patient who has a history of normal range task results, but the latest result indicates a trend heading out of the normal range, may be a priority result. Given a new set of pathology values for a U&E test we wish to determine the priority to be associated with this set of values.

The U&E data comprised a set of clinical patient records, $\mathbf{D} = \{P_1, P_2, \ldots\}$. Each record $P_j \in \mathbf{D}$ was of the form:

$$P_j = \langle PatientID, TestDate, Gender, T_{So}, T_{Po}, T_{Ur}, T_{Cr}, T_{Bi}, c\rangle \quad (1)$$

Where: (i) $PatientID$ is the ID for the patient in question; (ii) T_{so} to T_{Bi} are five three dimensional time series, one per task, representing, in sequence, pathology results for: Sodium (So), Potassium (Po), Urea (Ur), Creatinine (Cr) and Bicarbonate (Bi) and (iii) c is the class label taken from a set of classes C. Each time series T_i has three dimensions: (i) pathology result value, (ii) normal low and (iii) normal high. The normal low and high dimensions indicate a "band" in which pathology results are expected to fall. These values are less volatile than the pathology result values, but can change over time.

For the purpose of building prioritisation models training data was required. The data set $\mathbf{D}$ was used to create individual training data sets, one per task, D_{So}, D_{Po}, D_{Ur}, D_{Cr} and D_{Bi}. Two data formats were used, one for the point-based outlier detection method and one for the three time series methods considered. For the first each data set comprised a set of pathology result values $D_i = \{p_1, p_2, \ldots\}$ where each point p_i comprised a tuple of the form $\langle v, n_l, n_h\rangle$ (pathology result value, normal low and normal high respectively). For the other methods each data set comprised a set of time series $D_i = \{T_1, T_2, \ldots\}$ where each time series T_i comprises a sequence of tuples, of the form $\langle v, n_l, n_h\rangle$.

4 Anomaly-based Unsupervised Pathology Prioritisation

The fundamental idea under-pinning the anomaly-based pathology data prioritisation approach is that an anomalous result should be prioritised. More specifically the first approach presented in this paper proposes that this can be achieved by clustering existing records and attempting to assign new records to this cluster configuration. If a new record cannot be easily allocated to a cluster it is considered to be an *outlier* and hence a priority record. This is an approach that has been frequently adopted with respect to cyber security applications [1, 11, 15, 16].

The outlier/anomaly detection approach to pathology data prioritisation produces a binary classification, a new pathology record is either an outlier (a priority record) or not. This is in itself useful, but we would like a finer grained outcome. Thus, for the outliers the distance to the centroid of the nearest cluster is determined to produce a ranking which can be used to produce a more fine grained prioritisation. In the evaluation presented later in this paper outlier records are labelled either as “high priority” or “medium pririty” according to a predefined threshold λ which needs to be established; non-outlier records are labelled a “low priority”. Thus we have a three class prioritisation, $C = \{high, medium, low\}$.

In the context of the U&E test application focus for this paper, as discussed above, we have five tasks. Hence, we have five cluster configurations and consequently five predictions that need to be reconciled, and so five thresholds to be identified λ_{So}, λ_{Po}, λ_{Ur}, λ_{Cr} and λ_{Bi}.

The high level process is as follows, given:

1. For each data set $D_i \in \mathbf{D}$, $D_i = \{p_1, p_2, \ldots\}$, where each point p_i is a tuple of the form, $\langle v, n_l, n_h \rangle$, create a cluster configuration, one per task, hence five configurations.
2. For each configuration, given the set of outliers A, for each $a \in A$ calculate the distances to the nearest centroid. To give a set of distances.
3. For each set of distances calculate the average, these are then the thresholds, λ_{So}, λ_{Po}, λ_{Ur}, λ_{Cr} and λ_{Bi}, that will be used to determine whether an outlier record is high or medium priority.

Given a new pathology record, it will be compared to the cluster configuration, generated as described above, which will produce five class labels, one for each task. The following rule is then applied.

Rule 1 If one of the class labels is “high” the overall class label is high, otherwise use voting to derive the overall class label.

Any one of a number of clustering algorithms could have been adopted. However, for the evaluation presented later in this paper the DBSCAN clustering algorithm [14] was adopted with respect to the work presented in this paper, because it readily supports outlier detection, and because it is a well established and understood clustering algorithm.

Two variations of the anomaly-based pathology data prioritisation approach were considered.

Point Based: Assumes that any new record is independent of any previous records for the same patient and hence can be considered in isolation.

Time Series Based: Assumes that any new record is not independent of previous records for the same patient and hence should be considered in context; in other words as a time series.

Each is discussed in further detail in the following two sub-sections.

4.1 Point Based Outlier Detection Unsupervised Pathology Prioritisation

The point-based approach considers each pathology result, pertaining to the same patient, to be independent. In this case the data set used, to create a desired cluster configuration, simply comprised a set of pathology results obtained from historical patient data.

DBSCAN uses two parameters: (i) $minPts$, the minimum number of data points that can be held in a cluster, and (ii) ε, the maximum distance between two data points whereby they are considered to be neighbours and thus should appear in the same cluster. If $minPts = 1$ is used this will result in every record forming its own cluster; if $minPts = 2$ is used, this will result in a hierarchical clustering as clusters will be repeatedly split into two. Therefore $minPts$ needs to be greater than 2. In [26] it was suggested that the value of $minPts$ should be at least $|A|+1$ (where A is the attribute set). The reasoning was that each attribute represents a dimension in a $|A|$-dimensional space. To determine a value for ε the approach proposed in [21] was adopted. For each record the distance to the kth nearest neighbouring record was determined and plotted using an "elbow plot". A range of values for k was considered starting with $k = minPts - 1$. For each value of k, the input data records were listed in ascending order according to distance. The elbow plot has k plotted along the x-axis and distance along the y-axis. The plot will feature an "elbow" marking a significant change in the gradient of the slope. The most appropriate value for ε is then the distance associated with the point where the elbow first starts to appear.

4.2 Time Series Based Outlier Detection Unsupervised Pathology Prioritisation

The main difference between the time series approach and the point approach is that the time series approach considers patient history. In other words the "trajectory" for the patient in question. The intuition was that this would provide a better prioritisation. As in the case of the point based approach, the idea was to cluster existing trajectories to produce a cluster configuration. Given a new record this will be added to the time series for the corresponding patient (if the record does not belong to an existing patient, the point approach will need to be used) and the resulting time series compared to the cluster configuration. Again, if the time series associated with the new record is found to be an outlier the record is considered to be a prioritiy record.

The DBSCAN clustering algorithm was again adopted. However, whereas the point based approach used Euclidean distance as the distance measure with which to generate a cluster configuration, for the time series based approach Dynamic Time Warping (DTW) was used [25] which gives a distance measure (the *warping distance*) between two time series. For applying DTW the Sakoe-Chiba band, as also proposed in [25], was used as a global constraint to accelerate the algorithm. The same mechanisms for determining the most appropriate values for *minPts* and ε as used with respect to the point based approach described above were adopted, those given in [26] and [21] respectively.

5 Proxy Ground Truth-based Supervised Pathology Prioritisation

The fundamental idea underpinning the proposed proxy ground truth-based supervised pathology prioritisation approach was that although no ground truth training data was available, the final destinations of patients where known, and hence these could act as a proxy for a ground truth. Consequently, supervised time series learning could be used to generate a pathology data prioritisation model. For the evaluation presented later in this paper, three outcome events were considered: (i) Emergency Patient (EP), (ii) In-Patient (IP) and (ii) Out Patient (OP), which were correlated with the priority descriptors "high", "medium" and "low" respectively.

Two variations of the proxy ground truth-based pathology data prioritisation approach were considered.

KNN Based: Uses k Nearest Neighbour (kNN) classification, the most frequently adopted form of time series classification.

RNN Based: Uses Recurrent Neural Network (RNN) classification, a time series classification model that is gaining increasing popularity.

Each is discussed in further detail in the following two sub-sections.

5.1 kNN Proxy Ground Truth-based Supervised Pathology Prioritisation

The kNN classification model uses a parameter k, the number of best matches we are looking for. For the evaluation presented later in this paper, $k = 1$ was used, because $k = 1$ often provides better accuracy when comparing time series using DTW [4]. Note that DTW was used for similarity measurement because of its ability to operates with time series of different length and because it has been shown to be more effective than alternatives such as Euclidean distance measurement [29]. The disadvantage of DTW, compared to the Euclidean distance measurement, is its high computational time complexity of $O(x \times y)$ where x and y are the lengths of the two time series under consideration. The complexity for Euclidean distance time series comparison is $O(x)$ (x is required to be equal to y).

There are many techniques available for reducing the time complexity of DTW coupled with kNN classification. Two that were adopted with respect to the work presented here were: (i) early-abandonment and (ii) lower bounding. The first is a strategy whereby the accumulative distance between two time series is repeatedly checked as the calculation progresses and if the distance exceeds the best distance so far the calculation will be "abandoned" [22]. The second involves pre-processing the time series to be considered by comparing the time series using an alternative "cheaper" technique and pruning those that are unlikely to be close matches and applying DTW to the remainder. One example of this, and that adopted with respect to the work presented in this paper, is the lower bounding technique proposed in [28], the so called the *LB-Keogh technique*. This operates by superimposing a band, defined by a predefined offset value referred to as the lower bound, over each time series in the bank, and calculating the complement of the overlap with the new time series. Where the calculated value exceeds a given threshold ϵ the associated time series is pruned.

The traditional manner in which kNN is applied, in the context of time series analysis, is to compare a query time series with the time series in the kNN bank. In the case of the U&E pathology prioritisation scenario considered here, as noted in Section 3, individual pathology results comprised five values, one per task making up the overall A&E test. The kNN process thus involved five comparisons, once for each task time series in the query record, $T_{q_{so}}$, $T_{q_{po}}$, $T_{q_{ur}}$, $T_{q_{cr}}$ and $T_{q_{bi}}$. In addition, traditional kNN is applied to univariate time series, in the U&E pathology case each task time series was a three-dimensional multi-variate time series: (i) pathology value, (ii) normal low and (iii) normal high. Thus, from the foregoing, for each comparison five distance measures were obtained. These five distance measures therefore need to be combined to give a final prioritisation.

The overall process was as follows, given a new pathology result for a patient p_q, that has been appended to the patient's history of pathology results to give five three-dimensional component time series $T_{q_{so}}$, $T_{q_{po}}$, $T_{q_{ur}}$, $T_{q_{cr}}$ and $T_{q_{bi}}$.

1. Calculate the average LB-keogh overlap for the five component time series and prune all records in D where the overlap for any one time series was greater than ϵ, to leave D'.
2. Apply DTW, with early-abandonment to compare each T_{q_j} with each $T_{i_j} \in D'$, where j indicates the U&E task, and use the class label c associated with the most similar record to assign to each time series T_{q_j}.
3. Use the class label c to define a priority for p_q and then apply Rule 1 from Section 4 to determine the final prioritisation, "high", "medium" or "low".

With respect to the above the choice of the value for ϵ is of great importance as it affects the efficiency and the accuracy of the similarity search. According to [19], there is a threshold value for ϵ whereby the time complexity for the lower bounding is greater than simply using DTW distance without lower bounding. The experiments presented in [19] demonstrated that this threshold occurs when the value for ϵ prunes 90% of the time series in D. For the parameter setting in the work presented in this paper, $\epsilon = 0.159$ was used because, on average, this resulted in 10% of the time series in D being retained.

5.2 RNN Proxy Ground Truth-based Supervised Pathology Prioritisation

For the RNN-based approach a Long Short Term Memory (LSTM) architecture was adopted. Given the U&E pathology prioritisation scenario used as a focus in this paper the proposed approach commenced with the training of five LSTM models, one per task: $LSTM_{so}$, $LSTM_{po}$, $LSTM_{ur}$, $LSTM_{cr}$ and $LSTM_{bi}$. Figure 1 illustrates the construction and structure of the proposed LSTM approach to prioritise pathology data.

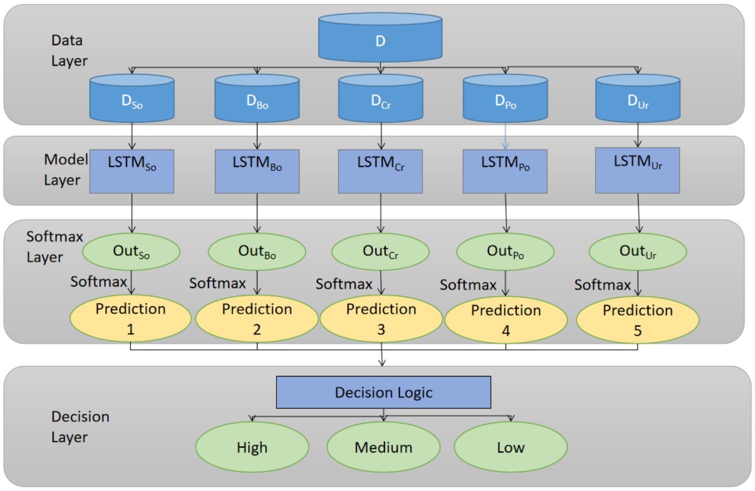

Fig. 1. LSTM architecture for proxy ground truth pathology prioritisation

With reference to Figure 1 the process and structure is expressed in terms of four layers: (i) Data, (ii) Model, (iii) Softmax, (iv) Decision. The input is the data set D divided into of its component parts D_{So}, D_{Po}, D_{Ur}, D_{Cr} and D_{Bi}. Each data set comprised a set of multi-variate time series $T_i = \{V_1, V_2, ..., V_m\}$, where V_j is a tuple of the form $\langle v, n_l, n_h \rangle$ where v is the pathology value, n_l is the normal low and n_h is the normal high. Where necessary each time series T_i was padded to the length of the longest time series in D_i, using the mean value of the v, n_l and n_h values, because the LSTM model requires all time series to be of the same length.

Next, for each of the five tasks, once a time series data set had been constructed, each was passed to the the model layer and the LSTM constructed. Note that each LSTM comprised two hidden layers. The output produce, Out_i, was used to define the softmax layer where predictions will be made. The Softmax Layer, and the Softmax function for normalising the output of each single task LSTM model was as follows:

$$y_i = \frac{e^{a_i}}{\Sigma_{k=1}^{|C|} e^{a_k}} \quad \forall i \in 1...C \tag{2}$$

Where: (i) $|C|$ is the number of classes (three in this case), (ii) a_i is the output of the LSTM layer.

The last layer is the decision layer where the final label is derived. After obtaining all of the five outputs and the predicted labels from the five LSTM models, a decision logic module was added to decide the final prioritisation level of the patient. This included the rule: *"If there exists a prediction that equates to 'High' for one of the tasks then the overall prediction is high, otherwise average the five outputs produced by the Softmax function and choose the class with the maximum probability"*.

For the LSTM, there are five parameters thaqt need to be tuned during the training process. The parameters belong to two categories: (i) optimization parameters and (ii) model parameters. The optimization parameters are: Learning rate, batch size and number of epochs. The model parameters are the number of hidden layers and the number of hidden units. For the optimization, Adam optimization was chosen due to its efficiency and the nature of the adaptive learning rate. For finding the optimal parameters, cross-entropy was used as the loss function, and the parameters tuned by observing the loss and accuracy plots of the training and validation data.

6 Evaluation

From the for going we have two approaches each with two associated variation:

1. Anomaly-Based Supervised Pathology Prioritisation.
   (a) Point-based
   (b) Time series-based
2. Proxy Ground Truth-Based Supervised Pathology Prioritisation.
   (a) kNN
   (b) RNN-based

A significant challenge of ranking pathology data without a ground truth is how to evaluate any proposed approach. There is also no previous work in this area, to the best knowledge of the authors, whereby any direct comparison between the proposed approaches and any existing approaches can be conducted. In the case of the anomaly-based approach we can of course measure the quality of the cluster configuration produced using cohesion and separation measures, such as the well-established Silhouette Coefficient [24]. However, this only tells us about the quality of the clusters configuration, not the quality of the classifications obtained using the cluster configuration.

For learning to rank methodologies, such as those proposed in the context of information retrieval, it is common to use metrics such as Mean Reciprocal Rank (MRR), Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG); but these all require a ground truth ranking. The pathology prioritisation problem was conceptualised as a three class problem, $C = high, medium, low$ which could therefore be evaluated using the "standard" accuracy, precision, recall and F1 metrics using the proxy ground truth (not an actual ground truth but the "nearest best thing"). Five-cross validation was used through out. All the experiments were run using a windows 10 desktop machine with a 3.2 GHz Quad-Core IntelCore i5 processor and 24 GB of RAM. For the LSTM, a GPU was used fitted with a NVIDA GeForceRTX 2060 unit.

As noted earlier, the application focus for the work presented in this paper was Urea and Electrolytes (U&E) pathology testing. For the evaluation U&E data was provided by Arrowe Park Hospital in Merseyside in the UK. Further detail concerning this data set is provided in Sub-section 6.1. The evaluation results are then presented and discussed in Sub-section 6.2.

6.1 Evaluation Data Set

A formalism for U&E data was presented in Section 3. The data set D provided by Arrowe Park Hospital comprised records for 3,734 patient records with five U&E task results (time series) per patient. To derive the proxy ground truth class label for each record $P_j \in \mathbf{D}$ reference was made to the outcome event(s) associated with each patient. As noted earlier, three outcome events were considered: (i) Emergency Patient (EP), an In-Patient (IP) or an Out Patient (OP). These were correlated to the priority descriptor class labels: "high", "medium" and "low". This resulted in 255 patients with high priority, 123 with medium priority and 3,356 with low priority, covering all five tasks. For the LSTM variation of the proxy ground truth based approach, re-sampling of the data was undertaken to give a total $8,192$ time series to address the class imbalanced problem. This was not needed with respect to any of the other three methods considered.

6.2 Results and Discussion

The evaluation results obtained are given in Tables 1 and 2, best results highlighted in bold font. Table 1 gives the precision and recall results obtained, whilst Table 2 gives the accuracy and F1 Score results obtained. Each table includes average values for the five folds and an associated Standard Deviations (SDs).

Fold #	Anomaly Detection Approach				Proxy Ground Truth Classification			
	Point-Based		Time series-Based		*k*NN		RNN	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
1	0.298	0.254	0.407	0.458	**0.661**	0.617	0.446	**0.755**
2	0.375	0.467	0.375	0.511	**0.703**	0.609	0.585	**0.663**
3	0.361	0.500	0.367	0.444	0.639	0.570	**0.695**	**0.629**
4	0.211	0.287	0.333	0.315	0.517	0.632	**0.693**	**0.663**
5	0.500	0.643	0.367	0.508	0.758	0.523	**0.762**	**0.626**
Average	0.349	0.430	0.370	0.448	**0.656**	0.590	0.636	**0.667**
SD	0.095	0.144	0.024	0.071	0.080	0.039	0.111	0.047

Table 1. Precision and recall results, best results in bold font

Fold	Anomaly Detection Approach				Proxy Ground Truth Classification			
#	Point-Based		Time series-Based		*k*NN		RNN	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
1	0.333	0.248	0.421	0.303	0.585	**0.638**	**0.671**	0.561
2	0.273	0.265	0.529	0.306	0.632	**0.653**	**0.642**	0.622
3	0.333	0.361	0.368	0.227	0.576	0.603	**0.622**	**0.660**
4	0.250	0.198	0.429	0.258	0.523	0.569	**0.608**	**0.678**
5	0.500	0.515	0.522	0.290	0.566	0.619	**0.645**	**0.687**
Average	0.338	0.317	0.454	0.629	0.576	0.616	**0.638**	**0.642**
SD	0.087	0.112	0.062	0.030	0.035	0.029	0.024	0.046

Table 2. Accuracy and F1 score, best results in bold font

From the two tables the first thing that can be observed is that the anomaly-based prioritisation approach performed poorly (regardless of which metric was considered) and which variation, point or time series. The reason why the anomaly detection-based prioritisation approach did not perform well might be because it featured the disadvantages that, given a large number of outliers with similar characteristics these might form there own clusters and no longer be considered to be outliers.

From Table 1 it can be seen that the RNN Proxy Ground Truth-based Supervised Pathology Prioritisation produced consistently the best recall, and in three of the five folds the best precision. From Table 2, it can be seen that the best average F1 scores, the harmonic mean of precision and recall and thus a good overall measure, were obtained using the RNN Proxy Ground Truth-based Supervised Pathology Prioritisation method, with the *k*NN method also performing well. Hence, in conclusion, it is argued here that the proxy ground truth-based Supervised method is the most appropriate method for addressing the challenge of pathology data prioritisation as defined in this paper.

7 Conclusions

The motivation for the work presented in this paper was the challenge of prioritising pathology data in the absence of any ground truth data. Two approaches were considered: (i) anomaly detection for prioritisation and (ii) proxy ground truth supervised learning for prioritisation. Two variations of both approaches were considered, point-based and time series-based for the first approach; and *k*NN and RNN-based, for the second. The four variations (methods) were fully described and evaluated using real data. From the results, the RNN proxy ground truth-based supervised pathology prioritisation method was argued to be the most appropriate. For future work the authors intend to investigate: (i) generate artificial evaluation data sets to provide for a more comprehensive evaluation, and (ii) collaborate with clinicians to obtain feed back regarding the prioritisations produced and to test the utility of the best performing mechanism in a real setting.

References

1. Bilal Ahmad, Wang Jian, Zain Anwar Ali, Sania Tanvir, and M. Sadiq Ali Khan. Hybrid anomaly detection by using clustering for wireless sensor network. *Wireless Personal Communications*, 106:1841–1853, 2019.
2. Mohiuddin Ahmeda, Abdun Naser Mahmooda, and Md. Rafiqul Islamb. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278–288, 2016.
3. S. Baek, D. Kwon, J. Kim, S. C. Suh, H. Kim, and I. Kim. Unsupervised labeling for supervised anomaly detection in enterprise and cloud networks. In *Proceedings 4th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud'17)*, pages 205–210, 2017.
4. Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery*, 31(3):606–660, 2017.
5. Keith Brawner and Michael W. Boyce. Establishing ground truth on pyschophysiological models for training machine learning algorithms: Options for ground truth proxies. In *Proceedings of the International Conference on Augmented Cognition*, pages 468–477. Springer, 2017.
6. R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, 10(1):5:1–51, 2015.
7. Dean Cerrato, Rosie Jones, and Avinash Gupta. Classification of proxy labeled examples for marketing segment generation. In *Proceedings of the 17th International Conference Knowledge discovery and data (KDD'2011)*, page 343–350. ACM SIGKDD, 2011.
8. Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.
9. Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33:917–963, 2016.
10. Zoltan Geler, Vloadimir Kurbalija, Miloš Radovanović, and Mirjana Ivanović. Comparison of different weighting schemes for the knn classifier on time-series data. *Knowledge and Information Systems*, 48:331–378, 2016.
11. B. S. Harish and S. V. Aruna Kuma. Anomaly based intrusion detection using modified fuzzy clustering. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(6), 2017.
12. Z. He, X. Xu, and S. Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9-10):1641–1650, 2003.
13. L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
14. Kamran Khan, Saif Ur Rehman, Kamran Aziz, Simon Fong, and Sababady Sarasvady. Dbscan: Past, present and future. In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pages 232–238. IEEE, 2014.
15. Vipin Kumar. Parallel and distributed computing for cybersecurity. *IEEE Distributed Systems Online*, 6(10), 2005.

16. Max Landauera, Florian Skopika, Markus Wurzenbergera, and AndreasRauberb. System log clustering approaches for cyber security applications: A survey. *Computers and Security*, 92, 2020.
17. Yen-Hsien Lee, Chih-Ping Wei, Tsang-Hsiang Cheng, and Ching-Ting Yang. Nearest-neighbor-based approach to time-series classification. *Decision Support Systems*, 53(1):207–217, 2012.
18. Lishuai Li, Santanu Das, R. John Hansman, Rafael Palacios, and Ashok N. Srivastava. clustering techniques to detect abnormal flights of unique data patterns. *Journal of Aerospace Information Systems*, 2015.
19. Zheng-xin Li, Shi-hui Wu, Yu Zhou, and Chao Li. A combined filtering search for dtw. In *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, pages 884–888. IEEE, 2017.
20. Christopher D. Manning, Raghavan Prabhakar, and Hinrich Schutza. *Introduction to information retrieval.* Cambridge University Press, 2008.
21. Nadia Rahmah and Imas Sukaesih Sitanggang. Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra. In *IOP Conference Series: Earth and Environmental Science*, volume 31, page 012012. IOP Publishing, 2016.
22. Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 262–270, 2012.
23. Adil Roohi, Kevin Faust, Ugljesa Djuric, and Phedias Diamandis. Unsupervised machine learning in pathology: The next frontier. *Surgical Pathology Clinics is published by Elsevier*, 13(2):349–358, 2020.
24. Peter J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 20:53–65, 1987.
25. Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.
26. Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data mining and knowledge discovery*, 2(2):169–194, 1998.
27. Srikanth Thudumu, Philip Branch, Jiong Jin, and Jugdutt (Jack) Singh. A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, 7:Paper number 42, 2020.
28. Sharad Vikram, Lei Li, and Stuart Russell. Handwriting and gestures in the air, recognizing on the fly. In *Proceedings of the CHI*, volume 13, pages 1179–1184, 2013.
29. Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2):275–309, 2013.
30. Albert C. Yang, Norden E. Huang, Chung-Kang Peng, and Shih-Jen Tsai. Do seasons have an influence on the incidence of depression? the use of an internet search engine query data as a proxy of human affect. *Plos One*, 2010.
31. Chong Zhou and Randy C. Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*, page 665–674, 2017.