Smart Diagnostic Algorithms for Automated Detection of Childhood Pneumonia in Resource-constrained Settings

Elina Naydenova, Althanasios Tsanas, Climent Casals-Pascual, Maarten De Vos (University of Oxford)
Stephen Howie (Child Survival Theme, Medical Research Control Unit)

***Abstract*—Pneumonia is the leading cause of death in children under five, with 1.1 million deaths annually - more than the combined burden of HIV/AIDS, malaria, and tuberculosis for this age group; the majority of these deaths occur in resource-constrained settings. Accurate diagnosis of pneumonia relies on expensive human expertise and requires the evaluation of multiple clinical characteristics, measured using advanced diagnostic tools. The shortage of clinical experts and appropriate diagnostic tools in many low and middle income countries impedes timely and accurate diagnosis. We demonstrate that the diagnostic process can be automated using machine learning techniques, processing several clinical measurements that could be obtained with affordable and easy-to-operate point-of-care tools. We evaluated our findings on a dataset of 1093 children, comprising 777 diagnosed with pneumonia and 316 healthy controls, on the basis of 47 clinical characteristics. Seven feature selection techniques were used to identify robust, parsimonious subsets of clinical characteristics, which could be measured reliably and affordably. Standard machine learning techniques, such as support vector machines and random forests, were used to develop a predictive algorithm based on the four jointly most predictive characteristics (temperature, respiratory rate, heart rate and oxygen saturation); this approach led to 96.6% sensitivity, 96.4% specificity, and an Area Under the Curve (AUC) of 97.8%. The proposed approach can be easily embedded in a mobile phone application, allowing for point-of-care assessment and identification of children in need of clinical attention by basically trained healthcare workers in resource-constrained settings.**

Keywords*—*Childhood pneumonia, diagnostics, machine learning, Random forests.

I. Introduction

PNEUMONIA is the leading cause of death in children under five. The disease is particularly prevalent in resource-constrained settings and kills 1.1 million children annually, more than HIV/AIDS, malaria, and tuberculosis combined [1]–[3]. One of the main reasons for these devastating statistics is diagnostic complexity: accurate diagnosis of pneumonia requires the evaluation of multiple clinical parameters, measured with advanced diagnostic tools, which are used as decision support tools by highly-trained clinicians [4], [5]. The shortage of clinical experts and appropriate diagnostic tools in many Low and Middle Income Countries (LMICs) impedes timely and accurate diagnosis. The gold standard for identification of childhood pneumonia includes a combination of physical signs (e.g. fever, chest recession and respiratory rate); cases with suspected severe pneumonia are further examined using blood culture and chest radiographs. However, in LMICs, many of these advanced diagnostic tools are either unavailable or the local clinical staff lacks the expertise to interpret many of these measurements.

The World Health Organisation (WHO) has developed a set of guidelines for diagnosis of childhood pneumonia in resource-constrained settings, in the context of a broader strategy for Integrated Management of Childhood Illness (IMCI) [6]. However, the integration of these guidelines into clinical practice worldwide has been reported to deliver a reasonably high sensitivity of diagnosis (69%-94%) but rather poor specificity (16%-67%) [7]–[9]. This leads to unnecessary antibiotic prescription, causing drug stocks depletion and increased microbial resistance. The IMCI-defined symptoms for identification of pneumonia include: fast breathing (≥ 50 breaths per min in a child aged 2-11 months and ≥ 40 breaths per min in a child aged 1-5 years) and chest indrawing.

The technological revolution has transformed multiple existing industries and defined entirely new ones. Healthcare systems worldwide have benefited from this phenomenon; in LMICs, technology is often pointed out as the tool for leapfrogging unaffordable and inefficient solutions from the past (e.g. mobile health as an alternative to conventional healthcare delivery). Specifically, computational algorithms, combined with appropriate hardware, offer a potential solution to the shortage of medical expertise. For example, data fusion of multiple clinical parameters has been applied to various health problems to improve diagnostic outcomes, e.g. intensive care unit (ICU) monitoring systems for mortality prediction and decision support systems for neonatal ICU [10]–[13]. A data based decision support approach has been previously applied to diagnostics of pneumonia, focusing on identification of patients suitable for treatment at home, reducing hospitalisation and healthcare costs [14]–[16]. All these studies have applied a wide range of machine learning techniques to datasets collected in hospitals, where a substantial number of clinical parameters have been recorded by highly trained staff (between 46 - 158 parameters) and this data is used to predict health outcomes.

On the other hand, no substantial research has been done on fusing a few clinical variables, which can be measured by basically trained health workers in resource-constrained settings, to automate diagnostics of childhood pneumonia. Previous work in this field has implemented basic analytical tools such as thresholding of individual clinical variables [17]–[19]; however, individually, none of these variables are sensitive or specific enough (Table I).

Other diagnostic problems, involving continuous and discrete clinical variables, have been investigated via a wide range

TABLE I. SUMMARY OF SENSITIVITY AND SPECIFICITY OF INDIVIDUAL CLINICAL PARAMETERS IN IDENTIFYING CHILDHOOD PNEUMONIA (ACCORDING TO THRESHOLDS DEFINED IN ACCEPTED GUIDELINES) AS REPORTED IN THE RESEARCH LITERATURE [17]–[19].

Clinical parameter	Sensitivity	Specificity
X-ray scan	Gold standard?	Gold standard?
Rales	15%	99%
Respiratory rate	50-70%	43-95%
Rhonchi	26%	98%
Oxygen saturation	26-63%	77-93%
Tachycardia	51%	70%
Fever	47%	68%
Retractions	26%	98%
Crackles	43%	73%
Wheezing	4%	98%
Biomarkers	82-96%	53-61%

of machine learning techniques [20]–[23]. This paper presents the development and application of a mixture of methods - feature selection and classification - that were used in realising a dual research objective: (1) identification of a minimal set of clinical parameters that improves diagnosis and can be measured with affordable and easy-to-operate point-of-care tools in resource-constrained settings; (2) design of machine learning algorithms that fuse information from individual measurements to deliver an accurate diagnostic outcome.

II. METHODS

The dataset used in this project was collected as part of a clinical study described by Huang et al. [24]. The 1581 participants were Gambian children aged 2-59 months: 780 were diagnosed with childhood pneumonia and 801 were recruited as healthy controls. Originally, the dataset consisted of 57 variables (i.e. clinical characteristics); these variables included measurable clinical parameters (e.g. white blood cell count, neutrophils, haemoglobin etc.), observational clinical parameters (e.g. sleepiness, sternal indrawing, cough heard etc.) and conventional vital signs (e.g. respiratory rate, heart rate, oxygen saturation etc.). The presence of pneumonia was identified by a clinician on the basis of WHO guidelines [25] and IMCI [6] and this was used as the outcome when training and testing an algorithm. All analysis of this dataset was performed in Matlab2014.

A. Preprocessing

The original dataset contained a substantial number of missing values, which necessitated preprocessing. The following steps were taken: (1) features and samples where more than 20% of the data were missing were excluded from the analysis; (2) for the remaining features, we used imputation methods [26]. Three different imputation techniques were investigated (feature mean, feature median and k-Nearest Neighbour); given that the dataset contained a mixture of continuous and discrete features, the latter two were more appropriate. However, no substantial differences were recorded in the distributions of features across imputation techniques; therefore, in cases of missing values, we simply imputed with the corresponding median for that feature. In order to investigate the effect of imputation, we created additional vectors for each feature that contained imputations - these vectors contained 'ones' where imputation was done and 'zeros' otherwise. For ease of comprehension, we will refer to these vectors as ghost vectors throughout the paper. After this initial preprocessing we obtained a design matrix with 1093 samples and 47 features (29 clinical characteristics and 18 ghost vectors).

B. Feature selection

Seven feature selection techniques were used to investigate the relevance of clinical variables to the outcome: maximum relevance on the basis of correlation coefficients, maximum relevance minimum redundancy (mRMR), Relief, Gram-Schmidt orthogonalisation, Least Angle Shrinkage and Selection Operator (LASSO), Elastic Net and sparse Linear Discriminant Analysis (sLDA). The performance of each technique has been documented in the literature and some techniques have been seen to perform better depending on the choice of classifier applied next [26-35]. For the purposes of this diagnostic challenge, we wanted to obtain a more objective selection of features, irrespective of the bias that each technique carries. We therefore developed a majority voting approach that consolidates results from all techniques, executed in the following way: (1) the data is randomly split in ten sections of equal size; (2) each feature selection technique is applied to each section, delivering a rank for each feature; (3) the first two steps are repeated 50 times to minimise the bias that comes from the arbitrariness of the data split; (4) an accumulative frequency was derived for each feature for each ranking position; (5) based on this accumulative frequency, the most represented features in the first 10 ranking positions were identified. Additionally, each feature was assessed according to three criteria: whether it is measurable in a quantifiable and objective manner; and in case this is true, whether it is measurable in a point-of care setting; and whether the measurement is affordable. Assessment was done based on currently available options as well as feasible design of novel tools. Finally, for the feature to be selected for analysis it should satisfy all of these three criteria as well as be present in the top 10 of at least three "fundamentally different" techniques. Namely, the pairs of techniques, not considered fundamentally different, were: correlation & mRMR and Lasso & Elastic Net, as they share a very similar theoretical basis.

To visualise the effects of feature selection, a dimensionality reduction technique originally defined by van der Maaten - t-Stochastic Neighbourhood Embedding (t-SNE) was used [27]. More conventional methods in this field, such as Principle Components Analysis (PCA) [28] and Multidimensionality Scaling (MDS) [29], are linear and therefore mainly preserve the separation between dissimilar data entries within a low-dimensional space. However, most medical data occupies non-linear manifolds, making it essential to also preserve the closure between similar data entries. t-SNE has been previously reported to capture aspects of both the local as well as the global structure, preserving the neighbouring probabilities

of samples. It calculates Euclidean distances between data entries and derives similarities (conditional probabilities) by assuming a Student-t distribution [27]. t-SNE was applied to the list of features, constructed after feature selection, and the formation of clusters was compared to the outcome. Additionally, the distribution of each feature across the two clusters was investigated in order to identify potential reasons for misclassification.

C. Classification

Three machine learning techniques were adapted and applied to this diagnostic challenge: Logistic Regression, Support Vector Machines (SVM) and Random Forests (RF). The implementation of all three techniques to various medical diagnostics problems has been documented in the literature [26], [30], [31]. The first technique only required the optimisation of beta-coefficients, whereas the other two involved more advanced fine-tuning of several parameters. Specifically, a Gaussian radial basis function kernel was used in SVM, determining the optimal values of the kernel parameter γ and the penalty parameter C though a grid search of possible values. The LIBSVM implementation was used [32]. Regarding RF, the number of trees and number of features for splitting were optimised through a grid search of parameters.

For each machine learning algorithm, we take an aggressive validation approach by that has three elements. First, 4-fold cross-validation was performed to test variance in performance, training the data on 820 randomly selected cases (training set) and testing on the remaining 273 (test set). Second, internal 5-fold cross-validation was performed to optimise parameters in each training set, based on Area Under the Curve (AUC). A separate algorithm was trained for a different number of features; each one was applied to the test set, recording the following performance metrics: sensitivity, specificity, AUC, balanced accuracy. Third, the first two steps were repeated 100 times to offset any bias in the split of the data. Algorithm performance in this paper is reported in terms of mean values as well as variance across repetitions.

The performance of the three algorithms was measured as a function of the number of features presented to the classifier, identifying both the optimal number of features as well as the best performing algorithm. Additionally, probabilistic predictions were derived to reflect the degree of certainty with which any given entry can be assigned to each of the two classes. These predictions were mapped back to the cluster structure established with t-SNE, contextualising classification accuracy to the global structures of the data.

III. Results

A. Feature selection

Using our multi-stage approach to feature selection, 17 out of the 47 features were selected by at least one technique; from those, 5 features were highlighted through the multi-stage feature selection approach: Temperature, HR, RR, Osat and Malnutrition Score (WHZ) (Figure 1). None of the ghost vectors were ranked amongst the Top 10 on either of the feature selection techniques, but relevant ghost vectors (the WHZ ghost vector) were added to the list to test their effect on classification. Applying t-SNE on the dataset of 1093 cases

		Relevance	mRMR	Relief	Gram-Schmidt	Lasso	Elastic net	sLDA	Measurable	Point-of-care	Affordable
Vital signs	Temperature	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
	HR	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	RR	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Osat	N	Y	Y	N	N	N	Y	Y	Y	Y
Measurable clinical features	WBC count	Y	N	N	N	N	N	N	Y	N	N
	Neutrophils	Y	Y	N	Y	Y	Y	N	Y	N	N
	Lymphocytes	N	Y	N	Y	Y	Y	N	Y	N	N
	WHZ	N	Y	N	Y	Y	Y	Y	Y	Y	Y
	Age	N	N	Y	N	N	N	Y	Y	Y	Y
Lung sounds	Grunting	Y	N	N	N	N	N	N	Y	Y	Y
Observational clinical features	Sleepy	Y	N	Y	N	N	N	N	N	N	Y
	Pallor	Y	N	N	N	N	N	N	N	N	Y
	Head nodding	Y	N	N	N	N	N	N	N	N	Y
	Cough	N	Y	Y	Y	Y	Y	Y	N	N	Y
	Yellow fever	N	N	Y	N	N	N	N	N	N	Y
	Unwell	N	N	N	N	N	N	Y	N	N	Y
	Cyanosis	N	N	N	N	N	N	Y	N	N	Y

Fig. 1. Features voted in the Top 10 across all 10 folds and 50 iterations, for each feature selection algorithm. Features were also assessed with respect to their measurability, point-of-care and affordability. Based on these criteria, the following features were selected: Respiratory rate (RR), Heart rate (HR), Temperature (T), Oxygen saturation (Osat) and Malnutrition Score (WHZ)

and 5 features, we were able to identify 2 clear clusters; when compared to the outcomes these were shown to be consistent with the separation between Pneumonia and No Pneumonia classes (Figure 2). With the exception of a few cases, the clustering outcome gave us confidence that separation between the two classes should be possible on the basis of these five features.

Additionally, the distribution of each individual feature across the two clusters was investigated (Figure 5). From Figure 3 and 4, Pneumonia can be confirmed in cases with: $RR > 65$, $HR > 175$, $T > 38.5$, $Osat < 87$, $WHZ < -5$ and $WHZ > 1.8$. However, the data contains a substantial amount of clinical uncertainty - there are many cases where this type of simple thresholding is insufficient to separate the two classes. A closer look at the overlap zones identified in Figure 2, illustrates some of this uncertainty. In Zone A: the No Pneumonia entries have elevated HR and low WHZ; the Pneumonia entries have low RR, low T. In Zone B: the No Pneumonia entries have elevated HR; the Pneumonia entries have low RR, no fever; and WHZ values are equally spread across both classes. In Zone C: the Pneumonia entries have low HR, no fever; and WHZ values are equally spread across both classes. This preliminary analysis indicated that more advanced

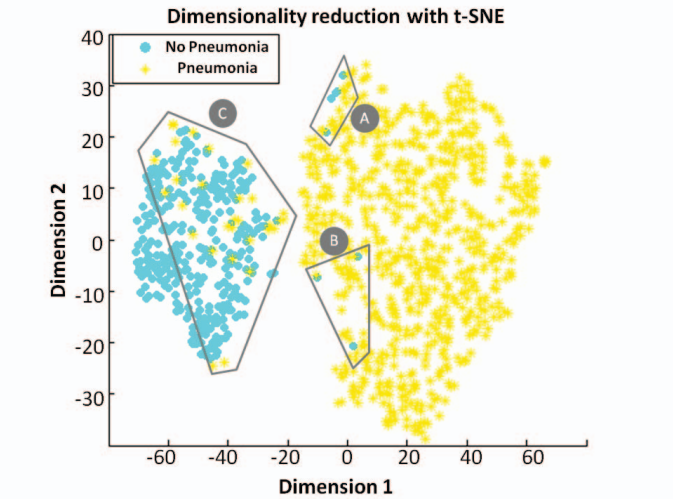

Fig. 2. Dimensionality reduction with tSNE using data from the five selected features. Data points were coloured according to the outcome for purposes of visualisation; however, the outcome was not used in constructing the clusters. Three zones of overlap (A, B, C) between the two classes were identified. Note that these regions have been arbitrarily defined to facilitate visual inspection and carry no mathematical meaning.

machine learning techniques might be needed to separate the two classes.

B. Classification using machine learning algorithms

Classification with all three techniques was optimised when the first four features were used: RR, HR, T and Osat. The following hyperparamters were found to lead to best classification performance: a γ of 0.1 and a cost parameter of 1000 for SVM; 750 decision trees and searching over 2 variables at each tree node for RF. Comparison between the ROC curves of each technique revealed similar performance, with somewhat more favourable sensitivity obtained via RF than SVM and Logistic Regression, for same specificity values (Figure 5).

The performance of the RF algorithm was investigated further. Through the internal cross-validation, we found that the optimal hyperparameter values in RF were 750 decision trees, and searching over 2 variables at each tree node. Increasing the number of features was seen to be particularly beneficial with respect to specificity (Figure 6). With four features, RF delivered: sensitivity of 96.6% (95% CI 95.8% - 97.6%); specificity of 96.4% (95% CI 95.3%-98.0%); AUC of 97.8% (95% CI 97.7%-97.9%); balanced accuracy of 96.6% (95% CI 96.1%-97.1%). The variance across iterations was observed to be below 5% across all metrics. The ghost vector present in the dataset was seen to have limited effect on classification, with changes in all metrics limited to +/- 1%.

Finally, probabilistic values were derived from the RF algorithm and these were mapped back to the cluster structure derived with t-SNE (Figure 7). The optimal threshold for conversion of this probabilistic outcome into a binary one was recorded to be 0.35. The majority of data points are classified

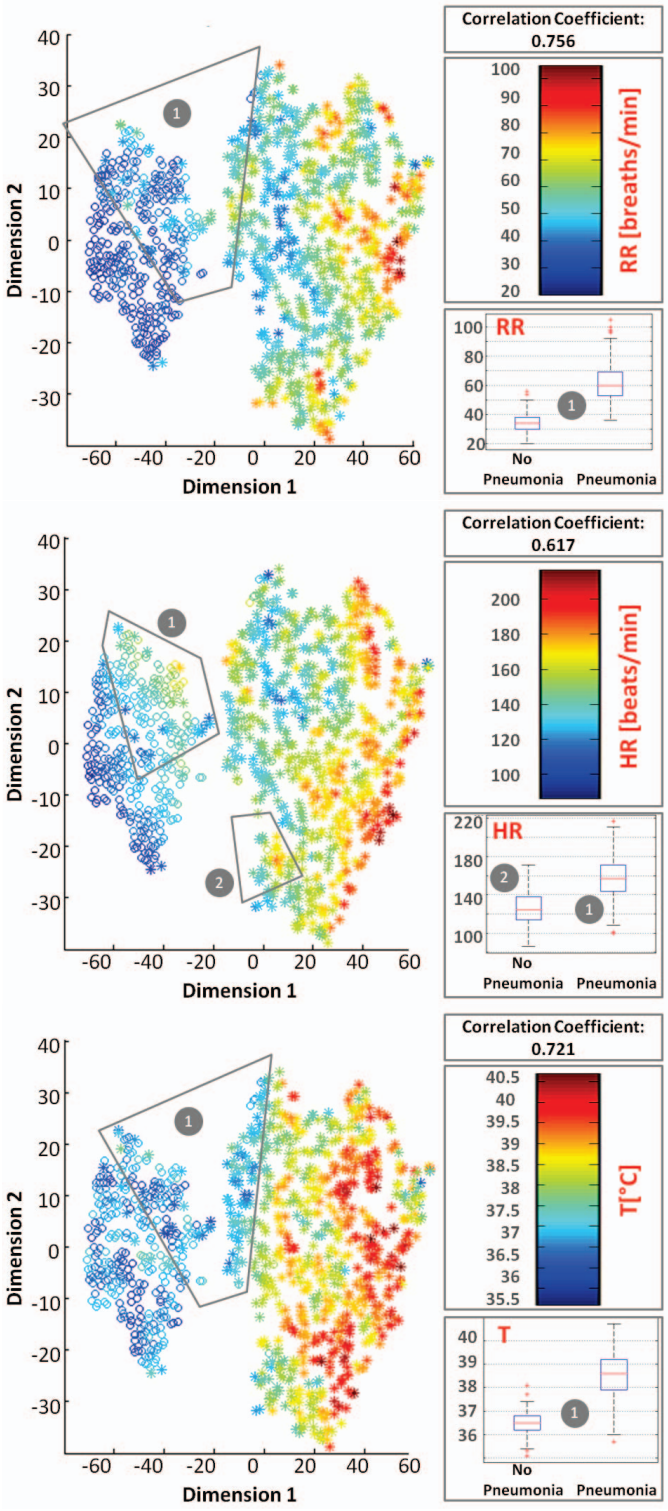

Fig. 3. Distribution of individual features (RR, HR and T) across both clusters. The shapes definition from Figure 4 has been preserved - ○ for No Pneumonia and * for Pneumonia cases. The color scale represents the corresponding magnitude of that feature per entry. Areas of interest are identified through the boxes around them. The boxes on the right of each plot contain the following information: (top) correlation coefficient between feature and outcome; (middle) colour bar for the colour scale used in the plot; (bottom) a boxplot of the data in each feature, split between the two classes is included. In each box, the central/red line represents the median, the edges are the 25th and the 75th percentiles, and the dashed lines extend to the most extreme data points.

with high certainty. Some data points are classified with less certainty and these zones have been highlighted in Figure 8. Zone A: A 'False Negative' (FN) case and a 'False Positive' (FP) case were observed, with the latter being very close to the threshold value; some of the Pneumonia cases were classified correctly but with less certainty. Zone B: contained no FN cases but some Pneumonia cases with quite low probabilistic values. Zone C: The most FP and FN cases were identified here as well as a few borderline cases, where the probabilistic assignment was seen to be quite close to the threshold value.

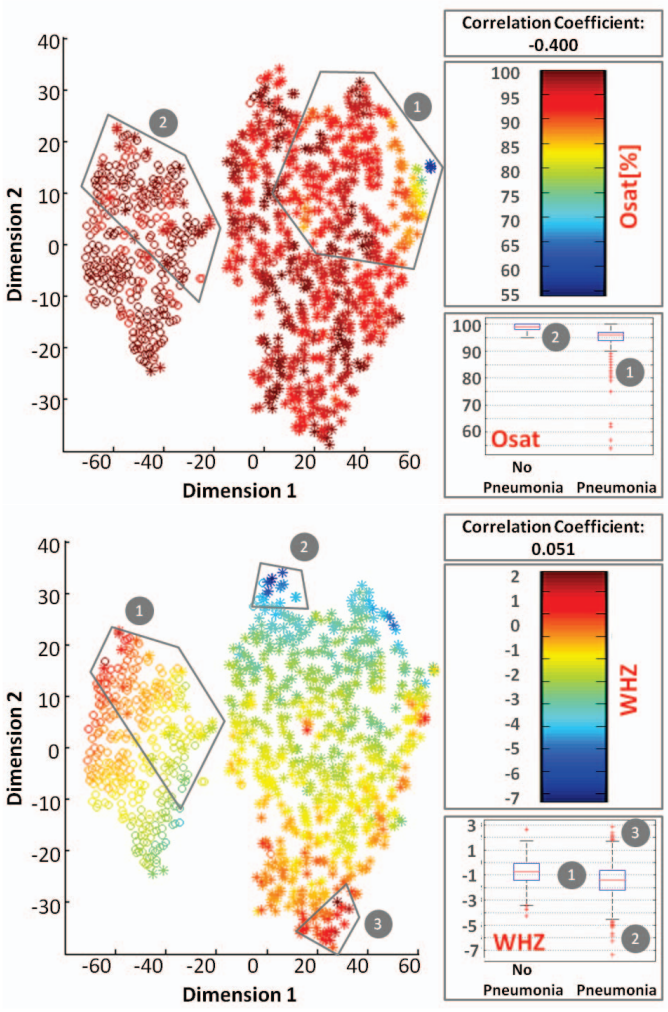

Fig. 4. Distribution of individual features (Osat and WHZ) across both clusters. The shapes definition from Figure 4 has been preserved - ○ for No Pneumonia and * for Pneumonia cases. The color scale represents the corresponding magnitude of that feature per entry. Areas of interest are identified through the boxes around them. The boxes on the right of each plot contain the following information: (top) correlation coefficient between feature and outcome; (middle) colour bar for the colour scale used in the plot; (bottom) a boxplot of the data in each feature, split between the two classes is included. In each box, the central/red line represents the median, the edges are the 25th and the 75th percentiles, and the dashed lines extend to the most extreme data points.

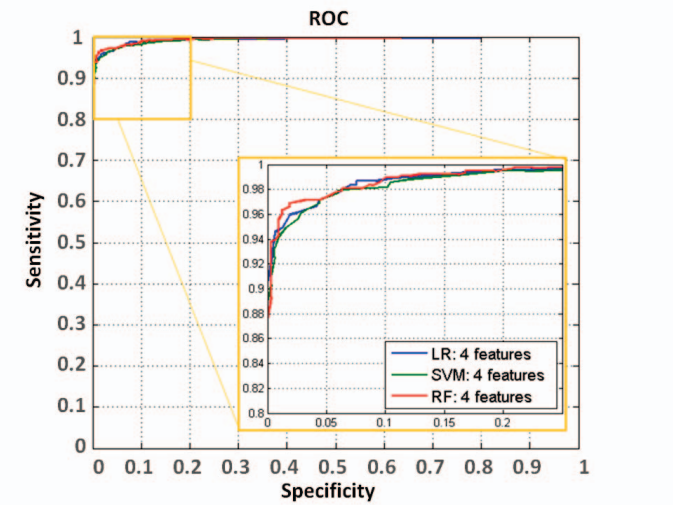

Fig. 5. ROC obtained via the three classification techniques (Logistic Regression (LR), Support Vector Machine (SVM), Random Forests (RF)), applied to four features (RR, HR, T and Osat) of the data. The yellow window contains a close-up view on the top section of the ROCs to highlight differences.

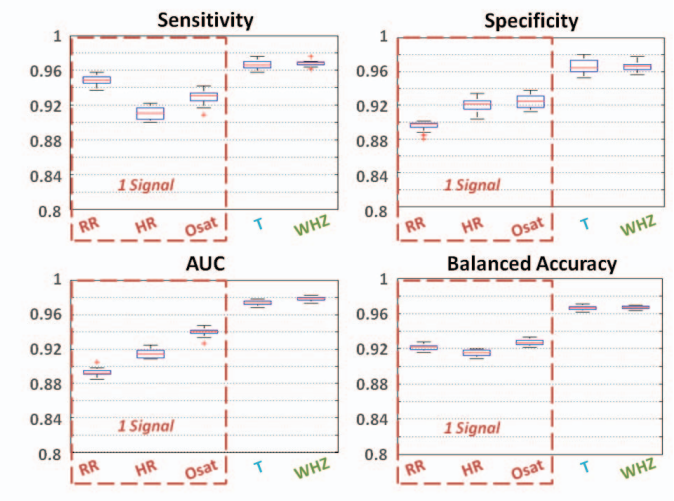

Fig. 6. Sensitivity, specificity, AUC and balanced accuracy plots for RF and an increasing number of features. Results are presented through a boxplot to assess robustness and variability of this technique. In each box, the central/red line represents the median, the edges are the 25th and the 75th percentiles, and the dashed lines extend to the most extreme data points. Along the x axis, features are listed in an additive manner, i.e. each x-entry represents classification performed on a dataset including that specific feature and all features on the left of it. Note that the y-axis has only been plotted in the 0.8-1 range to better visualise differences between different boxplot graphs. The red box around the first three features indicates that all three of them can be extracted from an individual signal/measurement (PPG signal)

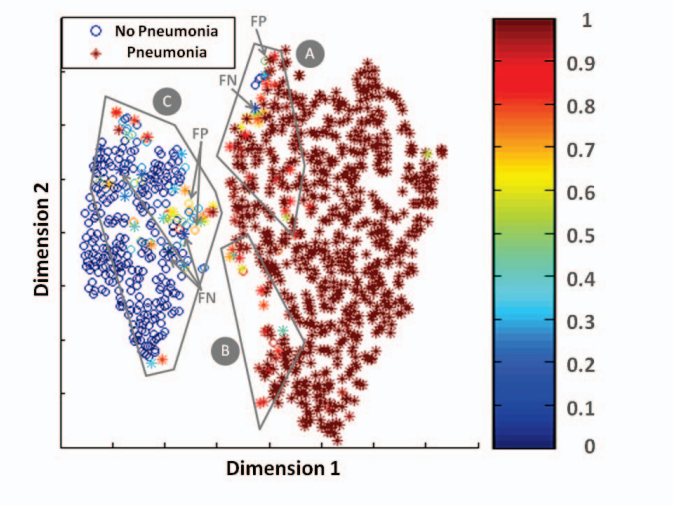

Fig. 7. A comparison between the two dimensional data representation established with t-SNE and classification with RF. The shape of each data point represents its class according to the real outcome. The colour of each data point represent the probabilistic prediction derived from RF, where 1 corresponds to 100% certainty of pneumonia and 0 corresponds to 100% certainty of 'No Pneumonia'. Zones of uncertainty have been identified on the plot and labelled as A,B & C.

IV. Discussion & Conclusions

This study highlighted a few interesting points in the context of algorithms for automated diagnosis of childhood pneumonia and these are summarised below.

This project reinforced the observation that certain clinical features can be individually predictive of pneumonia. However, it is only when these individual features are combined, that high levels of both sensitivity and specificity are attained.

Oxygen saturation was seen to be very specific but not very sensitive; low levels of SpO2 were seen to be highly indicative of pneumonia. However, normal levels of SpO2 were observed across both classes. This is consistent with the fact that hypoxia mainly presents in very severe pneumonia cases (in approximately 20%) of the cases [44].

Malnutrition has been documented to have high relevance to pneumonia diagnostics, especially in the context of biomarkers, where concentration levels can vary across subjects depending on their malnutrition status [24]. In the dataset analysed here, extreme malnutrition scores were seen to be related to the presence of pneumonia; however, the majority of cases with more neutral malnutrition scores were seen to be equally distributed across both classes. Given that this study uses vital sign measurements, this finding is unsurprising - the partial mutual information analysis saw little overlap between the malnutrition score and any of the other measurements.

Some features were identified by the feature selection techniques amongst the most predictive (Cough, Neutrophils, Lymphocytes) but were excluded from the analysis due to diagnostic application considerations. Specifically, the cough variable relies on a parent/accompanying adult identifying cough in the child prior to the health consultation. Additionally, Abeyratne et al. have reported on the use of cough recordings, in combination with a few other basic clinical parameters, achieving high sensitivity (94%) but moderate specificity (75%) [33]. Moreover, the approach requires continuous sound recording in a hospital setting; in practice, the large volume of patients in primary care facilities and the limited tolerance young children have for physical examinations mean that consultation times are typically less than two minutes. The "Neutrophils" and "Lymphocytes" variables are difficult to measure in low resource settings, let alone in a point-of-care fashion.

This analysis highlighted the fact that simple thresholding of clinical features is insufficient to provide reliable diagnostic outcomes. Clinical uncertainty in the presentation of individual childhood pneumonia cases means that advanced interpretation of several parameters simultaneously is needed to deal with cases which would otherwise require expert training to deal with.

The IMCI guidelines were used by a clinician in defining the gold standard in this study. However, simply automating diagnostics through that set of recommendations is insufficient. Therefore, this study aimed to identify measurable and quantifiable features that, combined through machine learning, would remove some of the ambiguity that stems from the use of observational parameters such as chest indrawing which often become obvious only in the later stages of disease.

In this context, the three standard machine learning techniques investigated in this study demonstrated an ability to automate childhood pneumonia diagnostics using quantifiable clinical features that could be measured in point-of-care settings. The superior performance of RF indicates non-linear interactions between the features. The use of more advanced machine learning methods, such as neural networks, could be investigated in the future but would possibly require a bigger dataset. Nevertheless, the classification performance recorded with RF indicated that it is possible to address some of the diagnostic challenges associated with childhood pneumonia through machine learning, including the high degree of clinical uncertainty, and provide clinical workers with an easily interpretable recommendation.

The dataset analysed in this project is not necessarily representative of a realistic medical setting. Namely, the control cases contain some clinical variability (e.g. some elevated HR and low SpO2) but overall are very healthy. For a machine learning algorithm to be reliably applicable to a realistic clinical scenario, it should be trained on a population where the control cases also suffer from other symptomatically similar diseases (e.g. malaria and tuberculosis). Nevertheless, this initial analysis provides a substantial foundation; we are currently working towards expanding these findings through the use of larger and more diverse datasets.

The analysis presented in this study confirmed the original hypothesis that fusion of individual clinical signs through machine learning methods has the potential to facilitate automated and improved pneumonia diagnostics. Moreover, the approach taken regarding feature selection allowed the identification of informative features that can also be measured in an affordable and reliable way. As highlighted in the results section, some of

these features can be derived from the same clinical measurement. Specifically, the first three features from the proposed set (i.e. RR, HR, Osat) could be all derived simultaneously from the PPG signal obtained through a pulse oximeter, keeping cost and clinical input required low. Furthermore, the designed algorithm could be embedded in a mobile application, which, equipped with the right hardware, would support community health workers in their clinical decision making process.

A brief market survey of off-the-shelf pulse oximeters, which could connect to a mobile phone and allow derivation of RR, HR and Osat, indicated a rough procurement cost of $100. A simple digital thermometer could be added for less than $5, increasing sensitivity by 3.5% and specificity by 4%. A comparison with current associated healthcare costs (blood culture and X-ray for a wide range of cases) suggests that such an approach could improve patient outcomes in a cost-effective way. More structured market research is needed to quantify the cost-effectiveness of this approach and identify existing tools that would be optimal for the application in mind.

ACKNOWLEDGMENT

MDV and EN acknowledge the support of the EPSRC and the RCUK Digital Economy Programme grant number EP/G036861/1(Oxford Centre for Doctoral Training in Healthcare Innovation). The authors would like to thank the Skoll Centre for Social Entrepreneurship at the University of Oxford for supporting this study and the researcher vision behind the larger project. This study was also supported by the Wellcome Trust through a Centre Grant No. 098461/Z/12/Z, "The University of Oxford Sleep and Circadian Neuroscience Institute (SCNi)". The authors would like to thank the MRC Unit The Gambia Severe Pneumonia Studies and Pneumococcal Surveillance Project field workers, nurses, laboratory staff, data staff, drivers and administrative staff; the staff at the MRC Hospital in Fajara; the government health staff at the Royal Victoria Teaching Hospital (RVTH), and the health centers at Basse, Fajikunda, Serekunda, and Brikama. Particular thanks to Mr Simon Donkor for his data management work and Mr Nuru Adams for his management of the plasma samples.

REFERENCES

[1] WHO/UNICEF, "Ending preventable child deaths from pneumonia and diarrhoea by 2025," *The Integrated Global Action Plan for Pneumonia Diarrhoea (GAPPD)*, 2013.

[2] I. Rudan, C. Boschi-Pinto, and H. Campbell, "Epidemiology and etiology of childhood pneumonia," *Bulletin of the World Health Organization*, vol. 86(5), pp. 408–416, 2008.

[3] UNICEF/WHO, "Pneumonia: the forgotten killer of children," 2006.

[4] WHO, "Community case management of pneumonia: at a tipping point?" *Bulletin of the World Health Organization*, 2008.

[5] ——, "Challenges to improving case management of childhood pneumonia at health facilities in resource-limited settings," *Bulletin of the World Health Organization*, 2008.

[6] WHO/UNICEF, "Integrated management of childhood illness (imci) handbook," 2005.

[7] M. R. Cardoso, C. M. Nascimento-Carvalho, F. Ferrero, F. M. Alves, and S. N. Cousens, "Adding fever to who criteria for diagnosing pneumonia enhances the ability to identify pneumonia cases among wheezing children," *Archives of Disease in Childhood*, vol. 96, no. 1, pp. 58–61, Jan 2011.

[8] M. Harari, V. Spooner, S. Meisner, M. Carney, F. Shann, and J. de Campo, "Clinical signs of pneumonia in children," *The Lancet*, vol. 338, no. 8772, pp. 928–930, 1991.

[9] M. Palafox, H. Guiscafr, H. Reyes, O. Munoz, and H. Martinez, "Diagnostic value of tachypnoea in pneumonia defined radiologically," *Archives of Disease in Childhood*, vol. 82, no. 1, pp. 41–45, 2000.

[10] S. Agarwal and G. N. Pandey, "Human computer interface design for neonatal intensive care with data mining," in *4th International Conference on Intelligent Human Computer Interaction: Advancing Technology for Humanity, IHCI 2012*, 2012.

[11] R. S. Hum, K. Cato, B. Sheehan, S. Patel, J. Duchon, P. DeLaMora, Y. H. Ferng, P. Graham, D. K. Vawdrey, J. Perlman, E. Larson, and L. Saiman, "Developing clinical decision support within a commercial electronic health record system to improve antimicrobial prescribing in the neonatal icu," *Applied Clinical Informatics*, vol. 5, no. 2, pp. 368–387, 2014.

[12] L. A. Celi, R. G. Mark, D. J. Stone, and R. A. Montgomery, ""big data" in the intensive care unit: Closing the data loop," *American Journal of Respiratory and Critical Care Medicine*, vol. 187, no. 11, pp. 1157–1160, 2013.

[13] J. Lee, D. J. Scott, M. Villarroel, G. D. Clifford, M. Saeed, and R. G. Mark, "Open-access mimic-ii database for intensive care research," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2011, pp. 8315–8318.

[14] G. F. Cooper, C. F. Aliferis, R. Ambrosino, J. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, C. Glymour, G. Gordon, B. H. Hanusa, J. E. Janosky, C. Meek, T. Mitchell, T. Richardson, and P. Spirtes, "An evaluation of machine-learning methods for predicting pneumonia mortality," *Artificial Intelligence in Medicine*, vol. 9, no. 2, pp. 107–138, 1997.

[15] G. F. Cooper, V. Abraham, C. F. Aliferis, J. M. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, J. E. Janosky, G. Livingston, T. Mitchell, S. Monti, and P. Spirtes, "Predicting dire outcomes of patients with community acquired pneumonia," *Journal of Biomedical Informatics*, vol. 38, no. 5, pp. 347–366, 2005.

[16] S. Visweswaran and G. F. Cooper, "Patient-specific models for predicting the outcomes of patients with community acquired pneumonia." *AMIA ...Annual Symposium proceedings / AMIA Symposium.AMIA Symposium*, pp. 759–763, 2005.

[17] E. Crain, D. Bulas, P. Bijur, and H. Goldman, "Is a chest radiograph necessary in the evaluation of every febrile infant less than 8 weeks of age?" *Pediatrics*, pp. 88(4):821–824, 1991.

[18] M. Ebell, "Clinical diagnosis of pneumonia in children," *Point-of-Care Guides*, vol. 82, p. 2, 2010.

[19] T. Lynch, R. Platt, S. Gouin, C. Larson, and Y. Patenaude, "Can we predict which children with clinically suspected pneumonia will have the presence of focal infiltrates on chest radiographs?" *Pediatrics*, pp. 113(3 pt 1):e186–e189, 2004.

[20] J. Young, M. Modat, M. J. Cardoso, J. Ashburner, and S. Ourselin, "Classification of alzheimer's disease patients and controls with gaussian processes," in *Proceedings - International Symposium on Biomedical Imaging*, 2012, pp. 1523–1526, cited By (since 1996):2.

[21] Q. H. Ye, L. X. Qin, M. Forgues, P. He, J. W. Kim, A. C. Peng, R. Simon, Y. Li, A. I. Robles, Y. Chen, Z. C. Ma, Z. Q. Wu, S. L. Ye, Y. K. Liu, Z. Y. Tang, and X. W. Wang, "Predicting hepatitis b virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning," *Nature medicine*, vol. 9, no. 4, pp. 416–423, 2003.

[22] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89–109, 2001.

[23] D. Agranoff, D. Fernandez-Reyes, M. C. Papadopoulos, S. A. Rojas, M. Herbster, A. Loosemore, E. Tarelli, J. Sheldon, A. Schwenk, R. Pollok, C. F. Rayner, and S. Krishna, "Identification of diagnostic

markers for tuberculosis by proteomic fingerprinting of serum," *Lancet*, vol. 368, no. 9540, pp. 1012–1021, 2006.

[24] H. Huang, R. C. Ideh, E. Gitau, M. L. Thezenas, M. Jallow, B. Ebruke, O. Chimah, C. Oluwalana, H. Karanja, G. Mackenzie, R. A. Adegbola, D. Kwiatkowski, B. M. Kessler, J. A. Berkley, S. R. Howie, and C. Casals-Pascual, "Discovery and validation of biomarkers to guide clinical management of pneumonia in african children," *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, vol. 58, no. 12, pp. 1707–1715, Jun 2014.

[25] WHO. (Accessed 28 May 2015) Pocket book of hospital care for children:second edition. 2013 guidelines for the management of common childhood illnesses. [Online]. Available: http://www.who.int/maternal_child_adolescent/documents/child_hospital_care/en/

[26] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "Review: A gentle introduction to imputation of missing values," *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[27] L. V. der Maaten, "Visualizing data using t-sne," *Journal of Machine Learning Research*, p. 25792605, 2008.

[28] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology*, p. 24:417441, 1933.

[29] W. Torgerson, "Multidimensional scaling i: Theory and method." *Psychometrika*, p. 17:401419, 1952.

[30] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[31] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[32] C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[33] U. R. Abeyratne, V. Swarnkar, R. Triasih, and A. Setyati, "Cough sound analysis - a new tool for diagnosing pneumonia," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2013, pp. 5216–5219.