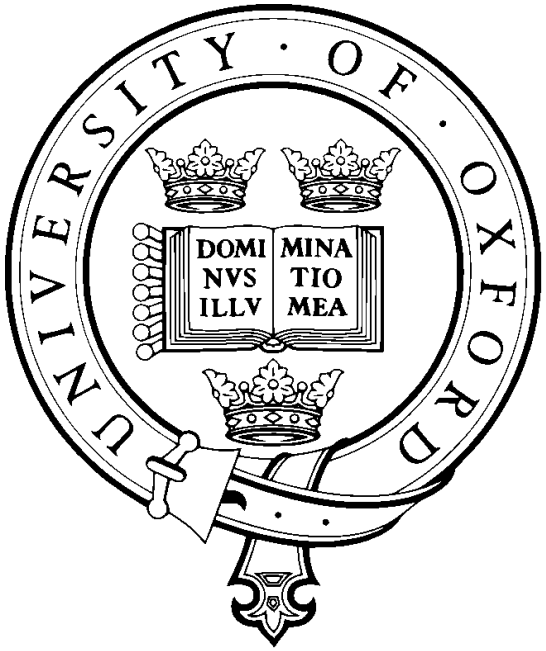
Machine learning for childhood pneumonia diagnosis



Elina Naydenova

University College

University of Oxford

Supervised by:

Professor Maarten De Vos

A thesis submitted for the degree of Doctor of Philosophy

Hillary 2018

Acknowledgements

"It is not the critic who counts; not the man who points out how the strong man stumbles, or where the doer of deeds could have done them better. The credit belongs to the man who is actually in the arena, whose face is marred by dust and sweat and blood; who strives valiantly; who errs, who comes short again and again, because there is no effort without error and shortcoming; but who does actually strive to do the deeds; who knows great enthusiasms, the great devotions; who spends himself in a worthy cause; who at the best knows in the end the triumph of high achievement, and who at the worst, if he fails, at least fails while daring greatly, so that his place shall never be with those cold and timid souls who neither know victory nor defeat."

During this DPhil, I had the unique privilege of going on a journey that was aligned with my pursuit for excellence but more importantly with my values and purpose to use science for social impact. None of this would have been possible without the incredibly people I met on this journey who gave me purpose, courage, opportunity, belief, and truth.

PURPOSE: In 2014, I was lucky to secure an internship at the World Health Organisation where I worked for Adriana Velazquez Berumen. Adriana gave me the most valuable gift you can give a young mind hungry for challenges – purpose. She introduced me to the challenges of childhood pneumonia and her passion & pain for this global health injustice became mine. Proving that we can address the shameful annual mortality from this treatable disease became my purpose and mission. For that and for her ongoing mentorship, I will be forever grateful.

COURAGE: My obsession with childhood pneumonia took me back to Oxford and made me unreasonable. Unreasonable in believing I could build a DPhil project around my newly discovered purpose. Unreasonable in believing I could find a supervisor willing to mentor me, clinical collaborators who would work with me, and funding for the field studies I knew were necessary. Luckily, at that time I met the remarkable Dr Pamela Hartigan who celebrated, cultivated and led unreasonable people (as beautifully described in her book, the Power of Unreasonable People). I had the immense privilege to work for her on the Social Innovation in Health Initiative in 2015/2016 and get to know many incredible health innovators transforming lives around the globe. Pamela & the Skoll Centre also provided an initial grant to seed my pneumonia research. But beyond that, Pamela's boundless energy and passion for socially transformative business and innovation infused me with courage, imagination and determination. I was devastated by Pamela's passing in 2016 and will forever cherish the time I got to learn from her.

OPPORTUNITY: Armed with purpose & courage, I spent months trying to find anyone who would take a chance on my unreasonable idea. This project would not have been possible if it hadn't been for Prof Maarten De Vos who bravely decided to take chance on me. Maarten was new to Oxford and took a risk in believing in my crazy idea. Throughout my time at Oxford he provided the perfect balance of support and nurture on one hand, and intellectual freedom on the other. I learned from both his technical guidance as well as the opportunity to be courageous and think outside the box. I will be forever grateful for Maarten's trust, his open-mindedness and his selfless support.

Along this journey, I was lucky to work with people who shared my belief in the project and generously provided support. Dr Athanasios Tsanas taught me so much about machine learning and made me a better scientist. Dr Jeremy Hull & Dr Climent Casals-Pascual taught me about medicine, making me a better biomedical engineer. Prof William Checkley and Dr Michael Seear generously provided datasets for analysis in this project. The teams from Apnalaya (Ninad & Arun) & the Sion Hospital (Dr Jadhav, Manasi, Dr Zope, Nilesha) in India worked tirelessly to deliver the clinical study in this project and I am incredibly grateful for all the learnings I gained about global health by watching them do their incredible work. None of these many activities would have been possible without Jo Armitage who kept the CDT programme running and provided the logistical support that glued everything together.

NURTURE: I am grateful to my college advisor, Dr Stephen Collins, without whom I would have not been able to navigate the Oxford University environment. Steve was always there to listen, advise and help resolve challenges. I will always remember our college lunches and be grateful for his selfless support and nurturing. I was also very lucky to meet Sarah McCloughry who through her professional training and mentorship made me a more confident and self-aware presenter, engineer and woman. Without Sarah, I would not have been able to beat many moments of self-doubt.

BELIEF: I owe an enormous amount to Prof Adrian Wilson. Adrian was my supervisor during my undergraduate days at Warwick University and was the first person to inspire me to want to become a biomedical engineer. I will forever remember the moment when in my second year of undergraduate Adrian told me I am good enough to be a biomedical engineer and do a PhD at a top university; it is his belief that carried me through many years to come. This included the clinical study in this project, when Adrian selflessly joined me in India to help fire-fight all the challenges that came our way. I am forever grateful to him.

FRIENDSHIP: During these four years, I grew alongside some amazing people whom I shared adventures with. I will always remember my first trip to India with Arvind and Johanna and be grateful to Arvind for introducing me to a new world full of life, energy and spirit that I fell in love with – India. For that I also thank Green Templeton College, Dr Georgina Murphy and the Centre for the Study of Social Change for supporting and collaborating on our research in Mumbai's slums.

TRUTH: Finally, the foundation of my success is my family's sacrifice and love. My parents put my dreams above their needs, even when my choices were beyond their comprehension. The opportunities I have had the privilege to take advantage of are rooted in my parent's hard work and enormous potential, lost in times of economic misfortune. I'm also grateful to my second family – Hattie & Keith – who relentlessly supported me as one of their own over the last 9 years. I will not be who are I am today without my best friend, my partner in life, my climbing partner with whom I face any mountain ahead of us, Adam, who continues to see me for the best I could be and expect nothing less. My truth is forever rooted in our dedication to turn the blessings of our lives into building blocks for a more just and equitable world.

Abstract

Pneumonia is the number one killer of children under the age of 5, causing more deaths than malaria, tuberculosis and HIV/AIDS combined. In 2015, over 920,000 children died of pneumonia and more than 95% of the incidence and 99% of subsequent mortality occurred in developing countries. Current gold standard diagnostic assessment of childhood pneumonia relies on the use of advanced tools (such as X-rays and blood culture) by a clinical expert who assesses and interprets a combination of clinical measurements. However, the shortage of clinical expertise and equipment in low-resource settings delay diagnosis, increasing the risk of mortality. This thesis investigates the role that machine learning could play in improving diagnosis of childhood pneumonia in low-resource settings.

Point-of-care devices such as a digital stethoscope and a pulse oximeter have become more affordable and hence abundant across the globe. The adoption of such devices, particularly in community settings, presents an opportunity to leverage information captured through their signals to provide early diagnosis for conditions such as childhood pneumonia. Hence, this thesis proposes a machine learning framework for the design of diagnostic models, built on a parsimonious set of symptoms which can be captured in a point-of-care setting, to deliver accurate and reproducible diagnosis of pneumonia. The approach is evaluated on three different datasets (with accuracy ranging between 81.8 – 97.9%), consistently outperforming diagnosis with the current gold standard guidelines for community diagnosis by the World Health Organisation.

A clinical study in Mumbai, India, is designed to capture a realistic dataset in a poor urban community and a resource-constrained hospital, where a mobile health toolkit including three point-of-care devices (a digital stethoscope, a pulse oximeter and a thermometer) is built to enable minimally trained users to capture essential health information. Diagnostic models developed on this rich dataset are seen to perform well when validated internally as well as generalise to other datasets from different geographies.

Finally, a set of tools for automated analysis of lung signals is proposed and validated on two big datasets, one from Peru and one from India. Combined with vital signs, information derived through this automated lung signal analysis was seen to outperform diagnosis via expert clinical annotation, delivering 81.3% sensitivity and 84.5% specificity.

Contents

List of figures

List of tables

List of abbreviations

ALRI = acute lower respiratory-tract infection

AR modelling = autoregressive modelling

ARI = acute respiratory infection

ASHA = Accredited Social Health Activists

AYUSH = ayurveda, unani, siddha, naturopathy, or homoeopathy

CHW = community health worker

DNN = deep neural network

EMD = empirical mode decomposition

EN = elastic net

GMM = gaussian mixture model

GSO = Gram-Schmidt orthogonalization

Hib = Haemophilus influenzae type b

HMM = hidden Markov model

HR = heart rate

IDW = initial deflection width

IMCI = Integrated Management of Childhood Illness

LASSO = least angle shrinkage and selection operator

Lcn2 = lipocalin

LDW = largest deflection width

mRMR, = minimum redundancy maximum relevance

NICE = National Institute for Health and Care Excellence

Osat = oxygen saturation

PCV = pneumococcal conjugate vaccine

RF = random forest

RR = respiratory rate

sLDA = sparse linear discriminant analysis

SVM = support vector machine

T = temperature

WHO = World Health Organisation

This page is intentionally left blank.

1. The Challenge of Childhood Pneumonia: Background & Motivation

1.1. Motivation

Pneumonia is a highly prevalent acute respiratory infection (ARI); this infectious disease affects the lungs and can be caused by a range of agents. Despite the biological complexity of childhood pneumonia, strategies for prevention, diagnosis and treatment are well documented and mostly effective in resource-rich settings. For example, Haemophilus influenzae type b (Hib) and pneumococcal vaccines can be used to prevent pneumonia directly, and measles and pertussis vaccines can be used to restrict sources of secondary infection (Madhi et al. 2008, UNICEF 2006). Furthermore, appropriate nutrition for pregnant women and infants, and limited exposure to cooking and tobacco smoke can further limit risk of contracting the disease (César et al. 1999, Caulfield et al. 2004, Smith et al. 2011). X-rays in combination with blood tests, auscultation (lung examination via a stethoscope) and respiratory rate count, are typically used as diagnostic mechanisms (Cherian et al. 2005, Rasmussen, Pio, and Enarson 2000). Treatment involves antibiotics in bacterial cases, and oxygen therapy is used in severe cases (Bari et al. 2011, Enarson et al. 2008). Given this entourage of tools, the disease should only have a limited impact on children (Bhutta et al. 2013). **However, childhood pneumonia remains the biggest killer of children under five years, accounting for 15% of childhood deaths globally** (Liu et al. 2016). More than 95% of the childhood pneumonia cases and 99% of subsequent deaths occur in developing countries (Rudan et al. 2013, Ghimire, Bhattacharya, and Narain 2012). Timely and accurate diagnosis that facilitates appropriate treatment has been reported to have the potential to reduce mortality by as much as 42% (UNICEF 2006). Yet, the shortage of clinical expertise and advanced equipment in low-resource settings prohibit the early identification of disease. The goal of this thesis is to develop and validate better tools for diagnosis of childhood pneumonia in low-resource settings, leveraging the power of modern technological advancements such as point-of-care devices and machine learning and signal processing algorithms.

1.2. Childhood pneumonia – problem dimensions

Before introducing the specific research objectives and work structure of this thesis, this chapter begins by outlining the challenges associated with childhood pneumonia and identifying gaps in the current approach to managing the disease. First, incidence and prevalence rates globally are investigated, highlighting major trends. Next, current approaches to diagnosis and treatment are introduced, focusing on key challenges in clinical practice, especially in low-resource settings. Finally, a set of research questions and objectives, emerging from this analysis are introduced, motivating the structure of the thesis.

1.2.1. Incidence and prevalence rates

Pneumonia is the number one killer of children under the age of 5, causing more deaths than malaria, tuberculosis and HIV/AIDS combined (Qazi et al. 2015, Rudan et al. 2013, WHO 2017) – see Figure 1.1[1]. Pneumonia is one of the most severe respiratory conditions; a range of pathogens (bacteria, viruses or sometimes fungus) can initiate an inflammation in the lungs leading to the disease. Consequently, the alveoli in the lungs (the microscopic air sacks where oxygen and carbon dioxide exchange occurs) get filled up with pathogen-secreted fluid and pus, affecting oxygen access to the bloodstream (Ali et al. 2013). Since young children have weaker immune systems, they are more susceptible to such severe respiratory infections (Smith et al. 2011). In 2015, over 920,000 children died of pneumonia; this equates to a child death every 35 seconds (WHO 2017).

More than 95% of the childhood pneumonia cases and 99% of subsequent deaths occur in developing countries (Rudan et al. 2013, Ghimire, Bhattacharya, and Narain 2012). Appropriate diagnostic assessment of childhood pneumonia typically relies on the use of advanced tools (such as X-rays and blood culture) by a clinical expert who assesses and interprets a combination of clinical measurements; the details of diagnosis are described in section 1.2.2. (Graham et al. 2008, Marsh et al. 2008). However, access to high-quality healthcare is often limited in many low- and middle-income countries (LMICs)

[1] The terms pneumonia, ARI and acute lower respiratory-tract infection (ALRI) are often used interchangeably in the global health space, even though the latter is a broader term encompassing other respiratory conditions such as lung abscess and acute bronchitis. Since most epidemiological data reports on ALRI, the two terms will be used interchangeably in this chapter.

due to a shortage of appropriate medical equipment and clinical expertise. Most childhood pneumonia deaths are reported to occur in a relatively early stage of disease progression and complications can develop quickly (Smith et al. 2011). In resource-constrained settings, hospital facilities are often distant and so community health workers (CHWs) need to differentiate between patients who can be managed locally (non-severe cases) and those in need of urgent referral (severe cases). Thus, the severity of disease is crucial for determining appropriate management.

Before we take a closer look at more specific incidence rates, prevalence rates, and symptom severity, it should be noted that accurate estimates of the pneumonia burden, globally, regionally and nationally, are very difficult to derive (Rudan et al. 2013). This is due to the complexity of studies required. Specifically, longitudinal community-based studies are required, ideally spanning across a full year to account for the seasonality of disease (Lanata et al. 2004, Rudan et al. 2004). Screening needs to be regular, all-encompassing and systematic (home visits no more than 2 weeks apart) to avoid underestimation due to recall bias, especially in larger families (Lanata et al. 2004). Yet such studies in LMICs are very rare due to their resource-intensiveness. Additionally, the choice of diagnostic definition of pneumonia, and the accuracy and consistency of its implementation, influence outcomes substantially (Campbell, Biloglav, and Rudan 2008). Variability also exists across diagnostic assessors, especially if they operate in different settings (e.g. hospital clinicians versus community doctors) (Lanata et al. 2004).

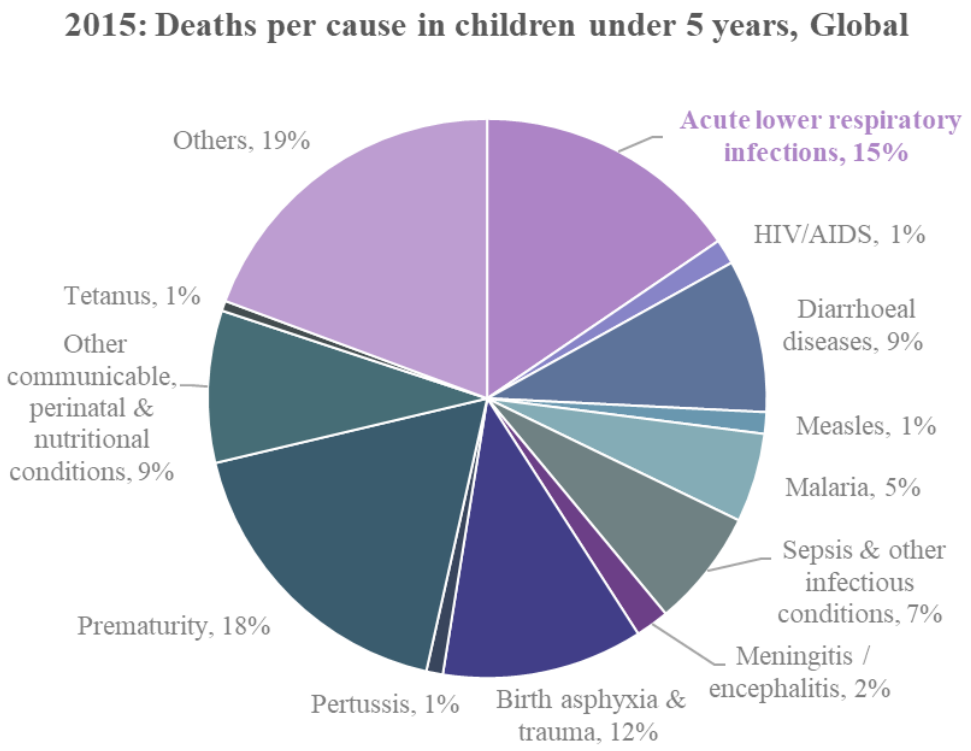


Figure 1.1: Distribution of childhood deaths across the main causes of mortality. Acute lower respiratory infection (ALRI) should be considered synonymous with pneumonia here (see footnote 1 in Section 1.2.1). Data from (WHO 2017).

The effects of these challenges have been previously demonstrated in a review of studies reporting on childhood pneumonia incidence rates (Rudan et al. 2004). From the 28 studies which satisfied the minimum quality criteria, incidence rates per child year varied by a factor of 100, which partly reflects the natural heterogeneity of disease but also highlights inconsistencies in the data collection process (Rudan et al. 2004).

Building upon this complexity, estimates of severity incidence (i.e. pneumonia cases in need of hospitalisation or referral for treatment) tend to be even more problematic. Such estimates cannot be based on reported hospitalisation rates as these already encompass a huge degree of bias due to differences in (1) care-seeking behaviour amongst parents, (2) access to hospital care, (3) admission criteria applied in any given hospital (Lanata et al. 2004). The World Health Organisation (WHO) has provided definitions for the severity of pneumonia (further discussed in Section 1.2.2.2), designed to primarily identify children in need of clinical attention (evaluation based on these definitions is sensitive but not very specific). Consequently, reliable estimates of the severity of childhood pneumonia in the community are rare (Rudan et al. 2004).

With regard to these challenges, data from the WHO Global Repository and a review conducted by Ruben et al. shed some light on incidence, severity and mortality rates and their distribution across various geographical regions (WHO 2017, Rudan et al. 2013). Deaths due to pneumonia represented between 10%-16% or mortality amongst children under five in 2015, across the different WHO regions (see Figure 1.2). Regional differences appear to be related to income level, with pneumonia representing a smaller fraction of mortality in Europe and the Americas (10% and 11% respectively) than other regions. In 2010, there were an estimated 120 million cases of acute lower respiratory-tract infection (pneumonia) amongst children; from those, 14 million were severe, requiring hospitalisation (12%), and 1.4 million (2%) were fatal (Rudan et al. 2013). However, the ratios between incidence, severity and mortality are different in different regions of the world, as illustrated in Figure 1.3. From this, pneumonia incidence was highest in Africa and South-Asia, and lowest in the Americas and Europe (this was the case even with data adjusted for immunisation against *H influenzae* type b and *S pneumoniae*). In fact, Southeast Asia accounted for 39% of the severe episodes and 33% of the mortality and Africa accounted for 30% of the severe episodes and 44% of the mortality, although the two regions only accounted for 28% and 21% of the child population (Rudan et al. 2013). Additionally, severity rates appeared to be substantially higher in Europe compared to all other regions; this is likely to be from under-diagnosis and under-reporting of severity in countries with weaker health systems. Finally, mortality was lowest in the Americas and highest in Africa. Access to high-quality treatment facilities is likely to be the driving factor behind this statistic. Yet, mortality amongst pneumonia cases appeared to be surprisingly high in Europe, higher than Southeast Asia and comparable to Eastern Mediterranean. This, combined with the high incidence of severe pneumonia in the region, could be related to the inability of the health system to identify sick children early and will be further discussed later in this chapter as well as subsequent chapters.

Going a step deeper than regional geographies, the incidence of pneumonia is further concentrated within certain countries. Fifteen countries account for 65% of total pneumonia episodes and 64% of severe episodes: Afghanistan, Angola, Burkina Faso, China, Democratic Republic of the Congo, Ethiopia, India, Indonesia, Kenya, Mali, Niger, Nigeria, Pakistan, Tanzania, and Uganda (Walker et al. 2013). India is the country with the highest disease burden, with 35.4 million new cases annually

amongst children under five (Rudan et al. 2013). Between 10% and 20% of episodes are reported to be severe (Acharya et al. 2003, Broor et al. 2007, Gladstone et al. 2008, Rudan et al. 2013). Mortality is also high – 303 per 100,000 children under five - amounting to more than 388,000 deaths in 2010. In comparison, China's mortality rate in 2010 was 53 per 100,000 children and a total of 43,000 deaths.

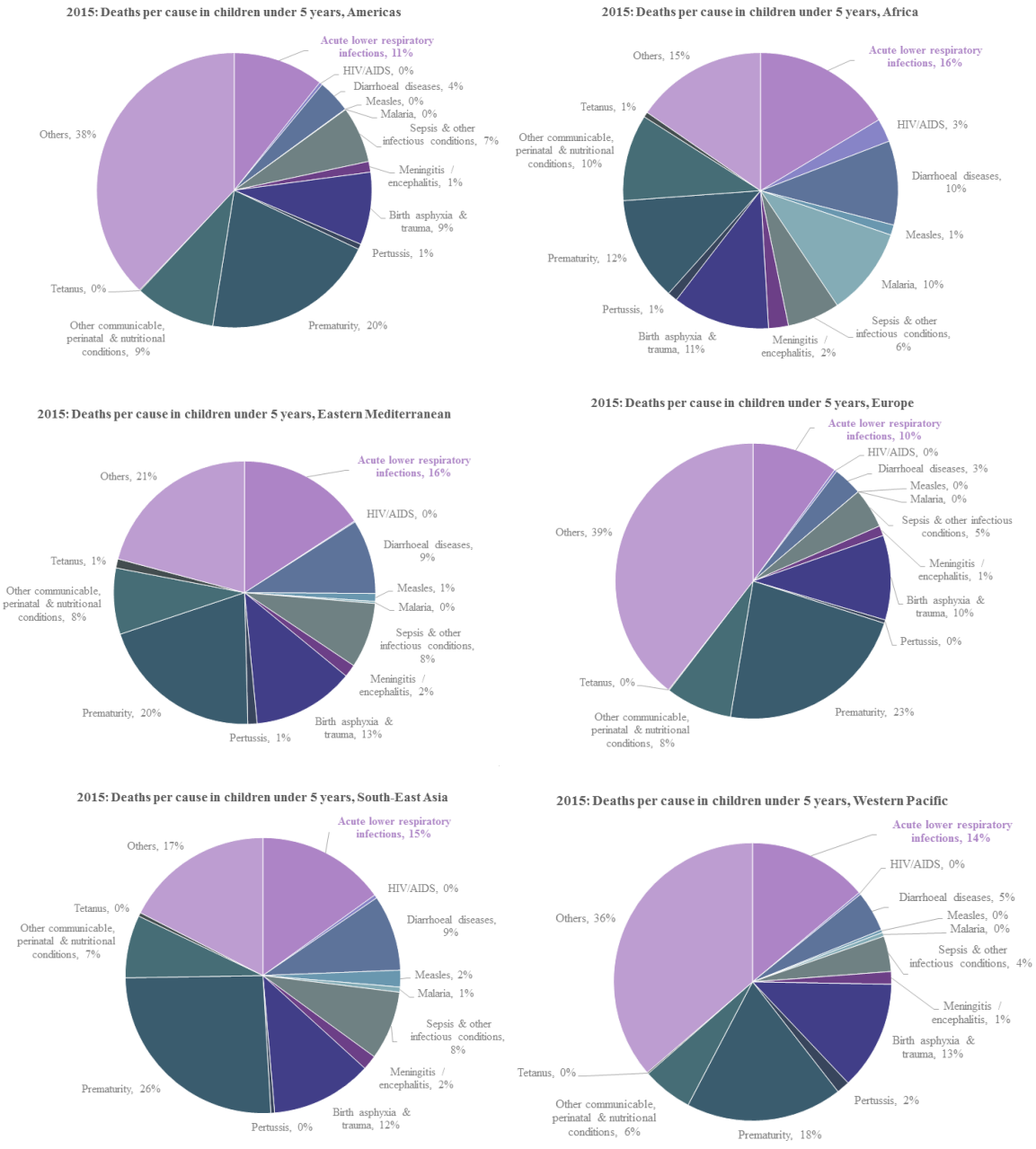


Figure 1.2: Distribution of childhood deaths across the main causes of mortality, for each WHO region. Acute lower respiratory infection (ALRI) should be considered synonymous with pneumonia here (see footnote 1 in Section 1.1.1. Incidence and prevalence rates). Data from (WHO 2017).

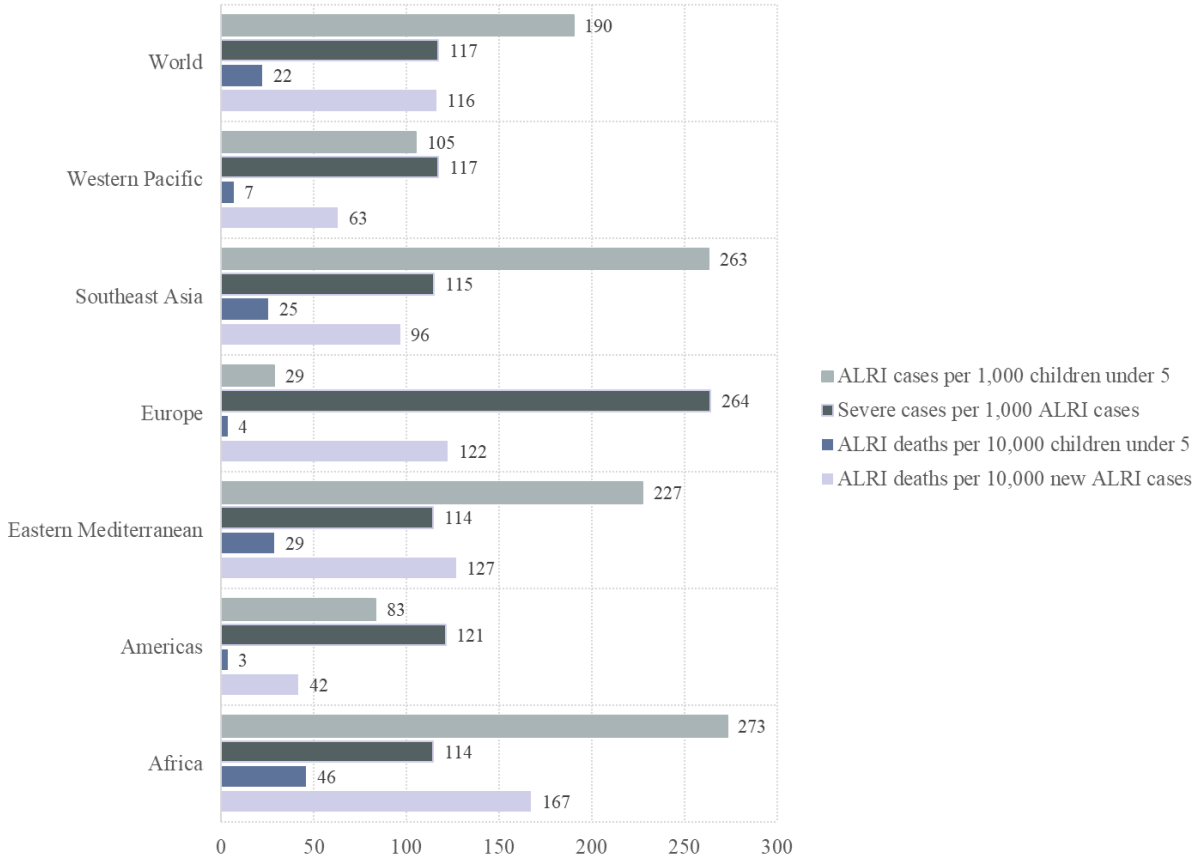


Figure 1.3: Incidence, severity and mortality rates globally and across the WHO regions in 2010. Raw data from supplementary material in (Rudan et al. 2013).

A number of large scale global health interventions, including targets formally defined by the Millennium Development Goals and the implementation of the WHO guidelines for Integrated Management of Child Illness (described in detail in Section 1.2.2.2), have driven a large reduction in mortality amongst children under five, over the last 15 years, as illustrated in Figure 1.4. However, mortality reduction has been more effective for some diseases than others. Additionally, mortality reduction has also not been consistent across the globe. Table 1.1 provides a detailed comparison of rates of mortality reduction across diseases and geographic regions. From this, diseases such as measles, malaria and diarrhoea have experienced higher rates of reduction than pneumonia. The African and Eastern Mediterranean regions have experienced lower rates of reduction than other regions. In a number of countries, progress has been very slow which, coupled with high birth rates, has resulted in an increase in the total number of deaths due to pneumonia. For example, in Burkina Faso, the pneumonia mortality rate decreased between 2000 and 2010 but was outnumbered by population growth, leading to growth

in childhood pneumonia deaths (17,389 to 21,763) (Walker et al. 2013). Similar outcomes have been recorded in Afghanistan, Cameroon, Chad, Democratic Republic of the Congo, and Mali (Walker et al. 2013).

In addition to geographical differences, the burden of childhood pneumonia has been suggested to also exhibit age and gender differences. Namely, younger children are disproportionately affected: 81% of deaths are said to occur in children younger than two years (Walker et al. 2013). The incidence of pneumonia is less age-dependent than mortality (Walker et al. 2013). Gender differences in incidence (higher amongst boys) and mortality have been suggested in some studies but the data are sparse, with the largest differences recorded in South Asia (Walker et al. 2013, Nair et al. 2013).

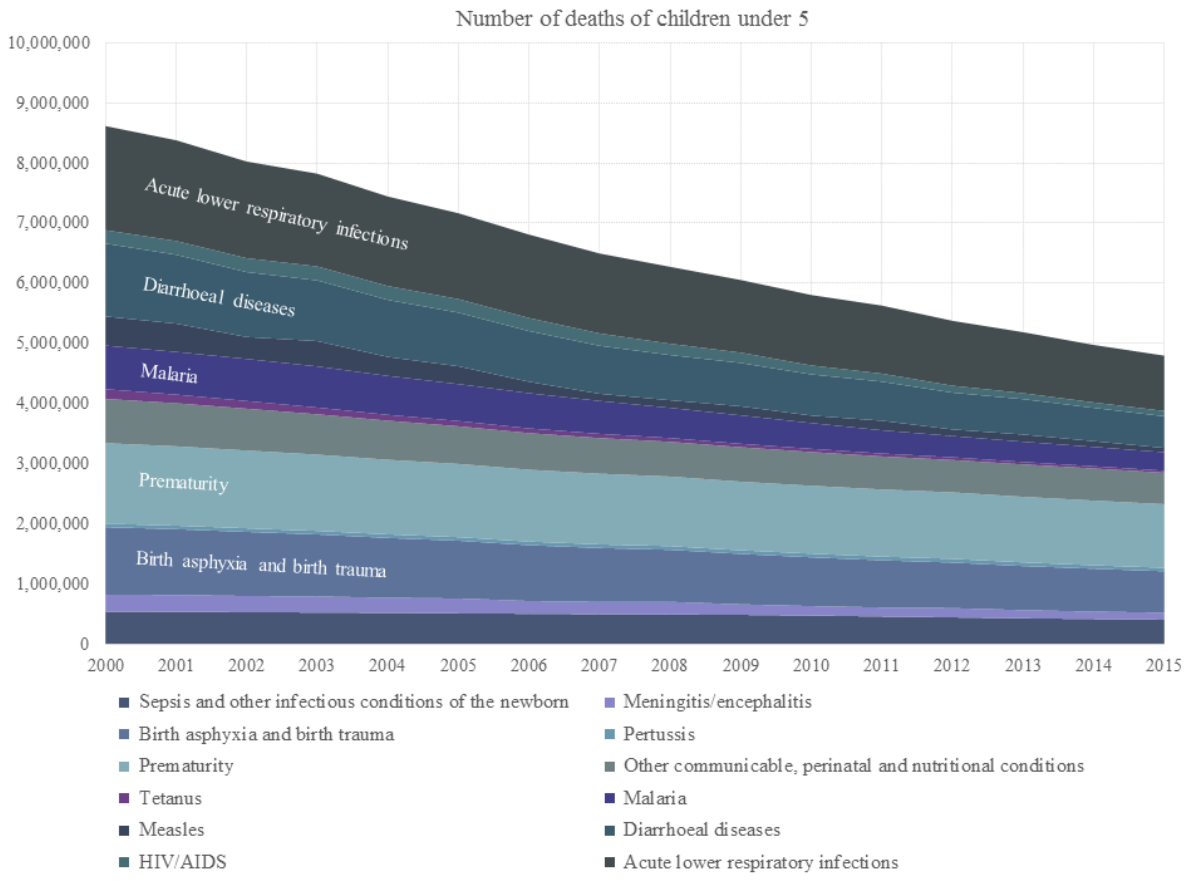


Figure 1.4: Number of childhood deaths, caused by some of the major disease contributors, between 2000 and 2015. Data source (WHO 2017).

	World	Africa Region	Region of the Americas	South-East Asia Region	European Region	Eastern Mediterranean Region	Western Pacific Region
Total deaths	-0.39	-0.29	-0.46	-0.52	-0.44	-0.19	-0.59
Acute lower respiratory infections	**-0.47**	**-0.28**	**-0.58**	**-0.63**	**-0.59**	**-0.35**	**-0.71**
HIV/AIDS	-0.61	-0.63	-0.79	-0.30	-0.54	0.81	-0.49
Diarrhoeal diseases	-0.57	-0.49	-0.75	-0.67	-0.64	-0.44	-0.72
Measles	-0.85	-0.89	-0.50	-0.73	-0.38	-0.82	-0.88
Malaria	-0.58	-0.58	-0.65	-0.49	-1.00	-0.59	-0.68
Sepsis and other infectious conditions of the newborn	-0.25	0.11	-0.39	-0.50	-0.35	-0.03	-0.37
Meningitis/encephalitis	-0.59	-0.52	-0.68	-0.70	-0.68	-0.55	-0.64
Birth asphyxia and birth trauma	-0.38	-0.11	-0.45	-0.57	-0.38	-0.21	-0.68
Pertussis	-0.04	0.05	-0.02	-0.24	0.11	-0.12	-0.08
Prematurity	-0.21	-0.03	-0.45	-0.28	-0.44	0.03	-0.52
Other communicable, perinatal and nutritional conditions	-0.29	0.07	-0.48	-0.61	-0.46	-0.06	-0.54
Tetanus	-0.79	-0.70	-0.70	-0.90	-0.70	-0.71	-0.89

Table 1.1: Rates of change of childhood mortality between 2000 and 2015, across a range of diseases and WHO regions. Data source (WHO 2017).

1.2.2. What is childhood pneumonia? – the status quo on diagnosis and treatment

1.2.2.1. Aetiology

Pneumonia can be caused by a bacterial or viral agent, as well as fungus in rare circumstances (Rudan et al. 2013). Pneumonia is the one paediatric respiratory infection for which a microbiological diagnosis is most difficult to determine (McCracken 2000a). Only a third of pneumonia cases can be attributed to a certain aetiology via culture, antigen detection and clinically available serological techniques (McCracken 2000b). For example, blood cultures have been reported to deliver evidence of pathogens in only 10-15% of bacterial pneumonia cases amongst children (McCracken 2000a). Depending on the depth of laboratory testing, the microbial cause of pneumonia has been reported to be identifiable in 20-60% of cases (Ruuskanen and Mertsola 1999). A large ongoing project, led by John Hopkins University and funded by the Gates Foundation, the pneumonia etiology research for child health (PERCH), tries

to collect better data to elucidate insights on the aetiology of pneumonia across the globe (Levine et al. 2012).

Nevertheless, some facts regarding the aetiology of pneumonia are well known. The most common bacterial agents that cause childhood pneumonia are *Streptococcus pneumoniae* and *Haemophilus influenzae type b (Hib)*; the most common viral agent is respiratory syncytial virus (McCracken 2000b). These pathogens can infect the child either through air droplets (e.g. from cough) or through contaminated blood. A number of studies have investigated the aetiology of pneumonia amongst children, reporting significant differences with age (Table 1.2 and Table 1.3) (Juven et al. 2000, Heiskanen-Kosma et al. 1998, Wubbel et al. 1999). In particular, the burden of viral pneumonia has been seen to be highest amongst younger children. A 3-year study from Finland found that a microbial cause was identifiable in 85% of the cases; from those, 62% were viral, 53% bacterial and 30% concomitant viral-bacterial (Juven et al. 2000). The most common pathogens were *S. pneumoniae* (37%), respiratory syncytial virus (29%) and rhinovirus (24%) (Juven et al. 2000). The latter is not typically associated with lower respiratory tract infections in children but sometimes leads to predisposition to superinfection. The incidence of viral causes was reported to decrease with age, whilst that of bacterial causes remained similar (Juven et al. 2000).

Both the pneumococcal conjugate vaccine (PCV) and the pentavalent vaccine (*Hib* protection) are commonly administered in developed countries and are becoming more available in developing countries too (IHME 2014). However, even with full vaccine coverage, pneumonia caused by other pathogens will continue to present a challenge.

Age (years)	**N**	**Viral (% of N)**	**Bacterial (% of N)**	**Mixed (% of N)**	**All (% of N)**
<2	108	80	47	34	93
2-5	84	58	56	33	81
>5	62	62	58	19	76
All	254	62	53	30	85

Table 1.2: Aetiology of pneumonia stratified per age. Source of data (Juven et al. 2000). N denotes the total number of pneumonia cases in the corresponding age bracket.

		Bacteria (% of N)			Viral (% of N)
	Age	*Streptococcus pneumoniae*	*Mycoplasma pneumoniae*	*Chlamydia pneumoniae*	
Heiskanen-Kosma	0-4	24	4	1	37
Wubbel	0-4	33	6	3	28
Heiskanen-Kosma	5-9	36	30	13	21
Wubbel	5-9	14	7	9	10
Heiskanen-Kosma	10-16	31	51	35	4
Wubbel	10-16	29	14	14	0

Table 1.3: Aetiology of pneumonia stratified by age, from two studies - (Heiskanen-Kosma et al. 1998) and (Wubbel et al. 1999). N denotes the total number of pneumonia cases in the corresponding age bracket.

1.2.2.2. Diagnosis

Timely and accurate diagnosis that facilitates appropriate treatment has been reported to have the potential to reduce mortality by as much as 42% (UNICEF 2006). However, diagnosis can often be challenging because the clinical manifestation of pneumonia in children is variable (Shah et al. 2017). Often, the disease will start with an upper respiratory tract infection, manifested through minor fever and rhinorrhoea (McCracken 2000a). These symptoms will be followed by tachypnoea, cough and higher fever. In addition, a number of lung sounds can accompany the disease (Chapter 6 discusses in detail the significance of lung sounds identified during auscultation to pneumonia diagnosis). The presence of crackles (also known as rales), although rare, is suggestive of pneumonia (McCracken 2000a). Wheezing, on the other hand, is said to be found in viral or mycoplasmal pneumonia (Ruuskanen and Mertsola 1999). Further details on crackles and wheezes are provided in Chapters 3 & 6.

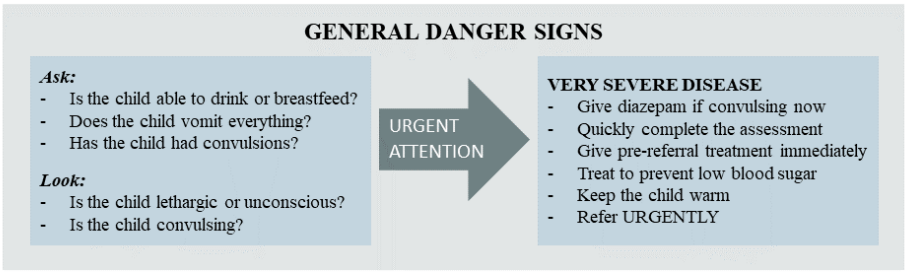
In the 1980s, WHO developed a case-management strategy to address the staggering number of childhood pneumonia deaths. This became the foundation of the acute respiratory infection (ARI) programme, later included in the integrated management of childhood illness (IMCI) – a comprehensive set of guidelines for both community and hospital management (WHO/UNICEF 2005). A number of assumptions underpinned this strategy (Graham et al. 2008):

- The overwhelming majority of acute respiratory infection deaths were due to pneumonia;
- Children with pneumonia required assessment by a well-trained health worker;

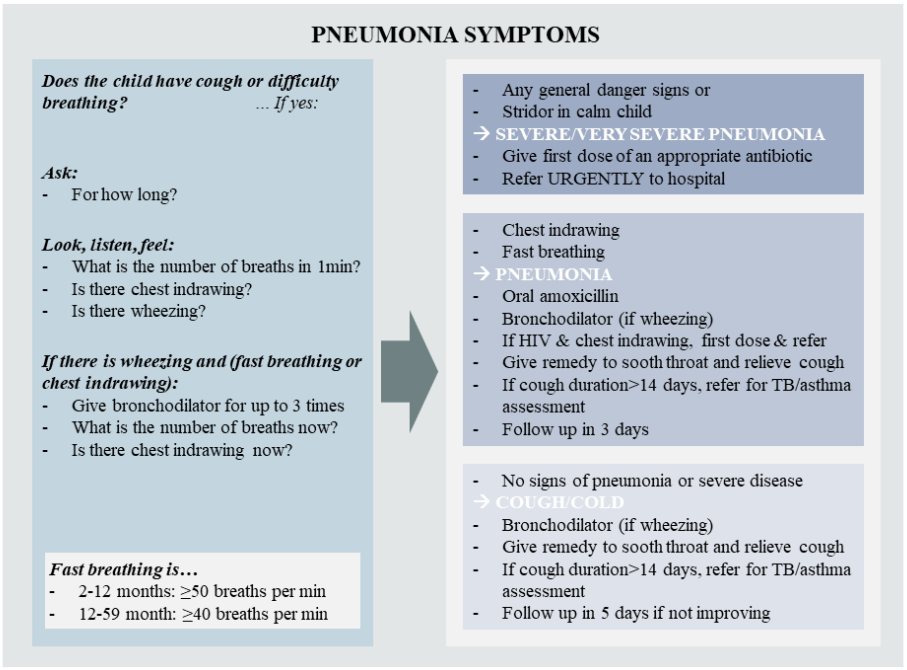
- Clinical signs, such as respiratory rate and chest indrawing, can be used to differentiate between pneumonia and other respiratory conditions (Shann 1995);
- The cause for many pneumonia deaths was bacteria - Streptococcus pneumoniae or Haemophilus influenzae (Shann 1995);
- Antibiotics should not be administered to children with a cough who do not have pneumonia to avoid causing antimicrobial resistance;
- Hypoxemia in children with pneumonia is indicative of an increased mortality risk (Duke, Mgone, and Frank 2001, Lozano 2001).

Figure 1.5 summarises the diagnostic process for childhood pneumonia according to the IMCI.

(a)



(b)



(c)

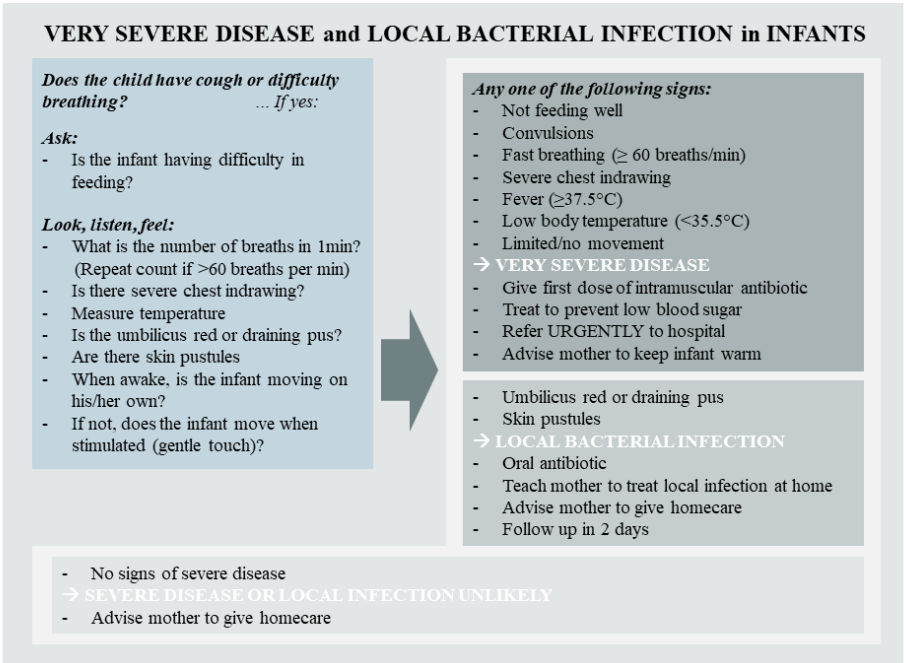


Figure 1.5: IMCI guidelines in the context of childhood pneumonia (WHO 2014b). (a) identifying general danger signs; (b) identifying pneumonia; (c) infants as a special risk group.

Beyond the WHO guidelines for identification of pneumonia in the community, some countries have their own guidelines for identifying pneumonia in clinical practice. The National Institute for Health and Care Excellence (NICE) in the UK has issued guidelines that recommend a combination of clinical characteristics in order to identify pneumonia (NICE 2013). Namely, pneumonia is likely if a child exhibits the following symptoms:

- tachypnoea (≥ 60 breaths per minute, age 0–5 months; ≥ 50 breaths per minute, age 6–12 months; ≥ 40 breaths per minute, age older than 12 months);
- crackles in the chest;
- nasal flaring;
- chest indrawing;
- cyanosis;
- low oxygen saturation.

Similarly, the British Thoracic Society guidelines define severe community acquired pneumonia using the following physiological parameters:

- temperature >38.5°C;
- respiratory rate >70 breaths/min in infants and >50 breaths/min in older children;
- moderate to severe recession or severe difficulty breathing;
- nasal flaring, grunting or apnoea;
- cyanosis or oxygen saturation <92%;
- significant tachycardia for age range;
- capillary refill time of ≥2 seconds.

Since the clinical manifestations of pneumonia often overlap with a number of other respiratory conditions, chest X-rays are often used to confirm diagnosis and localise lung consolidation. Hospitalised children who do not respond to initial antibiotic therapy are likely to be subjected to more invasive tests. Since children do not produce sufficient sputum for collection, nasopharyngeal mucus is used for antigen detection and culture for viral agent detection (McCracken 2000a). However, a virus

detected in the upper respiratory tract does not automatically imply pneumonia, as it could have created a predisposition for a bacterial superinfection (McCracken 2000a).

Alongside the issue of pneumonia identification and aetiology determination, classifying the severity of disease is both very important and challenging. The severity of disease is a crucial determinant of treatment approach – e.g. home-based treatment versus hospitalisation (more details provided in the treatment Section below). Studies have demonstrated that severity correlates with case-fatality rate (Pépin et al. 2001, McNally et al. 2007, Berkley et al. 2005). Although the majority of pneumonia cases are non-severe (from Figure 1.3, on average 117 out of 1,000 pneumonia cases are severe), the majority of mortality occurs amongst children with severe cases (Graham et al. 2008). The IMCI guidelines provide a mechanism for differentiation between severe and non-severe pneumonia, aimed at triaging cases in need of referral. Early warning scores have been developed to stratify severity in a hospital setting and are discussed in detail in Section 2.3.2.

1.2.2.3. Treatment

Since the aetiology of pneumonia is rarely known, treatment is often empirical (McCracken 2000a). Clinical practice related to antibiotic administration varies – some members of the clinical community advocate antibiotic therapy in all pneumonia cases, whilst others recommend withholding antibiotics in mild cases of pneumonia that are likely to be of a viral origin (Ruuskanen and Mertsola 1999, WHO 2010, Trautner et al. 2006). A number of antibiotics (e.g. amoxicillin, chloramphenicol) are administered orally, or intravenously in severe cases which require hospitalisation; the latter are often administered oxygen too (Ruuskanen and Mertsola 1999). Enarson et al. have published an example of the clinical protocol recommended for case management of pneumonia (see summary in Figure 1.6) (Enarson et al. 2008). A study by Wubbel et al. documented the use of a range of antibiotics (azithromycin, erythromycin estolate or amoxicillin/clavulanate), recording recovery in 143 out of the 147 patients. Yet, many of these cases are likely to have had viral disease (Wubbel et al. 1999).

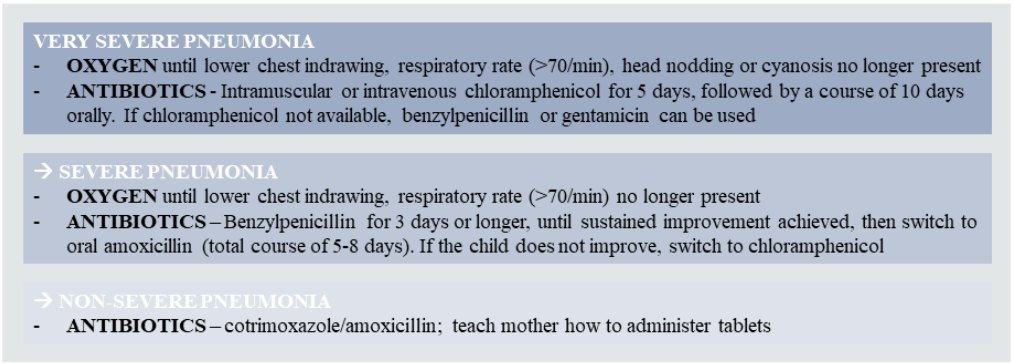


Figure 1.6: Case management protocol for children with pneumonia; note that the three categories match diagnostic criteria listed in the WHO IMCI guidelines (colours are kept consistent). Source (Enarson et al. 2008).

Similarly, the WHO recommends initiation of empiric antibiotic therapy based on clinical sign examination (Bryce, Victora, and Advisors 2005). This approach has been reported to reduce pneumonia-related morality amongst children under 5 by 35-40% (Sazawal, Black, and Group 2003). However, there are two main challenges associated with this approach: (1) treatment failure and (2) antimicrobial resistance fuelled by the over-use of antibiotics.

Jain et al. conducted a study, involving 181 children, which investigated predictors of treatment failure (Jain, Sarathi, and Jawalekar 2013). From the study cohort, 37 children (20.4%) experienced treatment failure at 48 hours, with 9 deaths. Reasons for treatment failure included: persistent worsening of lower chest indrawing and tachypnoea; hypoxemia (SpO2<90%); empyema/pneumothorax; lung abscess; meningitis. The study identified six factors at baseline which were associated with treatment failure 48 hours later: (i) infancy; (ii) measles immunisation not given by 9 months; (iii) severe malnutrition; (iv) very fast breathing; (v) hypoxemia; and (vi) bacteremia. Other studies have also highlighted the significance of malnutrition, fast breathing and hypoxemia at baseline as predictors of treatment failure (Rudan et al. 2008, Addo-Yobo et al. 2004, Fu et al. 2006). Additionally, prior antibiotic use, alternative diagnosis (e.g. malaria), immunosuppression (HIV) and antibiotic resistance have all been documented to lead to treatment failure (Graham et al. 2008). Most fatal cases of treatment failure have been recorded in infants, making this age group particularly high-risk (Graham et al. 2008). Interestingly, duration of illness has not been found to be a significant predictor of treatment failure. Finally, measles should be explicitly mentioned as a factor for pneumonia-associated mortality – 56%-86% of measles deaths are due to pneumonia (Duke and Mgone 2003).

1.2.3. Risk factors & challenges in low-resource settings

A number of barriers prevent children from receiving treatment for pneumonia. Under-resourced and over-crowded hospitals in developing countries often struggle to admit all children in need of care (Ashraf et al. 2008). Furthermore, access to care is further hindered by parents' inability to reach distant hospitals and limited financial resources (Ashraf et al. 2008, Ashraf et al. 2007). Hence, community-based management of pneumonia cases is essential. Successful delivery models in low-resource settings include day-care facility-based primary care services for childhood pneumonia, shown to deliver good health outcomes and to be cost-effective as part of a randomised controlled trial in Bangladesh (Ashraf et al. 2010). However, such community-based interventions are only possible on the back of early and accurate diagnosis of the disease that informs appropriate treatment course.

Large-scale studies in Bangladesh and Nepal have assessed the ability of CHWs to diagnose pneumonia after undergoing training (Ghimire, Bhattacharya, and Narain 2012, Hadi 2003). Intensive training and routine supervision were shown to deliver better diagnostic and treatment performance by the CHWs (Hadi 2003). Reduction in pneumonia incidence and severity rates were recorded between 2004 and 2006 across a number of districts in Nepal, where CHWs were trained to identify and treat the disease (Ghimire, Bhattacharya, and Narain 2012). Community-case management of pneumonia has been implemented in a number of countries, including a nationwide programme in the Gambia (WHO/UNICEF 2005). Similarly, in Honduras, community volunteers across 1800 communities deliver pneumonia treatment as part of their integrated community child care programme (WHO/UNICEF 2005). Other such examples come from Malawi, Ethiopia and Pakistan (Nsona et al. 2012, WHO 2014c).

The incidence of pneumonia has been linked to a number of risk factors, which disproportionately affect children in developing countries, including: stunting and underweight, lack of immunisation, insufficient breastfeeding (not exclusively in infants under six months), and indoor smoke and pollution from tobacco and household use of solid fuels (Rudan et al. 2008, Roth et al. 2008, Gakidou et al. 2007, Sazawal, Black, and Group 2003, Victora et al. 1999, Edejer et al. 2005). The latter has been reported to increase the risk of pneumonia by as much as 80% (Walker et al. 2013). Households in rural and poor urban communities across Southeast Asia often use traditional open ovens, fuelled by wood or dried

animal dung, for cooking and heating (Ghimire, Bhattacharya, and Narain 2012). A study in Nepal found that there is a statistically strong association between hours spent near the stove and the number of episodes of severe pneumonia in young children (Pandey et al. 1989). A further risk factor for both the incidence and health outcome from childhood pneumonia is maternal/parental lack of education (Tiewsoh et al. 2009). Parents' ability to recognise key danger signs is crucial for the early identification of the disease (Tiewsoh et al. 2009).

Malnutrition has been described as the underlining cause of death for 61% of all deaths of children under 5, and 52% of those with pneumonia (Caulfield et al. 2004, Rice et al. 2000, Bryce et al. 2005, Loeb and High 2005, Nannan et al. 2007). Nearly two-thirds of malnourished children who require hospitalisation have been diagnosed with pneumonia (Shimeles and Lulseged 1994, Ahmed et al. 1999). According to a recent review, children with pneumonia, who also suffer from moderate or severe malnourishment, are at higher risk of death (Chisti et al. 2009). For moderate malnutrition, the relative risk ranged between 1.2-36.5 and for severe malnutrition between 2.9-121.2 (Chisti et al. 2009). It has been suggested that the diagnostic criteria for childhood pneumonia require further specification for children with malnutrition. Namely, a study in the Gambia indicated that the respiratory rate cut-off between well-nourished and malnourished children differ by as much as 5 breaths per minute, linked to the lower body temperature that malnourished children exhibit (Falade et al. 1995). Similarly, lower chest indrawing has been found to be less common, and intercostal indrawing more common, in malnourished children (Falade et al. 1995, Chisti et al. 2009). Consequently, the use of lower chest indrawing for pneumonia diagnosis in severely malnourished children has been questioned (Chisti et al. 2009, Falade et al. 1995). Furthermore, it has been suggested that the use of classical clinical signs for pneumonia diagnosis is likely to underestimate the burden of pneumonia in malnourished children, due to reduced inflammatory responses and weakened respiratory muscles (Chisti et al. 2009, Suskind 1988).

1.3. Research aims of the thesis

This chapter defines the dimensions of the challenge of childhood pneumonia in a global health context, highlighting current diagnostic and treatment practice as well as barriers to health access in low-resource settings. Based on this analysis, novel solutions that bring evidence-based, accurate and robust clinical assessment to communities living in resource-constrained settings are urgently needed. Such solutions should address the challenges of reproducibility and user-variability encountered when observational guidelines such as IMCI are implemented at scale. Moreover, the specificity of such solutions should be much higher than that of existing tools, at no detriment to sensitivity, to rein in the unnecessary use of antibiotics and subsequent rampant antimicrobial resistance. In this context, this thesis has two main objectives: (1) develop tools for automated interpretation of clinical signals from point-of-care tools (e.g. stethoscope signals) that would enable minimally trained users to capture high quality clinical information in low-resource communities; (2) develop machine learning tools for fusion of a range of clinical symptoms, derived from the point-of-care device signals, into evidence-based diagnosis of childhood pneumonia. Consequently, the following research questions are investigated:

- Could machine learning of objective clinical characteristics (e.g. vital signs), rather than subjective observational health markers (e.g. nasal flaring), improve the diagnosis of childhood pneumonia in the community, enabling minimally trained users to identify the disease early?
- Could machine learning be used to derive data-driven stratification of severity of pneumonia, informing appropriate case management and treatment?
- Do differences in incidence of malnutrition necessitate the development of more specific diagnostic algorithms?
- How do machine learning models for diagnosis of pneumonia generalise across populations? Are models transferable across community and hospital populations, across geographies and settings?
- Can signal processing and machine learning enable the capture of important clinical insights, such as lung pathologies derived from auscultation, in the community?
- Are there new biomarkers that can be derived from data routinely collected using point-of-care tools to further improve the diagnosis of pneumonia?

1.4. Thesis structure

This thesis investigates the research questions and objectives defined above according to the following structure:

- Chapter 2 reviews existing approaches to improving the diagnosis of childhood pneumonia, covering device innovations, decision support tools such as data-driven clinical rules and early warning scores, and methods for computerised lung signal analysis.
- Chapter 3 introduces the existing datasets which were used for analysis in this thesis, including datasets of clinical characteristics and datasets of lung sound recordings, and highlights gaps in the available data which necessitated the design and implementation of a field study.
- Chapter 4 motivates the design of a field study in Mumbai, India, describes its implementation in the context of the local health system and presents details of the dataset collected.
- Chapter 5 introduces the theoretical principles behind a range of signal processing and machine learning techniques relevant to the analysis conducted in future chapters.
- Chapter 6 proposes a machine learning framework for the tri-fold diagnosis of pneumonia, including identification of disease, evaluation of severity and determination of aetiology. Results on a large dataset are reported, including assessment of generalisability on unseen data from the same population.
- Chapter 7 first investigates differences in the manifestation of pneumonia across populations. Using these insights, the chapter explores the issue of model generalisability across populations, highlighting two approaches to ensuring classification performance is transferable across different populations.
- Chapter 8 proposes a framework for the automated analysis of digital lung signals, which combined with the diagnostic models presented in Chapters 6 & 7 is seen to further improve diagnostic performance, whilst also removing the need for expert clinical evaluation of the lung signal measurements.
- Chapter 9 is a discussion of the findings presented in this thesis, highlighting research contributions as well as limitations and opportunities for improvement in future research.

1.5. Research publications & presentations

Publications

Naydenova, E., A. Tsanas, S. Howie, C. Casals-Pascual, M. De Vos, 2016. "Smart Diagnostic Algorithms for Automated Detection of Childhood Pneumonia in Resource-constrained Settings." *Proceedings of the Fifth IEEE Global Humanitarian Technology Conference GHTC* 377-384

Naydenova, E., A. Tsanas, S. Howie, C. Casals-Pascual, M. De Vos, 2016. "The power of data mining in diagnosis of childhood pneumonia." *Journal of the Royal Society Interface* 13 (120) DOI: 10.1098/rsif.2016.0266

Naydenova, E., Raghu, A., Ernst, J., Sahariah, S.A., Gandhi, M., Murphy, G. 2017. "Healthcare choices in Mumbai slums: A cross-sectional study." *Wellcome Open Research* 2:115 DOI: 10.12688/wellcomeopenres.13127.1 (under review)

Presentations

Naydenova, E., Automating the Diagnosis of Childhood Pneumonia, 3rd World Health Organisation Global Forum on Medical Devices, May 2017, Geneva, Switzerland

This page is intentionally left blank

2. Diagnostic Improvements for Childhood Pneumonia – a Literature Review

2.1. Introduction

Chapter 1 introduced the challenge of childhood pneumonia, quantifying its burden on children and health systems worldwide. Access to timely and accurate diagnosis was seen to be essential for the prevention of hospitalisation and mortality. Therefore, this chapter reviews existing approaches to developing solutions for childhood pneumonia diagnosis, including new devices, data-driven algorithms for interpretation of symptoms and existing attempts for computerised lung sound analysis. The latter aims to remove barriers to access to a powerful diagnostic tool for respiratory diseases – auscultation. Finally, this chapter highlights the gaps in existing approaches where this thesis aims to offer an impactful contribution.

2.2. Diagnostic innovations

There are a number of diagnostic innovations in the context of childhood pneumonia that have been recently developed, or are currently under development, as summarised in Table 2.1. The majority of these innovations are focused on improving individual measurements, e.g. better ways to record respiratory rate or better pulse oximeters. However, as illustrated in Chapter 1, there is no single clinical characteristic that is sufficient for reliable diagnosis of childhood pneumonia. Additionally, these product innovations do not resolve the challenge of balancing the need for a range of clinical measurements to identify pneumonia, on the one hand, and the shortage of clinical expertise, especially in low-resource settings, on the other.

Innovation Name	Description
ARI Counting Beads	Designed as a simple & cheap way to help community health workers count breaths. The beads are colour-coded to ease diagnosis. Should be used in conjunction with an ARI timer.
INSPIRE	An automated device that wraps around the child's chest and uses a sensor (accelerometery) to detect respiratory rate.
mPneumonia	An Android mobile application which automates the IMCI protocol, paired with a respiratory rate counter (based on tapping) and a pulse oximeter (Nonin 8000AP & Nonin Xpod).
Respiratory Rate Mobile application (*University of British Columbia, Prof. Ansermino & team*)	A mobile application that enables community health workers to press any button on the phone to register a breath; the mobile application then calculates a mean respiratory rate.
Solar & Cell Phone Pulse	A pulse oximeter device (measures oxygen saturation and heart rate) powered by old mobile phone batteries and charged through solar cells.
CliniCloud	A digital stethoscope that connects to a mobile phone and records data that can be assessed by doctors remotely.
ThinkLabs & Ekuore	Digital stethoscopes that can record lung signals onto a mobile phone.
The Nossal 'Smart Cable' Project	A set of sensors (oximetry, respiratory rate, thermometer) that connect to a device that enables data transfer onto a mobile phone.
The Phone Oximeter for Sepsis	A pulse oximeter that attaches to a mobile phone.
ThinkMD	A mobile phone platform that enables community health workers to collects a number of observational clinical characteristics. Diagnosis is based on empirical clinical rules, defined by a group of clinicians.

Table 2.1: Pneumonia-related innovations; sources (UNICEF 2013) and personal research.

2.3. Data-driven approaches for pneumonia diagnosis

Section 2.2 briefly reviewed some of the most relevant product innovations, developed to improve the measurements of individual (or a group) of symptoms (mostly vital signs). In addition to product innovations, some data-driven approaches have been proposed by the scientific community as tools of combining individual symptoms (that could be measured by these novel point-of-care tools) into pneumonia diagnosis and disease severity assessment.

2.3.1. Diagnosing pneumonia

Several decision rules have been proposed as tools for predicting childhood pneumonia on the basis of clinical characteristics. This section summarises most such studies and discusses potential issues that

hinder their implementation in clinical practice at scale. Table 2.2 includes a summary of the diagnostic performance of the range of approaches proposed.

For example, (Lynch et al. 2004) conducted a study with 570 children (1-16 years old) that identified four independent predictors of pneumonia: tachypnoea, fever, decreased breath sounds, and crackles. The diagnostic value of individual clinical characteristics was observed to be insufficient, with some of them being very sensitive and unspecific (e.g. cough – 88% sensitivity & 16% specificity) and others the reverse (e.g. grunting – 2% sensitivity & 100% specificity). A logistic regression model with fever, crackles and tachypnoea was recorded to deliver best amongst a range of combinations of clinical features; yet, sensitivity was high (97%) but specificity poor (15%). A full list of the diagnostic performance of individual clinical characteristics and logistic regression models of combinations is listed in Table 2.2. The gold standard used in the (Lynch et al. 2004) study was radiography evaluation by three experts. Children who presented at the emergency department and were evaluated by a clinician to require radiography were included in the study, with the exception of those who were under 1 year, or with asthma or critically ill. Matthews et al. conducted a similar study that also illustrated the inability of individual clinical characteristics to reliably identify pneumonia (see Table 2.2) (Mathews et al. 2009). Recursive partitioning of temperature (≥38°C) and oxygen saturation (<92%) was reported to produce 19% sensitivity and 95% specificity. The gold standard diagnosis of pneumonia was determined retrospectively for a cohort of 526 cases of which 26 had pneumonia (5% incidence). Mahabee-Gittens further expanded this analysis to investigate the diagnostic accuracy of a logistic regression model that combines tachypnoea and oxygen saturation, nasal flaring, alongside age (Mahabee-Gittens et al. 2005). The authors reported sensitivity of 20% and specificity of 98% (further details are listed in Table 2.2). This affirms findings from an older study that observed that absence of tachypnoea, nasal flaring, grunting, retractions, rales, and decreased breath sounds, eliminates any likelihood of pneumonia (100% specificity) (Leventhal 1982). Again, radiography-assessed pneumonia was used as a gold standard; the pneumonia incidence rate in the study was 9%. Bleeker et al. demonstrated that the addition of lab tests to clinical characteristics improves diagnostic performance, increasing the ROC area from 0.75 to 0.83 (Bleeker et al. 2001).

These studies provide some preliminary indication of the types of clinical features that are relevant to the diagnosis of pneumonia. However, there are a number of challenges. First, in both the study by Lynch et al. (Lynch et al. 2004) and the study by Mathews et al. (Mathews et al. 2009) there is already some pre-selection due to the recruitment protocol as only children who required an X-ray, as per the doctor's evaluation, were included in the study. For example, in the study by Lynch et al. (Lynch et al. 2004), cough, fever and crackles were part of the inclusion criteria. Hence, the results these studies present are not generalisable to a larger population of children who might have pneumonia. Second, validation of these models is poor, and it has not been reported how they perform in a completely independent clinical setting (Ebell 2010). A review by Maguire et al. has also questioned if many of these studies are statistically powered (most of them have populations of <500 cases) (Maguire et al. 2011). Additionally, it is unclear whether the choice of setting (secondary/tertiary hospital, primary care, community case) affects performance of the models. Related to this, the incidence of pneumonia across studies is variable; the authors have not proposed how this should be accounted for if the methods were to be generalised to other populations (e.g. resource-constrained settings).

The first issue has been examined in more detail. Namely, a number of studies have demonstrated that the incidence of occult pneumonia (i.e. pneumonia in febrile children who exhibit no signs of respiratory distress, tachypnoea or hypoxia) is not negligible. For example, Murphy et al. found this incidence to be 5.3% in a population of 1,084 children; Shah et al reported 6.8% of the population (Murphy et al. 2007, Shah et al. 2010). A similar study by Bachur et al. suggested an even higher incidence of occult pneumonia – 26% of febrile children with none of the traditional clinical characteristics indicative of pneumonia except positive radiographs (Bachur, Perry, and Harper 1999). Such cases would automatically be excluded by the studies described earlier which rely on a doctor's screening of classical clinical characteristics before an X-ray is administered. A number of additional characteristics, such as duration of cough, duration of fever and white blood cell count, have been suggested as beneficial for identification of occult pneumonia (Table 2.2) (Murphy et al. 2007).

Alongside the issue of under-detection of pneumonia (occult pneumonia), studies have also documented a challenge with over-diagnosis on the basis of traditional clinical characteristics. For example, (Neuman

et al. 2011) demonstrated that amongst children with a history of fever and crackles, who would typically be administered antibiotics, only 25% had radiographic pneumonia. Craig et al. conducted a large study (15,781 children younger than 5 years) to evaluate practices by which children with fever, who might be at risk of severe bacterial infection (urinary, pneumonia, bacteremia), are diagnosed and treated (Craig et al. 2010). The study found that 20% of children without bacterial infection were wrongly prescribed antibiotics; yet a large proportion of the severe cases were missed by early physician examination (only 10% - 50% sensitivity). Hence, the use of simple clinical characteristics for identification of pneumonia, avoiding radiation from radiography, should be carefully balanced against the risk of misdiagnosis, causing unnecessary use of antibiotics.

The issue of validation and generalisability has been tackled by Nijman et al. who developed a model that assesses the risk of pneumonia versus other serious bacterial infections or no infection (Nijman et al. 2013). A logistic regression model including oxygen saturation (<94%), body temperature, duration of fever, tachypnoea and C-reactive protein was developed and cross-validated on two Dutch populations. The model was then tested and validated on a third population from the UK (see results in Table 2.3). The model did not deliver as good a performance when applied to the third population: e.g. at 10% risk threshold, the pneumonia model gained sensitivity (from 60% in the derivation population to 75% in the validation) but dropped substantially in specificity (from 84% in the derivation population to 69% in the validation population). These differences in performance could be driven by population differences (i.e. differences in distribution of individual clinical characteristics and differences in incidence rates), as well as technical challenges (e.g. it is unclear how the authors dealt with normalising individual features across datasets). No studies have investigated how such models, developed on hospital-based studies in developed countries, generalise to (1) hospital care in developing countries or (2) community-based care in developed or developing countries.

Study	Clinical characteristic	Sensitivity	Specificity
(Lynch,2004)	Fever	47	68
	Fever history	92	20
	Tachypnoea	13	95
	Tachycardia	51	60
	Cough	88	16
	Crackles	43	27
	Wheeze	4	97
	Decreased breath sounds	54	55
	Bronchial sounds	7	96
	Grunting	2	100
	Fever + decreased breath sounds	96	11
	Fever + crackles	96	16
	Fever + tachypnoea	93	19
	Fever + decreased breath sounds + crackles	97	6
	Fever + decreased breath sounds + tachypnoea	95	8
	Fever + crackles + tachypnoea	97	15
	Fever + decreased breath sounds + crackles + tachypnoea	98	8
(Mathews,2009)	Tachypnoea	27	71
	Fever		
	≥38.0°C	50	75
	≥38.5°C	35	86
	≥39.0°C	19	94
	Duration of fever	23	87
	Fever history	81	42
	Oxygen saturation		
	<96%	38	70
	<94%	19	87
	<92%	15	95
	<90%	8	99
	Crackles	27	72
	Wheeze	15	89
	Decreased breath sounds	12	87
	Grunting	0	94
	Fever (<38°C) + Oxygen saturation (<92%)	19	95
(Mahabee-Gittens,2005)	Age (>12 months)	66	57
	Tachypnoea		
	≥40 per min	77	43
	≥50 per min	50	71
	≥60 per min	32	88
	≥70 per min	7	97
	Oxygen saturation		
	<96%	63	77
	<94%	42	88
	<92%	26	93
	<90%	12	96
	Nasal flaring	23	92
	Crackles	21	86
	Wheeze	21	84
	Decreased breath sounds	11	95
	Grunting	2	96
	Age (>12 months) + oxygen saturation (≤96%)	41	91
	Tachypnoea (≥50 per min) + oxygen saturation (≤96%)	34	92
	Age (>12 months) + tachypnoea (≥50 per min)	25	93
	Age (>12 months) + tachypnoea (≥50 per min) + oxygen saturation (≤96%)	18	97
	Tachypnoea (≥50 per min) + oxygen saturation (≤96%) + nasal flaring	20	98
(Murphy,2007)	Cough	96	23
	Duration of cough		
	>1	79	22
	>3	49	60
	>5	42	74
	>10	27	88
	Nasal congestion	65	30
	Ill appearance	2	97
	Fever		
	>39°C	71	36
	>40°C	32	71
	Duration of fever		
	>1	69	45
	>3	42	74
	>5	27	88
	WBC count (per mm3)		
	>15,000	74	58
	>20,000	61	72
	Wheeze	32	48
	Crackles	40	23
	Decreased breath sounds	32	89
	Respiratory distress	39	60
	Tachypnoea	72	29

Table 2.2: Diagnostic performance of individual clinical characteristics and models across studies. Data sources (Murphy et al. 2007, Mahabee-Gittens et al. 2005, Mathews et al. 2009)

Outcomes & risk thresholds	Sensitivity (95% CI)		Specificity (95% CI)	
	Derivation population	*Validation population*	*Derivation population*	*Validation population*
Pneumonia				
≥2.5%	90 (85 - 96)	99 (91 - 100)	40 (36 - 45)	14 (10 - 18)
≥5%	78 (71 - 86)	91 (81 - 100)	64 (61 - 67)	38 (33 - 44)
≥10%	60 (50 - 71)	75 (61 - 90)	84 (82 - 85)	69 (63 - 74)
≥15%	47 (37 - 56)	63 (47 - 78)	92 (90 - 93)	82 (78 - 87)
≥30%	16 (10 - 22)	34 (21 - 48)	99 (98 - 99)	95 (92 - 98)
Other serious bacterial infections				
≥2.5%	90 (84 - 96)	80 (68 - 93)	56 (53 - 60)	33 (28 - 39)
≥5%	82 (74 - 89)	74 (61 - 87)	72 (70 - 74)	45 (40 - 51)
≥10%	70 (61 - 79)	70 (57 - 83)	85 (84 - 87)	63 (57 - 68)
≥15%	55 (46 - 64)	62 (48 - 77)	91 (90 - 93)	72 (67 - 77)
≥30%	19 (12 - 26)	42 (28 - 55)	98 (97 - 99)	90 (86 - 93)

Table 2.3: Performance of diagnostic models for identification of pneumonia and other serious bacterial infections at different risk thresholds (Nijman et al. 2013).

2.3.2. Identifying severe disease

A number of research groups have focused on severity, suggesting potential methods for identifying children with severe disease, rather than focusing on the specific diagnosis of pneumonia. Paediatric early warning systems have been proposed on the basis of a combination of physiological parameters (Egdell, Finlay, and Pedley 2008, Haines, Perrott, and Weir 2006, Duncan, Hutchison, and Parshuram 2006, Tibballs et al. 2005, Monaghan 2005, Van den Bruel et al. 2007). These have primarily focused on predicting illness in hospitalised children. For example, a paediatric advanced warning score (PAWS) has been developed by (Duncan, Hutchison, and Parshuram 2006) combining dynamic information (heart rate, respiratory rate, systolic blood pressure, oxygen saturation, temperature, oxygen and fluid therapy) with medication and demographic information, to identify hospitalised children at risk of cardiopulmonary arrest; the authors reported 78% sensitivity and 95% specificity. Similarly, Egdell et al. have developed a PAWS chart for the assessment of children presenting at the emergency department and validated its performance retrospectively on data of 95 children, achieving 70% sensitivity and 90% specificity (Egdell, Finlay, and Pedley 2008). Thompson et al. have demonstrated that a mixture of vital signs can be used to differentiate between children with severe and mild disease, using the Manchester triage score and NICE guidelines as benchmarks (Thompson et al. 2009). The study cohort consisted of children who were taken to the Paediatric Assessment Unit at the University Hospital Coventry and

Warwickshire NHS Trust because a parent or a general practitioner suspected acute infection. The presence of one or more of the following clinical characteristics was found to differentiate between serious/intermediate and minor/no infection with 80% sensitivity and 39% specificity: fever (≥39°C), oxygen saturation (94%), tachycardia and tachypnoea. Whilst these systems/scores could improve hospital care, they are not directly applicable for a more general population (e.g. primary care) where children at risk of severe disease need to be first identified.

2.4. Lung sound analysis

Section 2.2 highlighted a number of innovations that have digitised the audio signal that can be captured via a stethoscope in order to enable remote recordings that can be sent for evaluation by an expert. Additionally, section 2.3 highlighted the contribution that individual lung sounds, identified from the lung signals by a clinical expert, can have towards the overall diagnosis of pneumonia. Consequently, it has become clear that an objective and standardised approach to interpretation of lung sounds could improve the diagnostic value of auscultation as a means of detection of lung pathologies as well as diagnosing pneumonia. Therefore, computerised analysis of lung sounds has been suggested and explored as a tool for automated classification of acoustic patterns and different respiratory conditions. This research first emerged as a topic in the scientific literature in the early 1980s, initially focusing on identifying specific lung sounds and more recently focusing on developing diagnostic tools for lung sound disorders. The analytical pipeline of digital lung sound signals has two main components: feature identification via signal processing and lung sound classification via machine learning; a number of techniques have been explored by the research community. This section presents an overview of these techniques, whilst also highlighting related challenges and opportunities for improvement.

2.4.1. Signal acquisition process

Sounds from the respiratory system are typically recorded in one of two ways: (1) microphones or (2) contact sensors that can be placed on the chest or mouth. Microphones can be used to record upper airway sounds such as cough or snoring. Lower airway sounds, adventitious or normal breath sounds, can be captured either through microphones (piezoelectric, electret or contact) or contact sensors such as accelerometers (Pasterkamp et al. 1993, Pasterkamp, Kraman, and Wodicka 1997). A range of digital stethoscopes, enabling the recording of the lung signal for analysis, are now commercially available,

including devices developed by Littmann, Ekuore, Thinklabs, Eko Devices. The emergence of such devices has resulted in the collection of lung sound signals that can be used for retrospective labelling by clinical experts and the development of automated algorithms.

2.4.2. Signal processing techniques for feature identification

A number of signal processing techniques have been used in the literature to characterise lung sounds. The subsequent sections introduce these techniques, with references to previous research papers, divided into four main categories: time domain, frequency domain, time-frequency and wavelet domain techniques. The outcomes and limitations of each of the previous attempts to implement these techniques are also highlighted. Moreover, the theoretical framework for the techniques that have been seen to deliver promising results and are relevant to analysis performed in Chapter 7 is introduced in detail in Chapter 5.

2.4.2.1. Time domain

A list of time domain approaches to deriving features from lung signals are presented in Table 2.4. Popular techniques in the time domain include autoregressive modelling (AR modelling) (Alsmadi 2002, Kahya 1999), statistical features and fractal analysis (Mondal 2014). Specific waveform characteristics have been explored in the literature as features for the detection of crackles. Crackles typically start with a width deflection and then include a long and damped sinusoidal wave (Yeginerand 2005, Vannuccini, Rossi, and Pasquali 1998). Murphy at al. have proposed a crackle counter based on the following wave characteristics, illustrated in Figure 2.1 (Murphy, Del Bono, and Davidson 1989):

- The amplitude of the largest peak is at least twice as big as the that of the background noise
- The crackle begins with a sharp positive or negative deflection
- After the initial deflection, zero crossings were increasingly wider

Ayari et al. have further expanded the number of waveform characteristics and used those to characterise the type of crackles (Ayari 2012). These characteristics are illustrated in Figure 2.1 and include:

- Initial deflection width (IDW)
- Two-cycle duration (2CD)
- Largest deflection width (LDW1)

- Largest deflection of the first three left and right neighbours (LDW2, LDW3, LDW4)
- Maximum amplitude of crackle waveform (Amp)
- Cycle duration points between extrema (T1 – T4)

Signal processing	Classifier	References	Sample size	Lung sound or disorder	Outcomes	Limitations
TIME DOMAIN						
Waveform characteristics	'	Murphy RL	100 signals	crackles & wheeze	0.78 correlation between detector and clinician	Sensitive to artefacts
	FC	Ayari, F.	25 signals	pulmonary fibrosis & chronic bronchitis	Fuzzy clustering used to characterise different types of crackles	Only used for crackle morphology classification
AR coefficients	kNN	Alsmadi, S.S.	unclear	healthy & pathological	Online digital processing of lung recordings appears to be possible.	A methodology paper with no sizable train or test set.
	ANN	Martinez-Hernandez, H.G.	29 patients	healthy & pathological	81% sens and 83% spec	Small sample size
	kNN	Sankur, B.	unclear	normal vs adventitious	100% sens and 85.7% spec	Small sample size
	'	Kahya, Y.P.	57 signals	COPD, normal, restrictive pulmonary	67-88% accuracy	Small sample size, substantial variability
	'	Castañeda-Villa, S.	2 simulated & 2 real cases	crackles	No quantification of performance	Small sample size
Statistical characteristics	ELM	Mondal, A.	30 signals	normal & abnormal	86.3% sens and 86.9% spec	Small sample size recorded in highly controlled environment (no noise or artefacts)
Fractal	'	Gnitecki, J.	5 signals	healthy	NA	airflow measurement at mouth, very small sample size
	'	Hadjileontiadis, L.J.	14 signals	crackle & squack	99-100% accuracy	no identification of lungs, only discrimination between different types of crackles

Table 2.4: Signal processing techniques for feature extraction in time domain. Abbreviations: sens = sensitivity, spec = specificity, kNN = k-Nearest Neighbour, ANN = Artificial Neural Networks, AR coefficients = Autoregressive model coefficients; FC = fuzzy clustering; ELM = extreme learning machine.

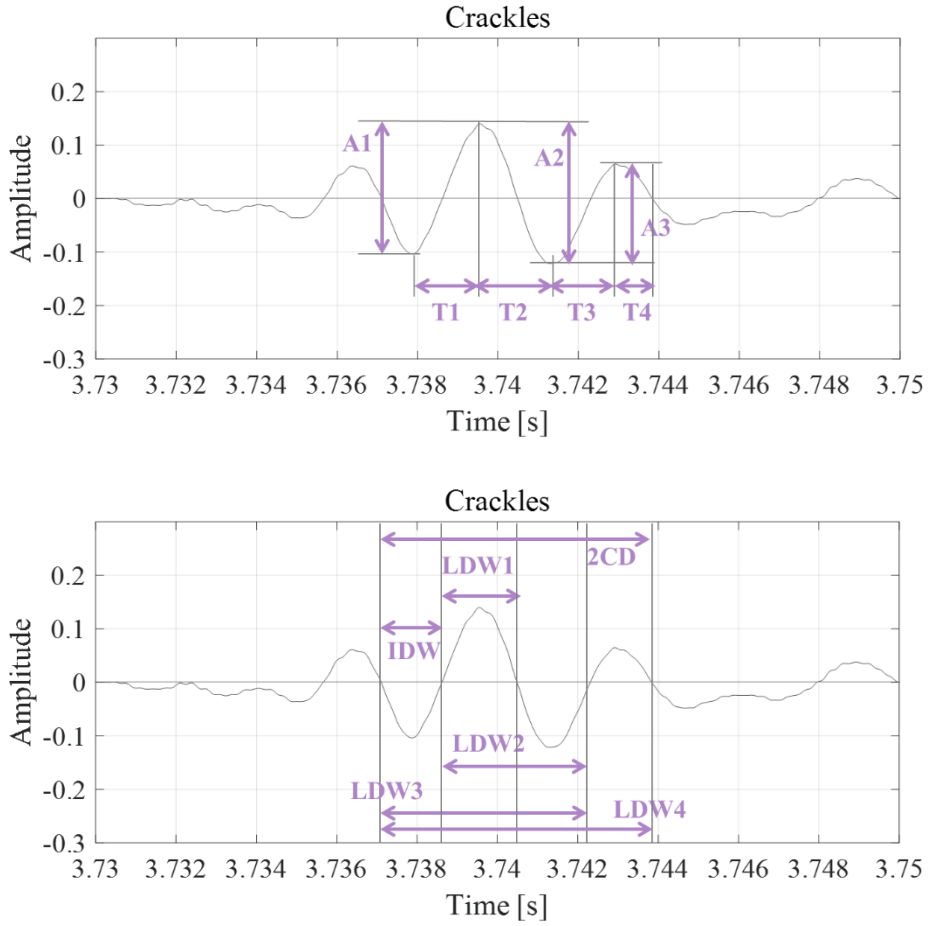


Figure 2.1: Waveform characteristics for crackles.

2.4.2.2. Frequency domain

A number of studies have highlighted that lung signals are primarily non-stationary, making classical frequency analysis unsuitable (Mondal 2014, Sovijärvi 2000). Hence, signal processing in the frequency domain is the least popular technique for lung sound analysis. However, signal windowing has been used to divide the signal into short frames which can be assumed to be semi-stationary. Hence, signal processing techniques in the frequency domain have been applied for lung sounds identification, making use of frequency differences between normal and pathological sounds (Gurung et al. 2011). A list of such approaches is included in Table 2.5.

A number of studies have developed algorithms for automated detection of wheezing based on a threshold defined to identify peaks in the frequency domain that are said to be characteristic of wheezing (Meslier, Charbonneau, and Racineux 1995). The definition of the threshold varies across the literature – from 3 to 15 times greater than the average value (Gurung et al. 2011). Oiu et al. have demonstrated

that frequency analysis alone is insufficient for correct identification of wheezing and have instead proposed a "frequency and duration dependent threshold" algorithm (Qiu et al. 2005). The threshold in their approach is no longer dependent on global power but rather power corresponding to a given frequency range.

A different approach has derived quantile vector from the Fast Fourier Transform of signals along 400ms; where lung sound features are based on the frequency with octile coefficient 0.125, 0.250…, 0.875 (Mayorga et al. 2012). Fourier transform techniques have also been used by Xu et al. and Wang et al. in the context of cepstral analysis (Xu, Cheng, and Wu 1998, Wang 2012).

Signal processing	Classifier	References	Sample size	Lung sound or disorder	Outcomes	Limitations
FREQUENCY DOMAIN						
	ANN	(Waitman 2000)	17 patients	normal vs abnormal	62% sens 85% spec	Small sample size, information leakage between train and test patients
	-	(Guntupalli et al. 2008)	14 signals	wheeze	83% sens 85% spec	multi-sensor mapping device suitable only for research; information leakage between train and test patients
	-	(Ono et al. 2009)	31 signals	fine crackles	75-80.5% AUC	only used for severity determination, not diagnosis
STFT	ANN	(Riella, Nohama, and Maia 2009)	28 signals	wheeze	84.8% accuracy	Small sample size from online database with "ideal sounds"
FFT	GMM	(Mayorga et al. 2012)	21 signals	normal & adventitious	0% equal error rate	Small sample size from online database with "ideal sounds"
PSD	-	(Jane et al. 2004)	23 signals	normal & asthma	"good correlation"	no statistically significant evaluation of performance
DFT	kNN	(Oud, Dooijes, and van der Zee 2000)	16 signals	airway obstruction	60-90% accuracy	large variability in performance, small sample size
MFCC	SVM	(Palaniappan 2013a)	66 signals	normal & disease	90.8% accuracy	Small sample size from online database with "ideal sounds"
	GMM	(Bahoura 2009)	24 signals	normal & wheeze	94.6% sens 91.9% spec	Small sample size from online database with "ideal sounds"
	-	(Ellington et al. 2014)	186 signals	normal & disease	differences in features	no quantification of classification accuracy
	HMM	(Yamashita 2014)	112 signals	normal & disease	83% accuracy	no indication if method can identify specific disease

Table 2.5: Signal processing techniques for feature extraction in frequency and time-frequency domain. Abbreviations: sens = sensitivity, spec = specificity, ANN = Artificial Neural Networks, SVM = support vector machines, MFCC = Mel-frequency Cepstral Coefficient, HMM = Hidden Markov Model, DFT = Discrete Fourier Transform, GMM = Gaussian Mixture Model, PSD = Power Spectral Density, FFT = Fast Fourier Transform, STFT = Short-time Fourier Transform.

2.4.2.3. Time-frequency domain

The time-frequency domain has been explored as a more appropriate technique for capturing the nonstationary behaviour of lung sounds, including techniques such as Short-Time Fourier Transform (STFT), empirical mode decomposition (EMD) and wavelets, as summarised in (Table 2.6).

The STFT is derived by performing the Fourier Transform on individual signal frames:

$$\mathbf{X}(\mathrm{t},\mathrm{f}) = \int_{-\infty}^{\infty} x(\tau) w(\mathrm{t}-\tau) \mathrm{e}^{-\mathrm{j}2\pi \mathrm{m} \mathrm{f} \tau} \mathrm{d}\tau$$

where $w(\mathrm{t}-\tau)$ is a window function, typically a Hamming window, and $\mathrm{e}^{-\mathrm{j}2\pi \mathrm{m} \mathrm{f} \tau}$ is a complex sinusoid form used to convert the signal into the frequency domain. A range of features have been suggested to be extracted from the STFT: peak frequency, discontinuity (Taplidou 2003, Taplidou and Hadjileontiadis 2007), mean, amplitude, local maxima, median frequency, entropy, high order frequency moment (Sanchez Morillo et al. 2013). An alternative approach has employed image processing techniques applied to the STFT as an image (spectrogram) analysed to locate wheeze in the signal (Lin 2006). STFT is computationally simple but has been associated with a number of drawbacks, including low resolution (i.e. unsuitable for short lung sounds like crackles) and highly susceptible to noise (Rizal 2015).

Falndrin et al. have reported on the ability of EMD to deal with noisy signals, demonstrating the techniques robustness to Gaussian noise by acting as a dyadic filter bank, similar to wavelet decomposition (Flandrin 2004). Hadjileontiadis has further expanded this concept by combining EMD with fractal dimension (FD) analysis in a EMD-FD filter (Hadjileontiadis 2007). Although reported results demonstrate high performance (sensitivity of 98.1% and specificity of 97.7%), the dataset used in the study was rather small (16 samples) and consisted of recordings from audio tapes accompanying medical teaching materials. The latter are already filtered and carefully chosen by the authors of the teaching materials as they are easy to interpret. Hence, the applicability of this approach to 'real-life' data is unknown.

Another approach to automated detection of lung sounds is based on the use of wavelets and a number of applications have been reported in the literature. Hadjileontiadis et al. have combined continuous

wavelet transforms with a scale dependent threshold as a tool for identification of wheezing (Taplidou et al. 2004). Kayha et al. have used wavelets in two different ways – as part of an adaptive filtering approach and as part of a wavelet packet transform to isolate crackles (Kayha 2001). Yeginerand have introduced the concept of wavelet networks as a tool for monitoring of pulmonary crackles (Yeginerand 2005).

2.4.2.4. Multi-domain

A few studies have explored the combination of signal processing techniques from multiple domains (Table 2.7). For example, Yilmas et al. have combined AR modelling in the time domain with quantile frequency in the frequency domain (Yilmaz and Kahya 2006). A number of other approaches have tried improving detection by extracting a larger number of characteristics through a combination of: (1) Discrete Wavelet Transform and AR modelling (Kahya, Yeginer, and Bilgic 2006); (2) STFT and Power Spectral Density via Welch's method (Mazic 2003); (3) STFT and WPD (Emrani 2013). Although results have indicated that the use of multi-domain features might improve performance, the main limitation with methods combining multiple domains has been the longer computation time (Rizal 2015).

Signal processing	Classifier	References	Sample size	Lung sound or disorder	Outcomes	Limitations
TIME-FREQUENCY DOMAIN						
STFT	-	Taplidou, S.A.	13 signals	wheeze	95.5% sens and 93.7% spec	Small sample size from online database with "ideal lung sounds"
	-	Jin, F.	14 signals	wheeze	85.3%-97.9% accuracy	Small sample size, information leakage between train and test patients
	SVM	Jin, F.	21 signals	normal & disease	97.7% accuracy	Small sample size, collected in a lab
	-	Lin, B.S.	30 signals	wheeze & normal	96.7% sens and 90.7% spec	Small sample size, information leakage between train and test patients
	-	Jain, A.	19 sounds	wheeze & normal	84% sens and 86% spec	Small sample size, information leakage between train and test patients
EMD	-	Charleston-Villalobos, S	2 patients	crackles	Fine crackles information in IMF2 and coarse crackles information in IMF3	Small sample size and validation of performance is via visual inspection of results.
	-	Hadjileontiadis	not specified	4 lung diseases	97% accuracy	Samples from database of "ideal" lung sounds
Wavelets	DA	L. Hadjileontiadis	16 signals	crackles types	100% accuracy	Small sample size, information leakage between train and test patients
	-	Hadjileontiadis	unclear	wheeze	anecdotal	No significant sample size quantification
	-	Du, M.	unclear	crackle types	99.8-100% accuracy	No significant sample size quantification
	-	Yeginerand, M.	2 subjects	crackle types	70-84% accuracy	Small sample size, information leakage between train and test patients
	-	Gross, V.	32 subjects	pneumonia vs other respiratory	anecdotal	No significant sample size quantification
	GMM	L. Xiaoguang	18 signals	crackle types	93.9% accuracy	Small sample size from online database with "ideal lung sounds"
	ANN & SVM	Abbasi, S.	48 signals	normal & adventitious	93.5-100% accuracy	Small sample size from online database with "ideal lung sounds"
	SVM	Emmanouilidou, D.	28 signals	normal & adventitious	90.2 sens and 73.5% spec	Small sample size, information leakage between train and test patients
	ANN	Hashemi, A.	140 subjects	wheeze	89.3% accuracy	Unclear gold standard
	ANN	Kandaswamy, A.	unclear	six lung sound classes	92.6% accuracy	information leakage between train and test patients

Table 2.6: Signal processing techniques for feature extraction in time-frequency domain. Abbreviations: sens = sensitivity, spec = specificity; ANN = Artificial Neural Networks; STFT = Short Time Fourier Transform; EMD = Empirical Mode Decomposition; DA = Discriminative Analysis; SVM = Support Vector Machines; GMM = Gaussian Mixture Model.

Signal processing	Classifier	References	Sample size	Lung sound or disorder	Outcomes	Limitations
MULTI-DOMAIN						
Mixed	SVM	(Icer 2014)	60 signals	normal, crackles & rhonchus	82.6-100% accuracy	Small sample size, collected in controlled setting
	FDA	(Aydore et al. 2009)	7 signals	wheeze	93.5-95.1% accuracy	Small sample size, information leakage between train and test patients
	kNN & ANN	(Kahya, Yeginer, and Bilgic 2006)	40 signals	normal & pathology	80% accuracy	Small sample size, information leakage between train and test patients, classification of frames but no analysis on patient
	kNN	(Yilmaz and Kahya 2006)	48 signals	normal & disorders	77.8% accuracy	Small sample size, information leakage between train and test patients
	PW	(Guler et al. 2005)	57 signals	normal & disease	17.9% classification error	Small sample size
	kNN & SVM	(Serbes 2013)	26	crackles	97.50% accuracy	Small sample size, information leakage between train and test patients

Table 2.7: Signal processing techniques for feature extraction in multi-domain. Abbreviations: sens = sensitivity, spec = specificity, kNN = k-Nearest Neighbour, ANN = Artificial Neural Networks, SVM = support vector machines, DA = Discriminant Analysis, PW = Parzan Window.

2.4.3. Machine learning techniques for classification of lung sounds

Partial reviews of existing approaches to lung sound classification have been conducted over the last 2 decades (Palaniappan 2013b, Gurung et al. 2011, Palaniappan 2013c). The Artificial Neural Networks (ANNs) and k-Nearest Neighbour (kNN) are the two most popular techniques for classification of lung sounds in the literature. Other techniques, such as Gaussian Mixture Model (GMM), Hidden Markov Models (HMM), Fuzzy logic appear in fewer studies and the use of support Vector Machines (SVM) was seen to be particularly rare.

Tables 2.4-2.7 include information on the choice of classification methods in combination with signal processing techniques for feature extraction. Popular classification approaches such as kNN and ANN focus on classifying individual frames of the signal as belonging to one of several lung sound classes (e.g. normal vs crackles vs wheeze etc.). In particular, using kNN, a frame is classified as belonging to the i^{th} class if the majority of its neighbours have been assigned to that class too. Yet, this type of analysis does not capture information on patterns in the sequences of observation. The latter is likely to be

important both for the identification of individual lung sounds, as well as the overall analysis of a signal and its classification as indicative of a specific lung conditions. Additionally, the classification of individual frames was seen to lead to "information leakage" in a large number of the existing research papers, where even though frames were divided into a training and a test set, it is not ensured that frames from the same patient do not feature in both.

2.4.4. Outstanding challenges in automated lung signal analysis

The literature investigation performed in this chapter led to the identification of four challenges and gaps in the current approaches to lung sound analysis:

- **Small and idealised datasets**: The development of analytical tools for lung sounds analysis is typically conducted on small datasets, including acoustic samples from clinical education databases, e.g. RALES and Littman databases introduced in Chapter 3. Whilst such datasets are a great teaching tool, as they clearly exemplify the signal properties of lung sounds, they are very idealised, not capturing the plethora of acoustic interferences that lung signals recorded in 'real' clinical settings experience (e.g. movement, verbal noise etc.). Consequently, most of these approaches have not been validated on larger datasets, collected in a realistic clinical or community setting. Moreover, many of these existing analytical approaches are developed to identify a specific lung sound (e.g. crackles) among a very clear background signal that does not contain any other pathologies or noise.
- **Model training and testing**: A common error in the methodology adapted in existing approaches is information leakage between the training and test sets. As acoustic signals are typically analysed in frames, many studies divide frames into a training and a test set, without ensuring that frames from the same patient should be kept in either of those sets. This leads to artificial improvement in performance but also poor transferability to large and unseen datasets.
- **Limited multi-domain approaches**: Analysis of the signal is often conducted in single domain – e.g. time domain, or frequency domain, or wavelet domain. However, the complexity of lung sounds and the great acoustic interference in 'real' lung signals is rarely captured in a single domain. Hence, such approaches are less robust and do not generalise well. *Multi-domain*

approaches in the context of lung sounds have been limited, with a few notable exceptions introduced in this chapter. Some of the main barriers to this type of analysis have been long computation time and large feature sets for a limited number of signals.

- **Other pneumonia-associated biomarkers in signal**: Most research effort in this domain has focused on automating the identification of the clinically defined lung sounds (crackles, wheezing, bronchial breathing). However, these are arbitrarily defined sounds in the signal that have emerged through empirical clinical practice and defining a robust gold standard for them is very difficult, as it will be illustrated in Chapter 8. A few of the studies introduced in this chapter attempted the classification of full signals into diagnostic classes, delivering good results. This suggests that there might be other biomarkers in the acoustic signal that contribute to the diagnosis of pneumonia.

2.5. Conclusions

As highlighted in this chapter, current approaches to improving the diagnosis of childhood pneumonia suffer from a number of limitations:

- Product innovations are mostly focused on optimising measurements of individual, or a limited number of, vital signs or observational health characteristics. Yet, studies have repeatedly demonstrated that these symptoms are insufficient for the reliable diagnosis of pneumonia.
- Decision rules and data-driven models proposed by the research community rely on simple thresholding or logistic regression involving conventional vital signs. Limited validation of these models, internal (on an independent test sets from the population) or external (on different populations), has been conducted; the few studies that have attempted to assess generalisability have reported poor results. Additionally, the majority of these studies include small datasets, collected in a hospital setting, with inclusion criteria that already impose some preliminary screening. The latter is likely to exclude cases of occult pneumonia, further reducing generalisability to more realistic clinical or community settings. None of these studies have explored the generalisability of results to diagnosis in community or low-resource settings.

- Early warning scores have been developed for hospital use but their generalisability to community settings is questionable for two main reasons: (1) they are designed to identify deterioration in sick children, rather than serve as screening tools for large populations containing a mixture of healthy cases and cases of different diseases; (2) the data required for these scores are mostly suited to examination in hospital settings and not community-based examination with point-of-care tools.
- Computerised lung sound analysis presents an opportunity to enable access to a diagnostic tool (auscultation) currently only accessible via highly trained experts. However, a number of technical challenges with current analytical approaches have been identified as barriers to accurate, reproducible and generalisable diagnostic performance.

This thesis addresses all of these challenges systematically, building up towards an integrated data-driven solution for automated diagnosis of pneumonia in low-resource settings. In particular, advanced machine learning techniques are proposed as an alternative to the existing decision rules and evaluated on a large dataset collected in low-resource settings (Chapters 6 & 7). The generalisability of these newly developed methods to different populations is assessed in Chapter 7, offering insights regarding the development of a global model for childhood pneumonia diagnosis. Finally, Chapter 8 proposes a new approach to lung sound analysis, evaluating performance on two large datasets collected in low-resource settings with two different point-of-care digital stethoscopes.

This page is intentionally left blank.

3. Existing Childhood Pneumonia Datasets

3.1. Introduction

This project initially used existing datasets for the development and validation of machine learning methods for diagnosis of pneumonia, before designing and implementing a clinical study in Mumbai, India (described in Chapter 4). Access to the existing datasets was secured by establishing partnerships with clinical teams that had already conducted global health studies on childhood pneumonia. Two types of datasets were identified and analysed – clinical characteristics datasets and lung recordings datasets.

Clinical characteristics datasets

- **Gambian dataset:** a dataset collected by Dr Climent Casals-Pascual & team (Nuffield Department of Medicine, University of Oxford) in the Gambia.
- **Indian (Bangalore) dataset:** a dataset collected by Dr Michael Seear & team (BC's Children's Hospital, Vancouver, Canada) in India.

Lung recordings datasets

- **Online data:** a mixture of lung recordings publicised in open source domains for the purposes of training and research.
- **Peru dataset:** a dataset of lung sounds collected by Prof William Checkley & team (School of Medicine, John Hopkins University) in Peru.

Summary characteristics of the studies behind the collection of the datasets are included in Table 3.1 and details of each dataset are presented next.

Table 3.1: Summary characteristics of studies responsible for the collection of the existing datasets.

Datasets	**Gambia**	**India (Bangalore)**	**Peru**
Purpose of study	To describe the plasma proteomic signature in samples from Gambian children with severe pneumonia, non-severe pneumonia & controls	To develop clinical criteria, suitable for low-resource settings, for differentiating between pneumonia and wheezy disease and identification of children at high high risk of death	To use electronic auscultation to record lung sounds of children with pneumonia, asthma, bronchiolitis and upper respiratory tract infection to develop tools for computerised lung sound analysis
Gold standard	Extended IMCI (see Figure 3.1) & chest radiograph + blood tests	Clinical characteristics + chest radiograph	Clinical characteristics + chest radiograph
Age group	2-59 months	0-59 months	2-59 months
Point of recruitment	Children presenting with pneumonia at five health facilities in Greater Banjul area in Gambia (the Royal Victoria Teaching Hospital, the Medical Research Council Hospital, and the major health centers at Serekunda, Fajikunda and Brikama)	Children who presented at ER with cough or difficulty breathing at four Indian public hospitals in India (King George Medical University, Lucknow; Regency Hospital, Kanpur; Vanivilas Hospital, Bangalore; Bowring and Lady Curzon Hospital, Bangalore)	Children presenting with a primary respiratory complaint to the Instituto Nacional de Salud del Nino, a tertiary care hospital in Lima, Peru
Control group	Controls from the community with no suspected respiratory or other infectious disease	Wheezy disease patients (39.8% of dataset) and non-respiratory disease patients (2.6% of dataset)	Other respiratory conditions (asthma, bronchiatilis)
Types of data	Clinical characteristics (57) collected by a doctor with hospital tools	Clinical characteristics (29) collected by a doctor with hospital tools	Digital stethoscope recordings collected by a doctor with hospital tools

3.2. Clinical characteristics datasets

3.2.1. Gambian dataset

This dataset was originally collected as part of a clinical study described by Huang et al. (Huang et al. 2014). The 1581 participants were Gambian children aged 2–59 months. Various characteristics were recorded for each case. The full dataset consisted of 57 features (clinical characteristics), including measurable clinical variables (e.g. white blood cell count, neutrophils, haemoglobin, etc.), observational clinical characteristics (e.g. sleepiness, sternal indrawing, cough heard, etc.) and conventional vital signs (e.g. respiratory rate, heart rate, oxygen saturation, etc.). Additionally, selected cases (84 cases) also contained four biomarkers (C-reactive protein (CRP), lipocalin-2 (Lcn2), haptoglobin (Hap) and CD163 protein). The incidence of different characteristics across the dataset is reported in Table 3.2.

The diagnostic outcome for each case was provided by a clinician, expanding on the WHO's IMCI guidelines introduced in Chapter 1. The criteria for identification of pneumonia and severity determination are listed in Figure 3.1 and have been discussed in detail by Scott et al. in their search for a widely accepted clinical criteria for pneumonia classification (Scott et al. 2012). Based on this, 777 cases had pneumonia and 791 were primarily healthy controls and 13 were excluded due to insufficient data. The severity of disease for each pneumonia case was also recorded – 455 pneumonia cases (58.6%) were severe and 322 non-severe (41.4%). Additionally, X-ray end consolidation and/or a positive blood culture result were used to differentiate bacterial from viral pneumonia cases. Such information was available for 84 cases, including 22 bacterial and 62 viral pneumonia cases.

Gambian dataset	**Controls**			**Pneumonia**		
	N = 804			*N = 777*		
	Mean	Median	Std	Mean	Median	Std
Gender [%, female]		-			-	
Age [months]	17.5	15	11.9	17.2	15	11.6
Respiratory rate [breaths/min]	34.4	34	6.6	61.4	60	12
Heart rate [beats/min]	128.2	128	16.5	157.3	157	18.9
Oxygen saturation [%]	96.3	98	1.6	98.4	96	32.7
Temperature [°C]	36.6	36.6	2.2	38.7	38.6	3.3
History of fever (%)		-			-	
Duration [days]	-	-	-	-	-	-
Malnutrition (%)		14.1			24.2	
Z-score	-0.8	-0.8	1.1	-1.4	-1.4	1.3
Unable to eat/breastfeed (%)		0			18.1	
Lethargic (%)		0			17.1	
Chest indrawing (%)		0			3.3	
Cough (%)		0			86.4	
Duration [days]	-	-	-	-	-	-
Difficulty breathing (%)		-			-	
Dehydration (%)		0			4.0	
Stridor (%)		0			0.1	
Grunting (%)		0			21.0	
Crackles (%)		0			39.1	
Wheeze (%)		0			0.4	
Bronchial breathing (%)		0			10.8	
Head nodding (%)		0			4.1	
Cyanosis (%)		0			0.3	
Pallor (%)		0			19.9	
Blood culture (%)		9.6			12.5	
Haemoglobin [g/dL]	7.8	9.7	5.7	8.6	9.3	3.4
White blood cell count [billion/litre]	8.1	8.9	6.6	15.6	13.8	11.3
Neutrophils [billion/litre]	23.0	26	19.7	47.1	52.1	25.9
Lymphocytes [billion/litre]	37.1	47.8	29.2	30.1	31	19.9
Monocytes [billion/litre]	4.6	5.1	6.4	5.8	6.4	3.3
Platelets [billion/litre]	-	-	-	-	-	-
Chest X-rays						
Normal (%)		-			-	
Hyperinflation (%)		-			-	
Pleural fluid (%)		-			-	
Lobar changes (%)		-			-	
Minor patchy changes (%)		-			-	
Major patchy changes (%)		-			-	
Positive X-rays [%]		NA			23.6	

Table 3.2: Summary of clinical characteristics in the Gambian dataset.

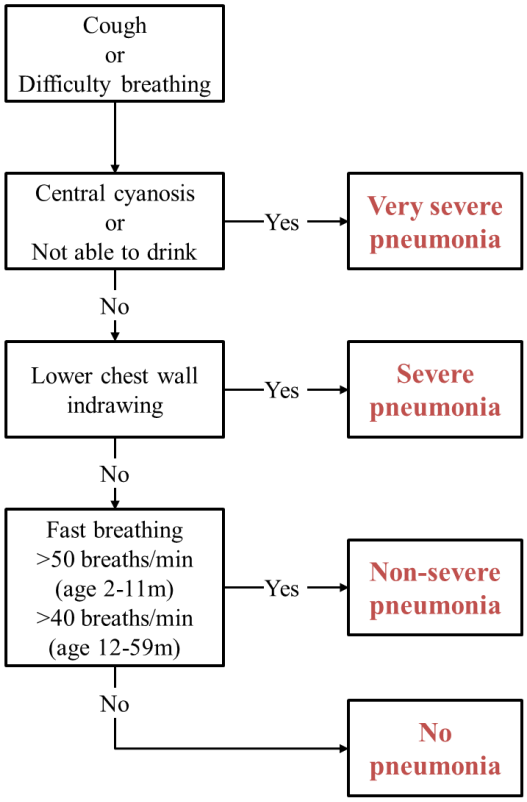


Figure 3.1: Diagnostic criteria used by the clinician to determine pneumonia outcome (Scott et al. 2012).

3.2.2. Indian (Bangalore) dataset

This dataset was collected as part of an observational study conducted at four Indian governmental hospitals – King George Medical University, Lucknow; Regency Hospital, Kanpur; Vanivilas Hospital, Bangalore; Bowring and Lady Curzon Hospital, Bangalore. The study was led by Dr Michael Seear (Division of Respiratory Medicine, Children's Hospital, Vancouver, British Columbia, Canada) and full details of the study are presented in (Gowraiah et al. 2014). In essence, all children under five years who presented at the emergency room with cough or difficulty breathing and were found to have tachypnoea (as per the WHO's IMCI criteria described in Chapter 1) were approached for recruitment in the study. Upon enrolment, 29 features of the child's history, examination and chest X-ray were recorded. An ER physician at each site was responsible for placing the child in one of four categories: (1) pneumonia; (2) wheeze disease (asthma bronchiolitis); (3) mixed (evidence of pneumonia and wheeze); and (4) non-respiratory (malaria, dengue, etc.). The ER doctor managed treatment following diagnosis, including identifying children who should be hospitalised from those who could be treated at home. For all hospitalised patients, five clinical characteristics (heart rate, respiratory rate, oxygen saturation,

temperature and chest indrawing) were measured twice a day for four days. A full list of the characteristics recorded and their incidence across the dataset are reported in Table 3.3. Four days post the original diagnosis, a paediatrician reviewed the clinical data at presentation and subsequent disease course to set a final diagnosis, blinded to the original diagnosis set by the ER doctor. The clinical diagnosis was based on clinical characteristics plus a chest radiograph. A member of the Canadian team conducted training at each one of the four sites, followed by a 1-week trial period, to ensure standardisation of the data collection and diagnostic process.

The dataset included: 199 pneumonia cases (36.5%), 217 cases of wheeze diseases (39.8%), 115 cases of mixed (21.1%) and 14 non-respiratory cases (2.6%). A total of 301 cases were admitted to hospital (and hence had measurements taken twice a day) and 244 were treated as outpatients. The severity of pneumonia cases was also recorded: 55 were found to be non-severe (29.0%), 92 to be severe (48.4%) and 43 to be very severe (22.6%).

Indian dataset (Bangalore)	**Controls**			**Pneumonia**		
	N = 804			*N = 314*		
	Mean	Median	Std	Mean	Median	Std
Gender [%, female]		36.8			35.7	
Age [months]	20.3	12	23.3	16.8	10	20
Respiratory rate [breaths/min]	57	57	11.3	59.5	58.5	11.5
Heart rate [beats/min]	138.9	138	24	138	137.5	25.2
Oxygen saturation [%]	96.7	98	3.4	93.5	95.5	5.4
Temperature [°C]	37.4	37.2	0.8	37.9	37.9	1
History of fever (%)		54.5			94.3	
Duration [days]	-	-	-	-	-	-
Malnutrition (%)		31.2			62.3	
Z-score				-	-	-
Unable to eat/breastfeed (%)		39.4			29.9	
Lethargic (%)		46.8			31.5	
Chest indrawing (%)		35.1			76.8	
Cough (%)		97.4			96.8	
Duration [days]	-	-	-	-	-	-
Difficulty breathing (%)		84.4			88.5	
Dehydration (%)		-			-	
Stridor (%)		-			-	
Grunting (%)		-			-	
Crackles (%)		11.3			76.1	
Wheeze (%)		88.3			37.9	
Bronchial breathing (%)		0.0			0.0	
Head nodding (%)		-			-	
Cyanosis (%)		-			-	
Pallor (%)		-			-	
Blood culture (%)		-			-	
Haemoglobin [g/dL]		-			-	
White blood cell count [billion/litre]		-			-	
Neutrophils [billion/litre]		-			-	
Lymphocytes [billion/litre]		-			-	
Monocytes [billion/litre]		-			-	
Platelets [billion/litre]						
Chest X-rays						
Normal (%)		26.4			18.2	
Hyperinflation (%)		67.1			7.0	
Pleural fluid (%)		6.1			42.0	
Lobar changes (%)		0.4			19.1	
Minor patchy changes (%)		0.0			10.2	
Major patchy changes (%)		0.0			6.7	
Positive X-rays [%]					-	

Table 3.3: Summary of clinical characteristics in the Indian (Bangalore) dataset.

3.3. Lung recordings data

Auscultation, the process of listening to lung sounds via a stethoscope, is used by clinical experts such as pulmonologists to identify a range of pathological lung sounds and to diagnose pneumonia, alongside other respiratory conditions. As illustrated in Chapter 2, lung sounds are particularly beneficial for

achieving improved diagnostic specificity. However, the interpretation of lung signals has historically only been performed by highly trained clinicians, hindering access to this diagnostic tool in low-resource settings. Consequently, this thesis explores the possibility of designing signal processing and machine learning algorithms that automate the interpretation of lung signal data, captured via a digital stethoscope. Two main sources of data were identified: open source libraries and data acquired in Peru by a research team from John Hopkins University. The clinical definitions of the relevant pathological lung sounds are introduced below, followed by a description of each of the data sources used.

3.3.1. Lung sounds – definitions

A number of pathological lung sounds are typically identified by clinicians and used as indication of different pulmonological conditions. The main lung sounds are defined below and will be referred to throughout this thesis.

3.3.1.1. Normal vesicular

Normal vesicular lung sounds are heard through auscultation on the chest wall; they are characterised by a low noise during inspiration and are hardly audible during expiration, as illustrated in Figure 3.2. Normal lung sounds are soft and low pitched. The inspiratory phase is normally twice as long as the expiratory phase, with no pause between the two. Inspiration is higher pitch and of greater intensity than expiration.

3.3.1.2. Crackles

Crackles are short, non-stationary and explosive sounds, indicative of a number of respiratory disorders including pneumonia (Piirila and Sovijarvi 1995). They originate in the airways that open or deform very abruptly. The movement of bubbles in airway fluid and secretion (e.g. patients with chronic bronchitis or oedema) can also produce crackles. Normally detected during inspiration, crackles are also characterised by their specific waveform and short duration (~20ms) (Sovijarvi 2000b), as illustrated in Figure 3.3. The crackle waveform, the number of crackles and their positioning in the respiratory signal are relevant to the diagnostic process (Sovijarvi 2000b, Reichert et al. 2008). Two main types of crackles are identified – fine and coarse crackles. Fine crackles are shorter in duration, higher pitched, dry, close to the ear and associated with pulmonary fibrosis or pneumonia. Coarse crackles are longer in duration,

lower pitched, wet and more distant; they are typically found in chronic obstructive lung disease and bronchiectasis (Piirila and Sovijarvi 1995).

3.3.1.3. Wheeze

Wheeze sounds are of great importance to respiratory diseases; it has been demonstrated that the acoustic characteristics of wheeze are relevant to the detection of abnormalities as well as severity and location of airway obstructions in conditions such as asthma (Sovijarvi 2000b, Taplidou et al. 2004). It originates from flutter and oscillations of the walls of narrowed airways, normally occurring in people with airway obstruction, asthma or chronic obstructive lung disease (Piirila and Sovijarvi 1995). The expiration in patients with wheeze appears to be longer since sounds are audible until the next inspiration. Wheeze sounds are pseudoperiodic signals that often appear as purely sinusoidal waves, as illustrated in Figure 3.4. The frequency of a wheeze lies between 100Hz and 2500Hz, with a dominant frequency (also referred to as pitch) reported to be between 100Hz (or 400Hz according to (Bahoura 2006)) and 1000Hz (or 1600Hz) (Taplidou et al. 2004, Reichert et al. 2008, Sovijarvi 2000a). Rhonchus is a sound similar to wheeze that mainly differs in terms of pitch, defined as ≤200Hz (Reichert et al. 2008). From a signal processing perspective, these frequency differentiations are rather arbitrary and the two sounds are often analysed jointly (Charbonneau 2000). A number of standards for the minimum sound duration of wheeze have been suggested – 250ms (Guler, Polat, and Ergun 2005), 100ms (Taplidou et al. 2004) and 80ms (Qiu et al. 2005).

3.3.1.4. Grunting

Grunting is caused by vibration of the vocal cords during expiration (Figure 3.5) and can be normally heard even without a stethoscope. If a newborn is experiencing respiratory distress, grunting prolongs expiration and helps keep the lung inflated. It can be found in primary respiratory disease as well as systemic illness such as sepsis.

3.3.1.5. Stridor

Stridor (illustrated in Figure 3.6) occurs when the larynx or the cervical trachea narrow, crating upper-airway obstruction, and can be heard during inspiration. Stridor can be heard in patients with whooping cough, and in laryngeal or tracheal stenosis. Wheeze and stridor can be heard simultaneously in patients with widespread airway obstruction.

3.3.1.6. Bronchial breathing

Bronchial breathing (Figure 3.7) is heard near large airways and is characterised by a hollow/tubular sound, resembling air being blown through a tube. The sound indicates that patient airways are surrounded by consolidated tissue, as it is the case in pneumonia or fibrosis patients. The length of the expiratory phase is elongated to twice as much as the inspiratory phase and there is a characteristic pause between them.

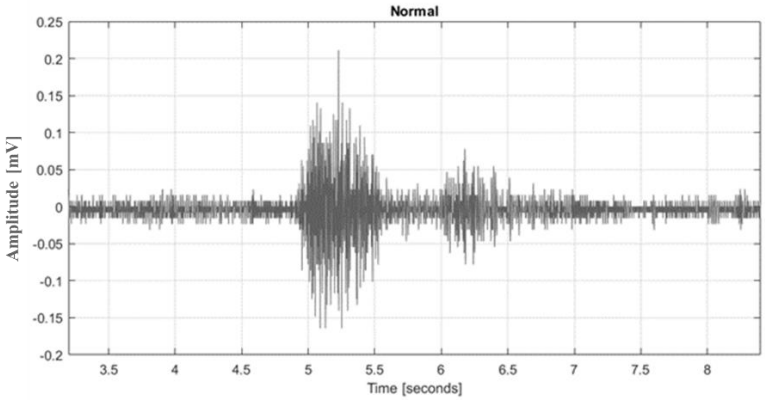


Figure 3.2: Normal vesicular sounds. Source: Littmann online library[2].

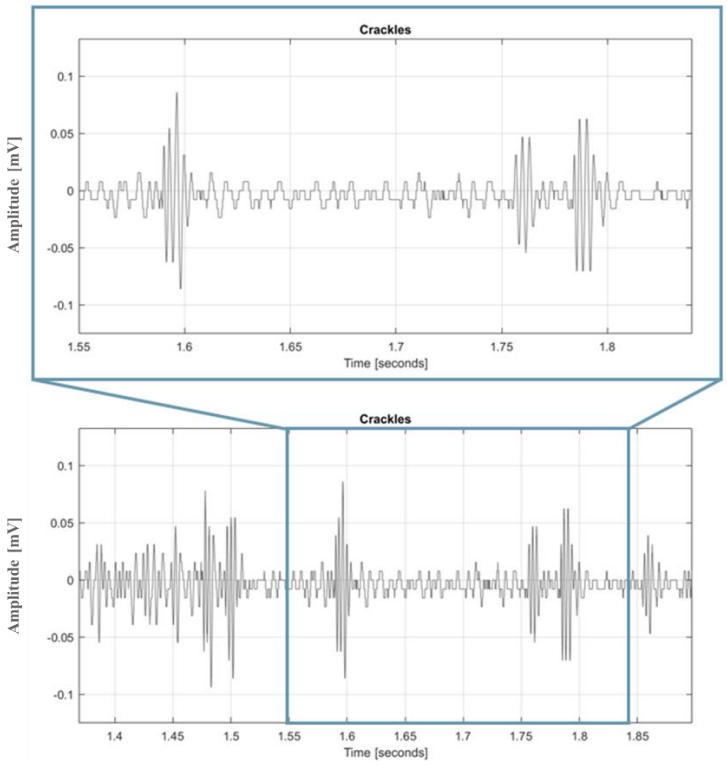


Figure 3.3: Crackle lung sounds. Source: Littmann library.[2]

[2] http://www.littmann.ca/wps/portal/3M/en_CA/3M-Littmann-CA/stethoscope/littmann-learning-institute/heart-lung-sounds/lung-sounds/

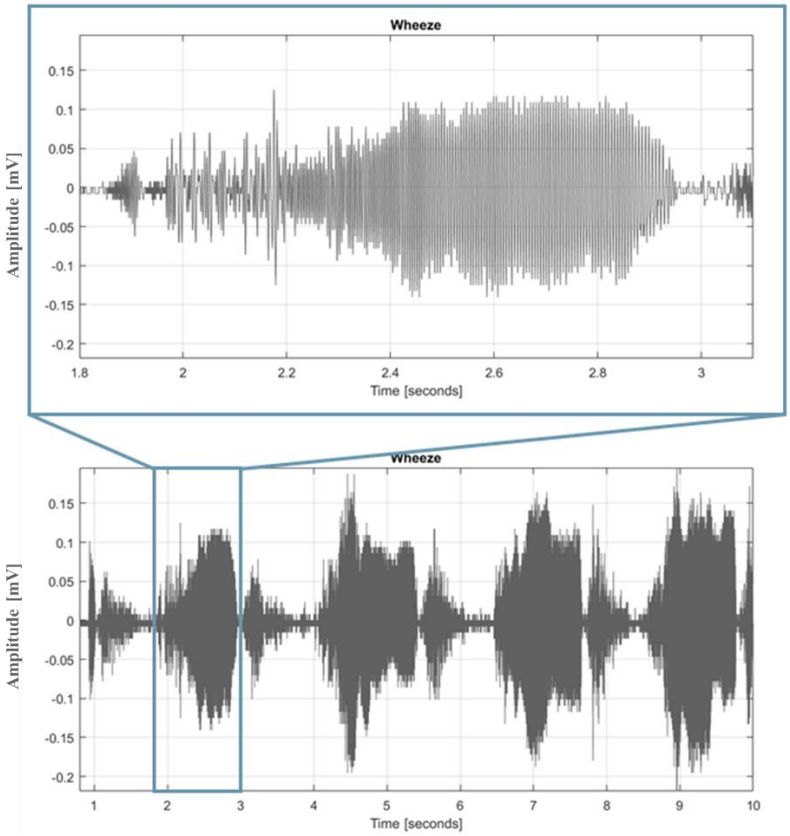


Figure 3.4: Wheeze lung sounds. Source: Littmann library.[2]

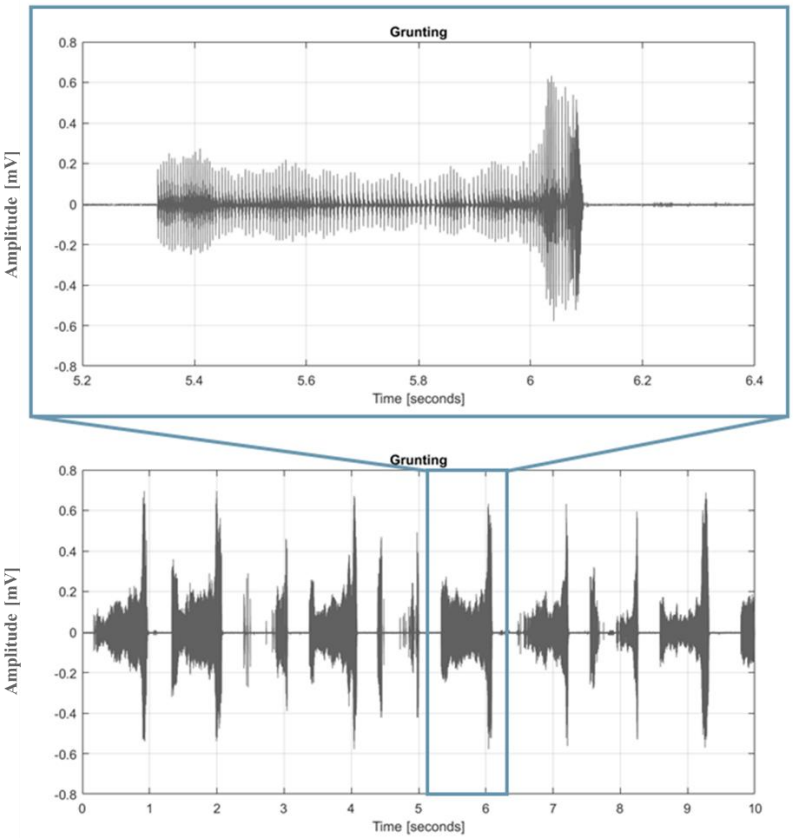


Figure 3.5: Grunting lung sound. Source: Littmann library.[2]

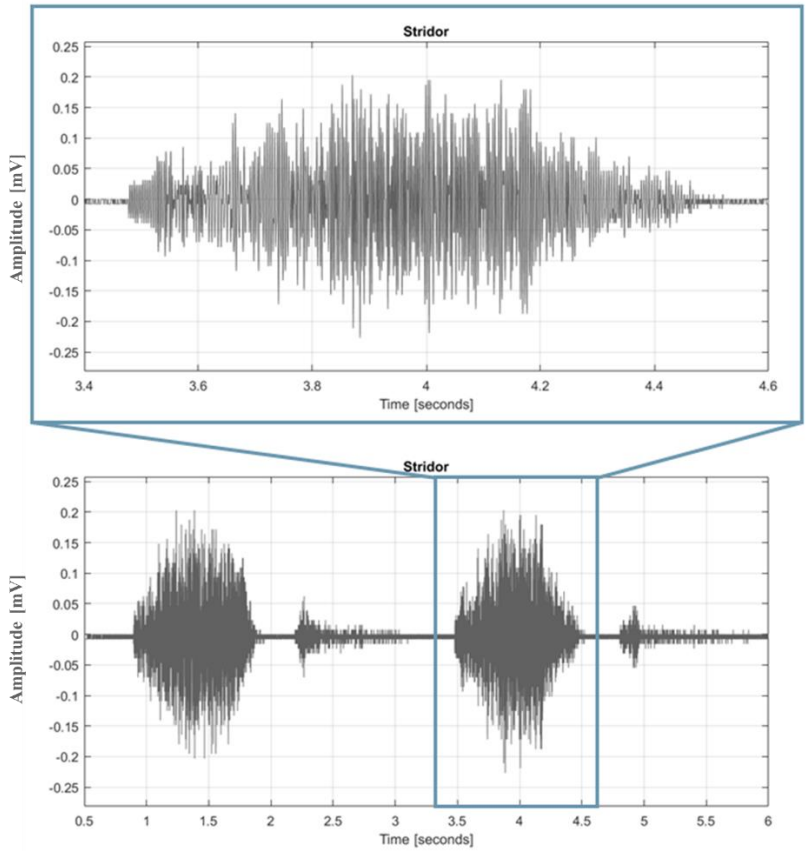


Figure 3.6: Stridor lung sound. Source: Littmann library.[2]

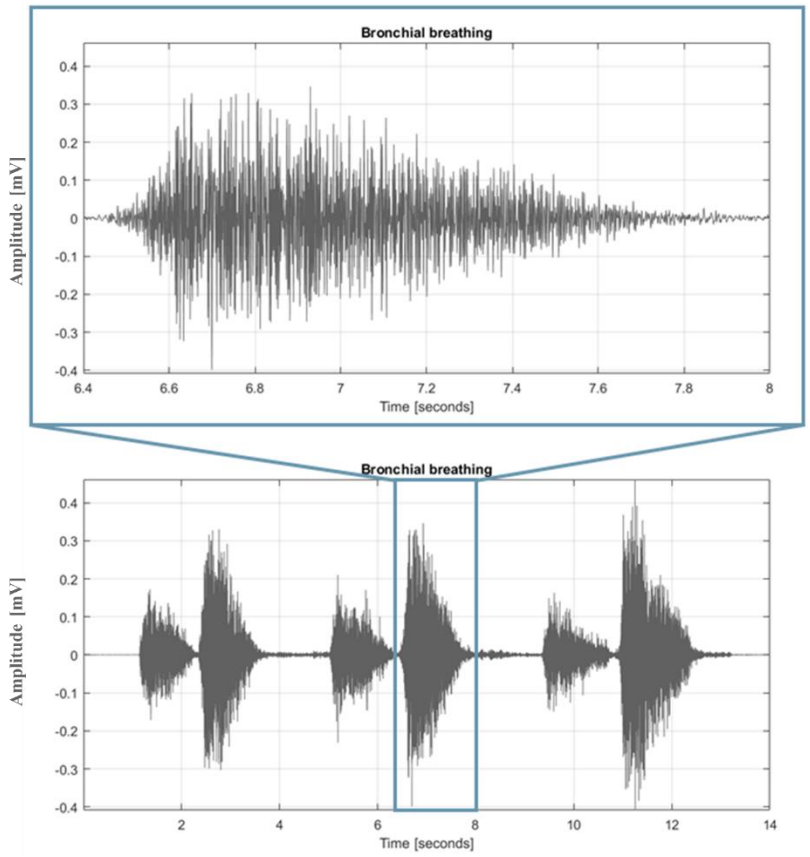


Figure 3.7: Bronchial breathing lung sound. Source: Littmann library.[2]

3.3.2. Open source lung sound recordings

A few open source libraries for lung sounds were found on the internet and used for the initial exploration of lung sounds (a full list below). These lung recordings are made publicly available for the purposes of training of doctors. Consequently, they are very idealistic, containing no noise, and present very obvious "text-book" examples of pathologies.

- **Vivre:** http://education.vivere.org.nz/
  Crackles (2 recordings); Wheeze (1 recording); Rhonchus (1 recording); Stridor (1 recording); Normal vesicular (1 recording)
- **University of Missouri:** http://breathe.missouri.edu/sounds/sounds.htm
  Crackles (1 recording); Wheeze (1 recording); Stridor (1 recording); Grunting (1 recording); Bronchial breathing (1 recording); Consolidation (1 recording); Tracheal (1 recording); Squawk (1 recording); Normal (1 recording)
- **University of Maryland:** https://itunes.apple.com/gb/itunes-u/spring-dpte-516-medical-issues/id431166351?mt=10
  Crackles (2 recordings); Wheeze (2 recordings); Stridor (2 recordings); Grunting (1 recording); Rhonchus (1 recording); Bronchial breathing (2 recordings)
- **Littmann Stethoscopes:** http://www.littmann.ca/wps/portal/3M/en_CA/3M-Littmann-CA/stethoscope/littmann-learning-institute/heart-lung-sounds/lung-sounds/
  Crackles (1 recording); Wheeze (1 recording); Stridor (1 recording); Normal (1 recording)

Examples of the lung recordings are included in Figure 3.8. Recordings are typically between 10 and 12 seconds, including a number of lung sounds repetitions. Background noise is minimal, allowing for the visual identification of the different sounds, in both the time and frequency domains.

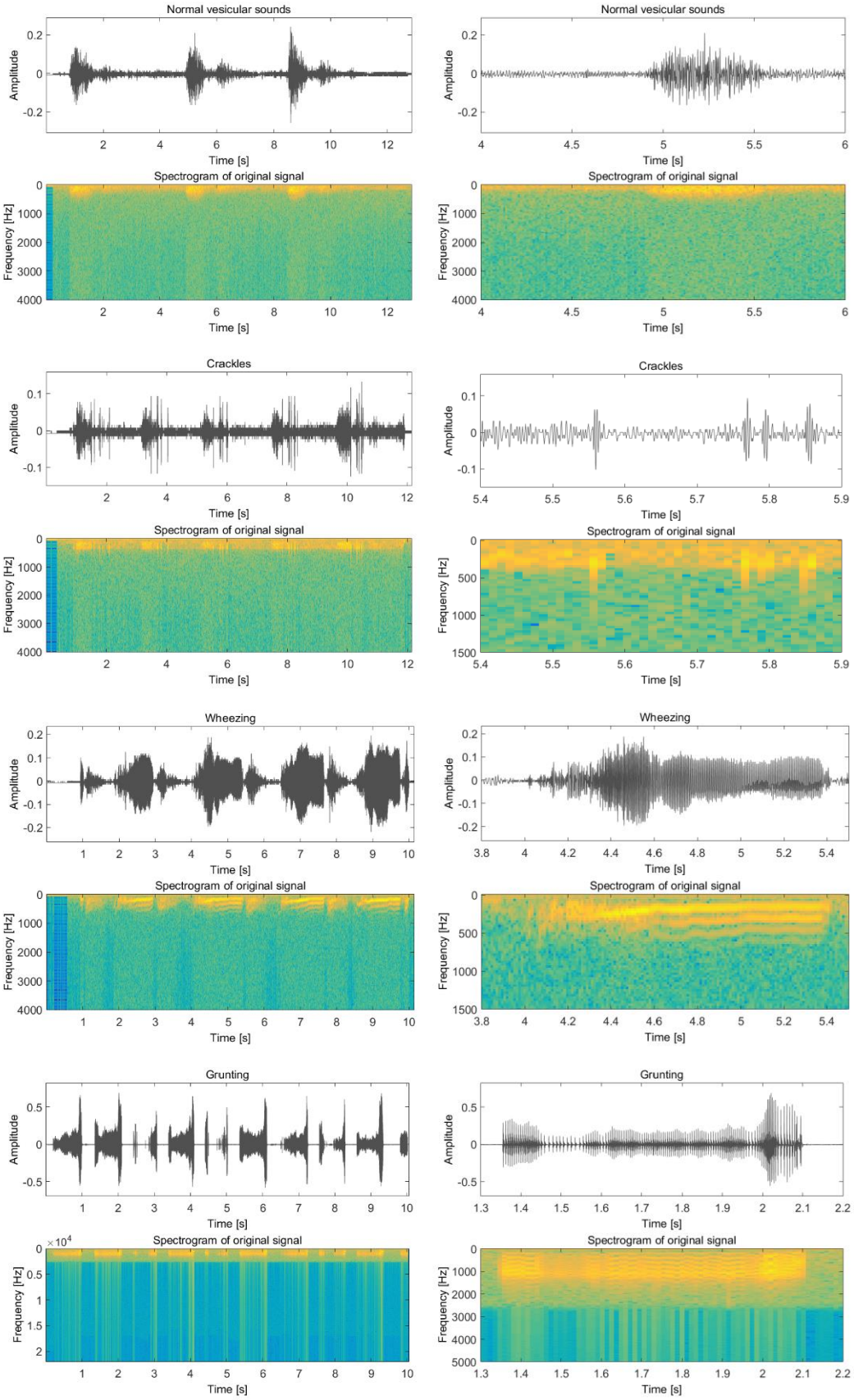


Figure 3.8: Examples of lung recordings containing pathological lung sounds from open source libraries. All sounds are plotted in the time domain and frequency domain (directly below). The right column contains zoomed-in sections of the signals in the left column, highlighting on individual sounds.

3.3.3. Peru dataset of lung sounds

A team from John Hopkins University conducted a cross-sectional study in Peru involving children, aged 2 to 60 months, who presented with primary respiratory complaints. The study was conducted at the Instituto Nacional de Salud del Nino, a tertiary care hospital in Lima, Peru, where children presenting at the emergency department, the asthma or pulmonary ward were recruited. The study group included children without a history of chronic lung disease, excluding asthma, or significant cardiac disease. The control group included children with no respiratory complaints and no history of acute respiratory illness within a period of a month. Pneumonia was originally clinically diagnosed by an examining paediatrician at the Instituto Nacional de Salud del Nino and later radiologically confirmed/rejected by a blinded radiologist at Johns Hopkins University, interpreting the final chest X-ray. A ThinkLabs digital stethoscope was used to obtain 80-second recordings at 8 different sites – 4 at the front and 4 at the back. Details on the study protocol have been published by (Ellington et al. 2012).

A pulmonologist from John Hopkins University was responsible for annotating all signals from the study in Audacity (an open source audio software)[3]. The pulmonologist inspected both the visual and the audio of each signal and marked the exact time positions of different lung sounds (crackles, wheeze, grunting etc.) as well as other events (verbal noise, child crying, stethoscope movement). Those labels could then be exported in a standardised text format and be processed in other software, such as Matlab.

Examples of the full signals, in time and frequency domains, and the annotator's labels are included in Figure 3.9. A summary of the incidence of different lung sounds across four disease categories represented in this dataset (pneumonia, asthma, bronchitis and non-respiratory controls) is included in Table 3.4. It should be noted that the clinical annotator in this dataset was not blinded to the final diagnosis which would influence their decision making, as evidenced by the zero incidence of pathological lung sounds amongst the non-respiratory disease controls.

[3] Audacity: https://www.audacityteam.org/

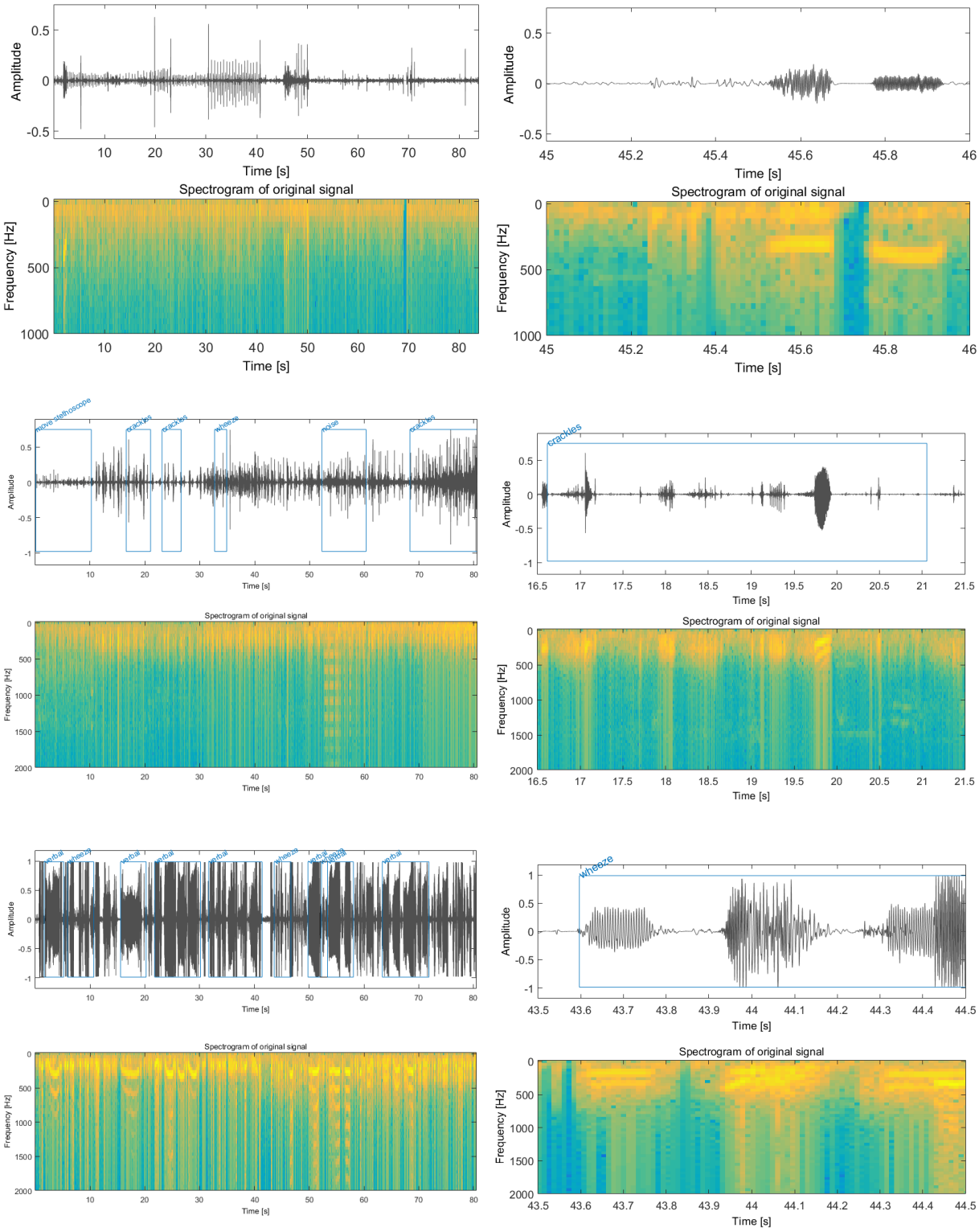


Figure 3.9: Examples of lung recordings in the Peru dataset. Clinical annotations are overlaid in blue. The signal is plotted in the time domain and in the frequency domain (immediately below each time domain plot). Zoomed-in sections of the signal are included in the right column.

	Crackles	Wheeze	Bronchial breathing
Control	0	0	0
Pneumonia	0.361	0.525	0.328
Asthma	0.417	0.604	0.375
Bronchitis	0.367	0.633	0.333

Table 3.4: Incidence of different lung sounds amongst the four disease categories in the dataset.

3.4. Conclusions

This chapter introduced the existing datasets used for analysis in future chapters, including a tri-fold diagnostic pipeline using machine learning techniques, described in Chapter 6 and an automated approach for lung sound analysis, described in Chapter 8. These datasets contain a number of gaps:

- **Insufficient disease variability**: the non-pneumonia class in both clinical characteristics datasets did not reflect a realistic clinical environment, where children present with a wide range of other diseases. The Gambian dataset contained mostly healthy controls and the Peru dataset contained two other respiratory diseases. Hence, models designed to differentiate between pneumonia and these two control groups are likely to generalise poorly to more realistic scenarios where they have to also differentiate between pneumonia and diseases like malaria or diarrhoea.
- **Incomplete data**: the Gambian dataset contained clinical characteristics but no signals and the Peru dataset contained lung signals but no clinical characteristics. To develop and validate a complete set of tools that automate the derivation of key symptoms and their fusion into diagnosis, a dataset including both types of data was required.
- **Point-of-care tools**: Chapter 1 identified a strong need for community-based diagnosis in order to identify pneumonia early and reduce mortality. Consequently, any proposed algorithms for automated diagnosis should ideally be designed on data captured with point-of-care tools, rather than hospital-based equipment, to fully reflect the challenge of capturing and interpreting clinical information in the community. Such tools might produce signals of poorer quality and hence necessitate specific signal processing approaches.

Consequently, a field study was designed to capture a more suitable dataset, addressing the gaps highlighted above; the details of the study are presented in Chapter 4.

This page is intentionally left blank.

4. Indian Field Study

4.1. Introduction

Chapter 3 introduced a number of existing datasets that were used for analysis and also highlighted gaps in the available data which motivated the design and implementation of a field study in Mumbai, India. The study had the following objectives:

- Design a mobile health application that can be used to collect medical data essential for the diagnosis of childhood pneumonia (both clinical characteristics and device signals);
- Collect data via affordable point-of-care medical devices that interface directly with the mobile application (pulse oximeter, digital stethoscope and thermometer) to be used for the development of algorithms for automated analysis;
- Identify challenges associated with the use of such tools by minimally trained community health workers and develop potential technical solutions.

The data collected from this study were used to develop: (1) signal processing tools for the automated analysis of the signals generated by the point-of-care devices; and (2) machine learning tools for the fusion of clinical features extracted from the signals to derive a diagnosis.

Before describing and discussing the study design and implementation process, this chapter first takes a look at the Indian health system, particularly in the context of child health and pneumonia. This introduction aims to describe the positioning of the study in the health system, as well as to inform the decision-making process when designing, iterating and implementing the mobile application and accompanying clinical examination process.

4.2. Health system considerations in low-resource settings

4.2.1. The Indian health system – challenges and opportunities

4.2.1.1. Overview

In the new millennium, India has achieved some significant achievements in health. Between the years 2000 and 2013, life expectancy at birth has increased from 62.5 years to 66 years (WHO 2017). Similarly, the infant mortality rate decreased by a third between 2003 and 2013, reaching 40 per 1000

livebirths (India 2014) (General 2005). The maternal mortality rate was cut by almost a half between 2001 and 2013, reaching 167 per 100,000 livebirths (General 2003, India 2015e). In March 2014, the World Health Organisation declared India polio free, and tetanus free in August 2015. But there is still a lot of work to be done. India's performance on health remains the poorest amongst the BRICS nations (Brazil, Russia, India, China, and South Africa) (Marten et al. 2014). In 2013, India accounted for 20% of the global burden of disease, despite housing 17.5% of the global population (GBD 2015). Additionally, 27% of the neonatal deaths globally and 21% of the child deaths (younger than 5 years) take place in India (Liu et al. 2015). Furthermore, although most Indian states have seen an accelerating reduction in infant mortality, Millennium Development Goals 4 and 5 were still unmet.

There are five conditions which account for 68% of child deaths in India: pneumonia, diarrhoea, neonatal sepsis, preterm birth complications, and birth asphyxia (Liu et al. 2015). On top of this, malnutrition, which is often an underlying cause of child mortality, is a very severe problem in India; between 2013 and 2014, 38.7% of children under 5 years were chronically malnourished (and consequently stunted) and 29.4% of children were underweight (India 2015b).

Underlying these conditions is the influence of inequality in health outcomes, evidenced by large morbidity and mortality differentials across: socioeconomic status, caste, class, sex, and geographic location (Patel 2015). As an example, in the 2000s, the risk of a child dying was 75% higher for girls than boys (Patel 2015). Overall, infant mortality rates between urban and rural areas differ by 17 points (EXIST 2014). Stark differences in infant mortality could be observed even between two districts of the same state (Assam): 37 per 1000 livebirths in Dhemaji and 74 per 1000 livebirths in Kokrajhar (Commissioner 2014). The incidence of severe malnutrition in children born to parents in the bottom wealth quantile was 25%, compared to 5% for children born to parents in the top wealth quantile (Sciences 2007).

The Indian health system is extremely complex and varied; providers range from under-resourced facilities, through reputed and highly reliable hospitals such as the All Indian Institute of Medical Sciences, AIIMS, to world renowned organisations such as the Narayana Hrudayalaya Hospital and the Aravind Eye Care System (HLEG 2011, School 2009, Economist June 2012). The following few

sections discuss in more detail the specific challenges faced by the Indian health system and potential opportunities in the context of child health and pneumonia.

4.2.1.2. The role of the public sector

The National Health Mission in India combines the National Rural Health Mission and the National Urban Health Mission. The former was established in 2005 to address challenges in the public health system and provide equitable access to quality healthcare for people in rural areas. Similarly, the National Urban Health Mission was set up in 2014 to address the health needs of poor people living in urban areas (Patel 2015). Consequently, infrastructure was expanded and 7629 sub-centres (5% of current availability), 2072 primary health centres (8% of current availability), and 2050 community health centres (38% of current availability) were added (Patel 2015). However, a recent study which evaluated basic primary care infrastructure, availability of services, clinical staff, and equipment, reported poor standards across most facilities (Powell-Jackson T 2013). In 2015, only 21% of primary health centres and 26% of community health centres fulfilled the Indian Public Health Standards (Patel 2015). The use of public health services has been seen to decline, especially in urban areas (from 43% in 1995–96 to 32% in 2014) (India 2015c). The socioeconomic background of families has been seen to relate to their propensity to use public services, i.e. the poor are most dependent on public health facilities (Patel 2015).

Despite a recent increase in the uptake of students in the medical profession, India continues to experience extreme shortages in its health workforce (Table 4.1) (Patel 2015). Adding to this complexity, in 2014 only 11.3% of modern medicine (allopathic) doctors were reported to work in the public sector (India 2015a). Practitioners of traditional medicine, AYUSH practitioner (ayurveda, unani, siddha, naturopathy, or homoeopathy) are increasingly less preferred by the population; in 2014, 90% of the population sought care through the allopathic system (India 2015b). Another attempt to address the human resource gap has been the introduction of nearly 900,000 Accredited Social Health Activists (ASHAs) at the village level; these are community health workers trained to deliver mostly maternal and child health services (Patel 2015). When the ASHA programme was introduced in 2005, it was intended that each ASHA cover a population of 1,000 people. ASHA compensation is performance-

based, according to facilitation of universal immunisation, referrals, general child health service, etc. (Mission 2005). Community health workers like these are a tremendous resource that, if trained properly, have been shown to deliver reduction in mortality, particularly through case management (Haines et al. 2007, Kumar et al. 2008, Sundararaman 2007, Bang et al. 1999). Case studies documenting programmes of appropriate community-based or home-based health programmes have been seen to deliver good results, particularly in areas with poor care seeking behaviour and limited access to health facilities. For example, Sachdev et al. have reported reduced risk of neonatal mortality and stillbirth as a result of home visits by community health workers (Gogia and Sachdev 2010). The knowledge level of health workers has been shown to be correlated with outcomes of health programmes, highlighting the importance of tools that strengthen their capabilities (Agrawal et al. 2012).

	Availability (2011-12)	**Needed**	**Shortfall**
Doctors	691,633	1,031,383	32.94%
AYUSH practitioners	534,091	594,562	10.17%
Dentist	88,370	182,009	51.45%
Pharmacist	492,923	849,374	41.97%
Nurses and midwives	743,324	2,062,765	63.96%
Auxiliary nurse midwives	361,879	1,031,383	64.91%

Table 4.1: Availability of health professionals in India – availability and gaps. Adapted from (Patel 2015).

4.2.1.3. The role of the private sector

In response to the insufficient healthcare provision from the public sector, the private healthcare sector has majorly expanded, although it remains extremely fragmented and mostly unregulated. Between 2002 and 2010, 70% of the increase in hospital beds across India came from the private sector (Gudwani 2012). Consequently, the private sector accounted for 70% of outpatient and 60% of inpatient care in 2014 (EXIST 2014). In particular, private providers are the first point of contact for a number of conditions, including acute illnesses and neonatal diseases (May, Roth, and Panda 2014, Willis et al. 2011). Yet, it is estimated that 55% of private providers either have no formal medical training or are not registered with the government to practice medicine (Sudhinaraset et al. 2013).

As a by-product of this expanding private sector, out-of-pocket expenditure on healthcare has been growing (Balarajan, Selvaraj, and Subramanian 2011). Out-of-pocket expenditure on healthcare in India

was responsible for approximately 58% of total health expenditure in 2013 (WHO 2014a). The impact of such expenditure is often catastrophic for families; 55 million people were estimated to have fallen into poverty due to health costs between 2011-2012 (Shepherd-Smith 2012).

4.2.1.4. Health information challenges

Health data in India are gathered by a multitude of providers, both public and private, as well as agencies and surveillance systems. However, there is limited coordination between these stakeholders, and as such little integration and consolidation of the data they collect (Patel 2015). For example, health workers are generally required to collect health metrics according to different local, state and central governments requirements (Analysis 2013, Consulting 2009). Simultaneously, doctors treating patients at public health facilities have limited visibility on the health data collected in the community or by private doctors, who are often the first point of contact (Patel 2015). Yet a fully integrated, population-based health-care system is required to provide equitable access to healthcare for all Indians (India 2011, Reddy et al. 2011). However, such a system is only possible if a framework for alignment and coordination between these stakeholders (public, private, allopathic, traditional, agencies, and surveillance systems) is put into practice, creating an opportunity for innovations that help resolve current divisions and barriers to information exchange.

4.2.2. Childhood pneumonia in India

Figure 4.1 presents a breakdown of the causes of childhood mortality in India. More than 50% of deaths amongst children under 5 years take place during the neonatal period, with pneumonia (9% of deaths), prematurity (13% of deaths) and asphyxia (10% of deaths) being leading causes (Paul et al. 2011). Amongst children between 1 and 59 months, pneumonia (11% of deaths) and diarrhoea (11% of deaths) are the main killers (Black et al. 2010).

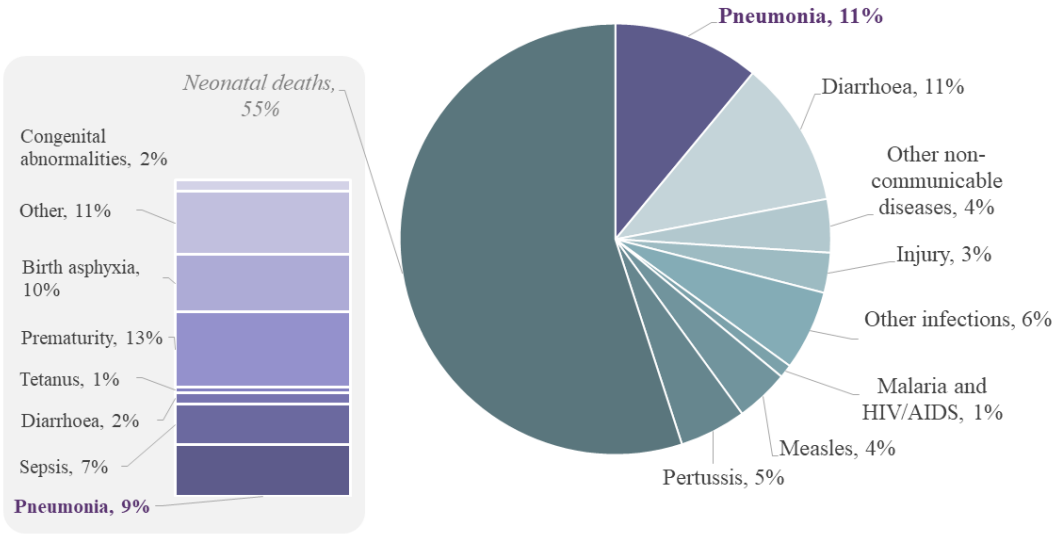


Figure 4.1: Main causes of death in children under 5 years in India. Data from (Black et al. 2010) & figure adapted from (Paul et al. 2011).

Childhood pneumonia in India results from infection by bacteria, viruses or atypical organisms like Chlamydia and Mycoplasma (Walia 2013). Some reports estimate that Pneumococcus accounts for 12-35% of childhood pneumonia cases, whilst H. influenzae and RSV account for 10-15% each (Walia 2013). The determination of pneumonia aetiology remains an unsurmountable challenge in India, where most labs are not equipped with the diagnostic tools required to stratify disease management. Therefore, clinical symptoms remain the main driver of diagnosis and management of the disease. Consequently, both inappropriate administration of antibiotics in cases of viral pneumonia, as well as delayed administration of antibiotics in cases of bacterial pneumonia, are huge challenges (Walia 2013).

The implementation of the WHO's IMCI in India is known as Integrated Management of Neonatal and Childhood Illnesses, and was first introduced in 2005 (Paul et al. 2011). The scale-up of the programme has proven to be challenging; only 43 districts (out of 707) have achieved 80% coverage in all categories of workers, although training has been initiated in 223 districts (Paul et al. 2011). Additionally, more work is required to strengthen referral pathways to ensure that a child who has been identified as suffering from pneumonia or diarrhoea can be referred to a doctor and prescribed the right treatment. The strategic approach - Reproductive, Maternal, Newborn, Child and Adolescent Health (RMNCH+A) - was launched in 2013 by the Indian government to improve equity, harmonisation, and coverage of

health services (India 2014). Within this approach, an integrated management of common childhood illnesses (including pneumonia, diarrhoea, and malaria) intervention was included.

A systematic review by UNICEF and the Public Health Foundation of India has highlighted some of the main gaps in the Indian health system that have resulted in high mortality amongst children, even in spite of the above-mentioned programmes (Mathew et al. 2011). Namely, community-based screening and management for pneumonia has been shown to be effective but has low coverage in India (Bang et al. 1990, Reddaiah and Kapoor 1991). Further studies have found that more than 45% of babies receive no treatment due to a range of barriers hindering access to care, including: lack of doctors, home remedies, and financial constraints (Dongre, Deshmukh, and Garg 2008).

4.3. Study design

The analysis presented in Section 4.2 highlights the gaps in the Indian health system that contribute towards high childhood mortality, especially due to childhood pneumonia. Hence, to address these gaps, new solutions need to:

- be appropriate for use by minimally trained health workers, and bridge the skills gap between highly-trained-but-sparse clinicians and health workers;
- be affordable and minimise the consumables required to perform individual tests;
- enable visibility of data by a variety of providers, both at the community and the hospital level.

Consequently, this project included the design of a mobile health toolkit to collect data in a clinical study in India, incorporating these three requirements. Funding for the study was secured through the Global Challenge Research Fund.

4.3.1. Study positioning within the health system

In September 2015, a needs assessment in the context of childhood pneumonia was conducted in India in order to: (1) map the clinical pathways children with suspected pneumonia follow; (2) identify key stakeholders operating in the field and their specific roles; (3) identify clinical gaps that can be addressed through technical automation. Informal conversations were conducted with NGOs, as well as public and private hospitals involved in the diagnosis and/or treatment of childhood pneumonia. Findings were used to design a research study, aligned with the needs of the local health system.

4.3.1.1. Clinical pathways

A child from a poor urban community in India, who is suffering from pneumonia, would normally follow one of three clinical pathways with their disease. Depending on the community where the child lives, he/she would most likely see a community health worker first. This could be as part of regular screening activities conducted by some NGOs in urban slum communities, or from a parent noticing their child is unwell and approaching one of their local community health workers. Multiple interactions with the community health worker might be required before they identify that the child needs clinical attention. A community doctor, operating either privately or as part of a child health scheme run by a local NGO, is the next most common interaction with the child. Parents from these poor urban communities are typically reluctant to take a child straight to the large hospitals, as this requires more time, resulting in daily wage loss, and transport expenses. The community doctor might be able to diagnose the disease and treat locally, but if the child does not improve, they will be referred to a public hospital. Depending on the severity of their disease, the child might require hospitalisation. The multiple layers of these clinical pathways often introduce delays which can be detrimental to a child suffering from severe pneumonia. Hence technology-driven interventions should be designed in a way that mitigates delays and clinical uncertainty, and optimises the early and accurate identification of pneumonia and its severity.

4.3.1.2. Study structure

A four-month study was conducted in Mumbai, India, between March and June 2017. Data collection took place at two locations; (1) in a poor urban community (Shivaji Nagar) via a community partner organisation, a local NGO called Apnalaya; and (2) at a large governmental hospital (Lokmanya Tilak Municipal General Hospital, also known as the Sion Hospital). A mobile health toolkit, including a mobile phone and a set of point-of-care devices (full details in 4.3.2.1), was developed and used for data acquisition in both locations. In the community, two types of staff engaged in the data collection process: (1) community health workers who used the toolkit to collect data, as well as to screen children to identify those with any danger signs (using definitions from the IMCI guidelines, as described in Chapter 1); and (2) community doctors who determined whether a child could be treated locally or needed to be referred to hospital for further examination and/or treatment. It should be noted that the community

health workers were engaged in the project on a full-time basis and did not conduct any other activities during its duration, whilst the community health doctors' engagement was alongside of existing health consultation services (e.g. malnutrition screening & treatment) they were conducting with young children. At the hospital, a nurse and a paediatrician collected data via the toolkit and the doctor provided a final diagnosis, after performing a number of additional tests (X-rays and blood rests). The data collection process was designed to run in parallel to all standard clinical procedures at the hospital, without interfering with any clinical activities and without affecting care. Depending on which and how many of these stages a study participant went through, they were allocated to one of three cohorts, as visualised in Figure 4.2.

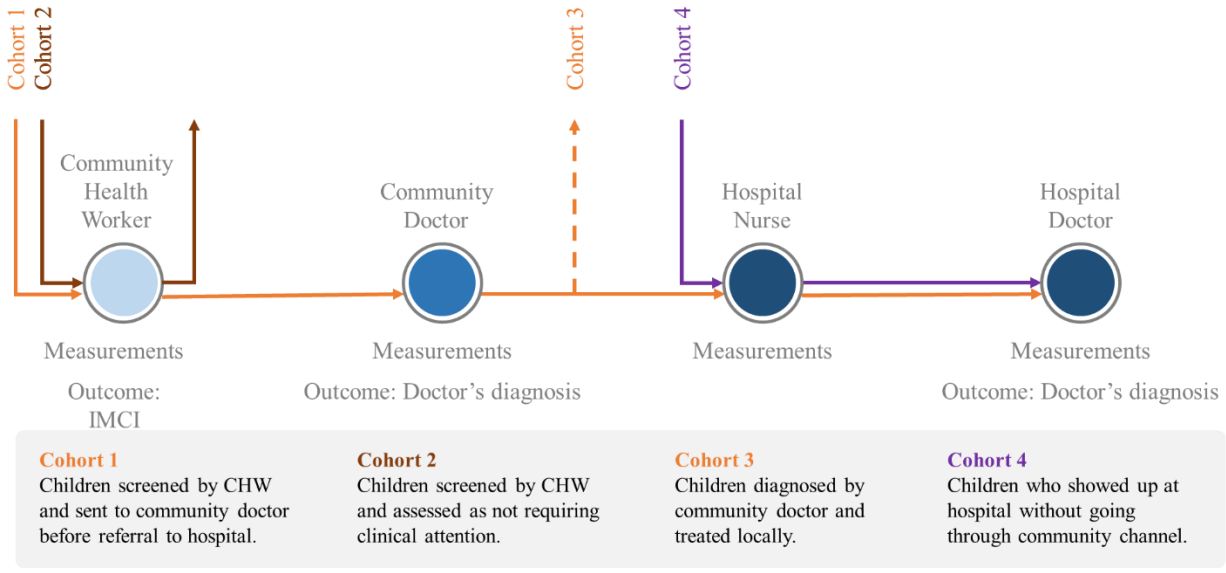


Figure 4.2: Study cohorts depending on the number of interactions and outcomes.

4.3.1.3. Outcomes and gold standards

As illustrated in Figure 4.2, three main types of outcomes were involved in the study. The gold standards used for each are described here:

(1) **General illness / referral to community doctor**: The IMCI guidelines, as described in Chapter 1, were implemented in the mobile application to enable the community health workers to perform general screening. The sensitivity of the guidelines was increased, so that any child with any of the following symptoms was referred to the community doctor: cough/difficulty breathing, fever≥37.5, chest indrawing, fast breathing, stridor, vomiting, unable to eat.

(2) Community diagnosis / referral to hospital: The community doctors used both IMCI guidelines and the medical device measurements to reach a diagnosis, specifically focusing on severity and deciding whether a child could be treated locally or should be referred to hospital. The doctors were instructed to refer all children with suspected pneumonia to hospital; for other milder respiratory conditions, they were instructed to make a decision on whether or not to treat locally as per their normal clinical practice.

(3) Full diagnosis / hospital diagnosis: The hospital doctors conducted blood tests, X-rays and other hospital-based tests to reach a final diagnosis. For pneumonia, the gold standard used was a lung consolidation in the X-ray and positive blood culture. Additionally, the doctor recorded the administered treatment course; medication prescribed, hospitalisation, oxygen etc.

4.3.1.4. Sample size

To calculate the required sample size, the statistical protocol for performing power calculations in diagnostic tests, described by Jones et al. was followed (Jones et al. 2003). The literature was surveyed to obtain a reliable estimate of the childhood pneumonia prevalence amongst the types of urban communities under investigation. Incidence rates between 25% and 55% were identified. The lowest acceptable sensitivity of the diagnostic test was assumed to be 85% (SN) and the lowest acceptable specificity – 80% (SP). A confidence interval of 5% (W) was selected. The required sample size can be determined via the sensitivity or the specificity limit, depending on priorities:

$$\mathrm{N(SN)} = \frac{\mathrm{TP} + \mathrm{FN}}{\mathrm{P}} = 1.96^2 * \frac{(\mathrm{SN}(1 - \mathrm{SN}))}{\mathrm{W}^2}$$

$$\mathrm{N(SP)} = \frac{\mathrm{FP} + \mathrm{TN}}{(1 - \mathrm{P})} = 1.96^2 * \frac{(\mathrm{SP}(1 - \mathrm{SP}))}{\mathrm{W}^2}$$

where N(SN) is the sample estimate based on the acceptable sensitivity value, N(SP) is the sample estimate based on the acceptable specificity value and the rest of the variables are as defined above. For this study we opted out for selecting the larger sample estimate of the two.

Table 4.2 reports on N(SN) and N(SP) estimates for a range of pneumonia prevalence values, where a sample size of 784 was seen to be sufficient for prevalence as low as 25%. Assuming that approximately 20% of participants drop out or do not follow up with referral, this equates to 980.

	Pneumonia prevalence									
	0.1	*0.15*	*0.2*	*0.25*	*0.3*	*0.35*	*0.4*	*0.45*	*0.5*	*0.55*
N(SN)	1959	1306	980	784	653	560	490	435	392	356
N(SP)	218	230	245	261	280	301	327	356	392	435

Table 4.2: Sample size estimates for a range of pneumonia prevalence values.

Census style sampling was used in the data collection, where any individual who satisfied the eligibility criteria (detailed in Section 4.3.1.5) and came into contact with the study staff through the existing clinical workflow was invited to participate. Formal consent was obtained from each participant (the accompanying adult with each child) before they entered the study.

4.3.1.5. Inclusion and exclusion criteria

The following inclusion criteria were used in this study:

- Children aged between 0-59 months;
- Children experiencing common symptoms of childhood disease, including: fever, cough, chest indrawing, diarrhoea, vomiting, malnutrition, breathing difficulties, difficulty swallowing, etc., as well as children appearing to be healthy;
- Children brought to a community health worker/doctor by a parent concerned with recent symptoms and behaviour.

The following exclusion criteria were used in this study:

- Children older than 59 months;
- Children suffering from trauma-related symptoms such as bleeding, broken bones, cuts, bruising etc.;
- Children experiencing extremely severe symptoms in need of immediate medical care, such as unconsciousness and convulsions;

- Children not accompanied by an adult or accompanied by an adult care-taker without the legal right to make health related decisions for the child.

4.3.1.6. Subject recruitment

At the hospital, participants were recruited in the waiting room, where the study staff approached each child and accompanying adult to introduce them to the study and offer them the opportunity to participate. Each participant was asked to sign a formal consent form after being given a full introduction to the study. The consent form and accompanying information sheet were available in the local language (Hindi and Marathi); the mobile application governing the data collection process included audio recordings for parents unable to read the information themselves. In the community, participants were recruited by community health workers from the community partner organisation, Apnalaya, who were trained to perform home visits and approach parents whose children satisfied the inclusion/exclusion criteria of the study.

Participants in the healthy control group were recruited in two ways: (1) in the hospital, when healthy children who satisfied the inclusion criteria for the study came in for a vaccination or when accompanying sick siblings; (2) in the community, community health workers would approach families with children having no apparent health conditions, and who satisfied the inclusion criteria for the study, and invite them to participate in the study, all the while making it very clear that the child is not suspected to have pneumonia.

To incentivise participants to comply with referral recommendations, the study staff were trained to carefully explain to parents that: (1) a consultation with a trained clinician is vital to the health of their child; (2) a designated clinical team, i.e. a doctor in the community or a nurse/doctor at the hospital, would attend to their case and they would not be required to wait in long lines. Moreover, transport costs to and back from the hospital were covered by the study. Additionally, the community partner on this study (Apnalaya) had already been working closely with the community, so people were well accustomed to the services of their community health workers, who already made referrals to the local doctor if children exhibited any health danger signs.

4.3.1.7. Ethical approval

The study design required rigorous review by Oxford University as well as in India. Approval was granted by the Oxford Tropical Research Ethics Committee (OxTREC Reference: 58-16) on the 21st December 2016. The ethical approval process in India was more complex, requiring approval by both the hospital participating in the study (institutional approval) and by the Health Ministry's Screening Committee (HMSC) at the Indian Council of Medical Research (national level). The latter conducts an extremely diligent screening of studies which use foreign funds to ensure research projects which include foreign collaborators are fully aligned with the interests and needs of the local population and the local health system. The HMSC review process took 3 months and approval was granted (5/7/1574/2016-CH) on the 22nd February 2017. Delays in the HMSC review resulted in delays in the study launch day. As a result, the study took place between March and June 2017.

4.3.2. Mobile health toolkit

The mobile health toolkit consisted of a set of three point-of-care devices (detailed in Section 4.3.2.1) and a mobile application, which was preloaded onto an Android mobile phone. A picture of the equipment is included in Figure 4.3.

Figure 4.3: Mobile health toolkit consisting of 3 point-of-care devices and a mobile phone. Each set was kept in a small travel bag which helped protect the equipment.

4.3.2.1. Point-of-care devices

Three point-of-care devices were used – a thermometer, a digital stethoscope and a pulse oximeter. The three devices used in this study were commercially available (covered by both FDA and CE mark), and presented no risk to the health of the child or the user. Details of each device are presented in Table 4.3.

The manufacturers of the digital stethoscope and the pulse oximeter provided an Application Programming Interface (API) which was incorporated into the mobile application to enable connectivity and immediate-transfer of data from the point-of-care devices onto the phone. The digital stethoscope connected to the phone via its own WIFI network and the pulse oximeter connected via Bluetooth; hence, it was possible to have both devices connect simultaneously.

	Pulse oximeter	**Digital stethoscope**	**Thermometer**
			
Manufacturer	Shanghai Berry Electronic Tech, China	eKuore, Spain	iProven
Distributor/ Vendor	Smart Care Sleep, UK	eKuore, Spain	Amazon, UK
Certification	CE	CE & FDA	CE & FDA
Technical specifications	Sensor precision of 1%	Suitable for both lung and heart sounds	Reading time = 10 sec Tiny tip suitable for neonates and young children
Connectivity	Connects to phone via Bluetooth	Connects to phone via WIFI	None
Consumables	 Finger probe suitable for neonates and young children	None	None

Table 4.3: Point-of-care devices used in the study

4.3.2.2. User-centred design

User-centred design principles were used to work with the user to iteratively develop the design of the mobile application. Feedback from a wide range of users (community health workers, nurses and doctors) was incorporated to ensure the data collection process ran smoothly and to mitigate errors. For example, results from hospital tests (blood tests and X-rays) often took time to obtain and hence diagnosis for many cases was delayed or revised. Consequently, for the hospital doctor, the mobile application featured the option to revise and add data for each case retrospectively. Feedback from study participants and community health workers was also incorporated. In particular, the community health workers often struggled to keep the children calm during examinations. Some children found the devices intriguing and insisted on playing with the pulse oximeter probe rather than keeping it on. Other children were terrified by them and cried. To address these challenges, a series of children's videos were added to the mobile application, to distract the children during the examination process. Additionally, toys were provided to each community health worker and a gamified routine was developed for the examination; a baby-sized doll was used as a demonstration to convince the child that the examination was a game which involves wearing the devices (Figure 4.4).

Figure 4.4: Examination with a community doctor demonstrating the use of toys to keep the child calm.

4.3.2.3. Mobile application

A mobile application was designed in Android Studio to govern the data collection process, including the collection of measurements via the digital devices. The mobile application was translated and presented in three languages – English, Hindi and Marathi – to facilitate use by a range of users and participants. The interface and structure was iterated multiple times to incorporate feedback from the different users (community health workers, nurses and doctors). The user experience of the final version of the mobile application is presented below:

- **Login**: Each user had a unique username and password which they used to log into the application;
- **Activities**: At the start of each case, the user could select if they want to pursue one of three options: (1) register a new case; (2) add information to an existing case; or (3) review an existing case.
- **Consent:** When a new case was registered, the adult accompanying the child was presented with full information regarding the study, including its purpose and objectives, as well as the examination process involved. This information was presented in the mobile application in both written and audio forms (for illiterate participants) in three languages (English, Hindi and Marathi); each participant was also given a study information sheet for their reference, including a contact number for any queries. The accompanying adult was then asked to confirm that they have the legal right to make decisions regarding the child's health, and to provide formal consent. The latter was accomplished by both filling in their name in the mobile application as well as signing a paper consent form. Screenshots of the consent screens of the mobile application are included in Figure 4.5.

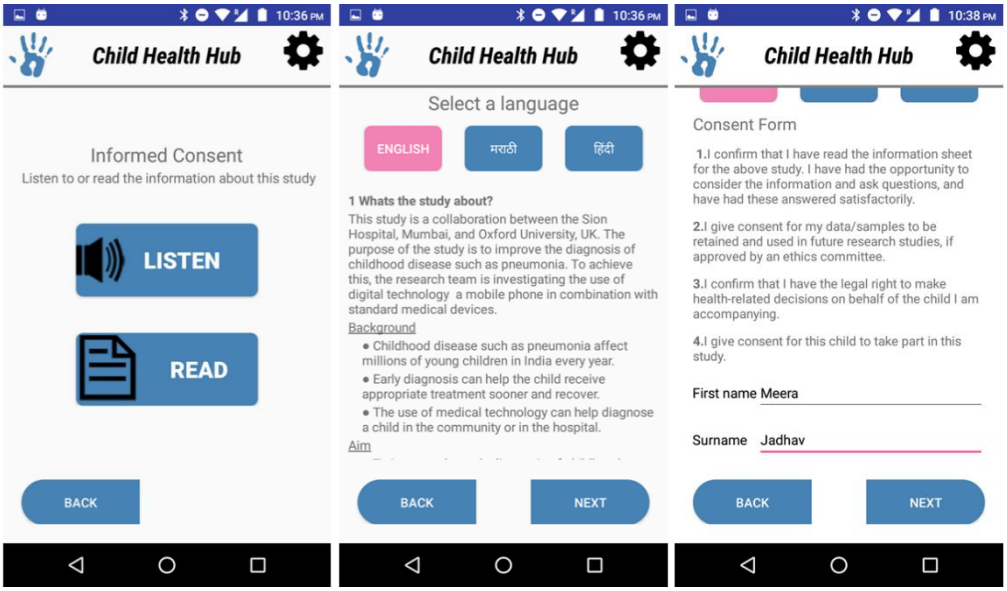


Figure 4.5: Consent screens of the mobile application.

- **Demographic data:** The age and gender of each child were recorded, as illustrated in Figure 4.6. This information was combined with the first letter of the child's name, surname and their mother's name to generate a unique ID number, e.g. abanmo026. The ID number was provided to the accompanying adult alongside the study information sheet. However, if they lost the sheet, the child's record could still be accessed by providing the same information to reconstruct the ID. The identifiable personal information (names of child and mother) was NOT stored.

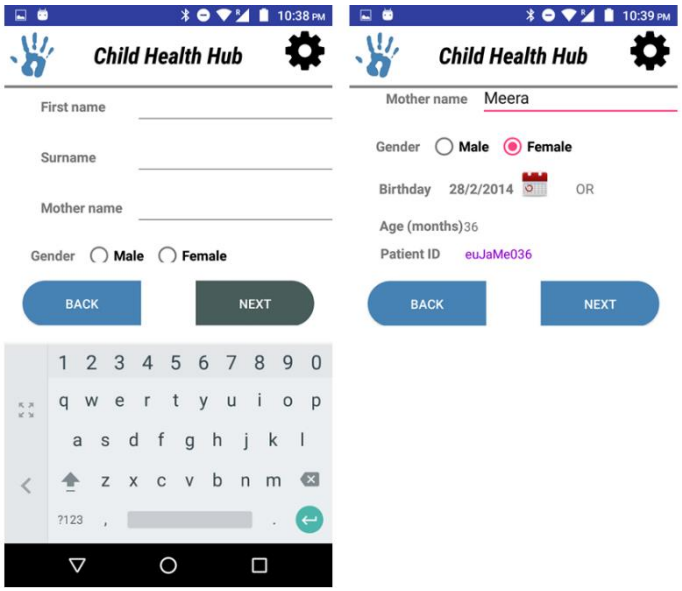


Figure 4.6: Demographic data screens.

- **Questionnaire:** A series of questions were answered by the accompanying adult; these included child health history and observational data as per the IMCI guidelines (see Chapter 1 for full details). The layout of the questionnaire is illustrated in Figure 4.7.

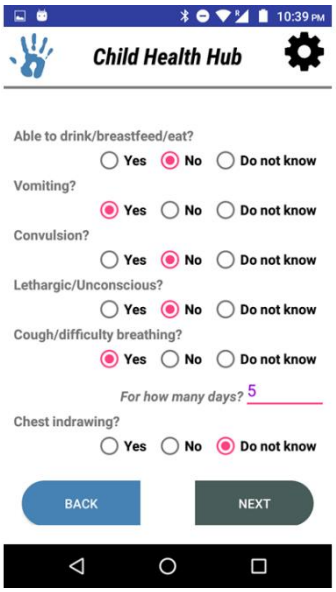


Figure 4.7: IMCI questionnaire within the mobile application.

- **Measurements:** Videos were included to illustrate how the devices should be utilised and to explain any device settings the user might face. These were used as supplements and reminders against the initial training all users received (details in Section 4.3.3).

Measurements with the pulse oximeter and the digital stethoscope were collected simultaneously (Figure 4.8, left). First, the pulse oximeter was placed on the child's finger and recording was initiated via a single button on the mobile application. Second, a connection was established with the digital stethoscope via another button. Measurements with the digital stethoscope were collected at four chest locations on the child's front and four chest locations on their back. At each of these positions, the user placed the stethoscope on the child chest, after tapping the corresponding position on a sketched schematic on the mobile application screen; this recorded the measurement against that position and started a displayed timer for 15 seconds to guide the user. The mobile application requires the user to perform all eight measurements before allowing them to advance using the 'Next' button. Some children started crying during this examination, which made collecting all eight lung recordings difficult. Consequently, a 'crying' button was added to the mobile application which, when pressed, reduced the required number of recordings

to four. Additionally, popular, and locally appropriate, children's videos were added to the mobile application to help the user calm down the children and keep them distracted during the examination.

In addition to the pulse oximeter and digital stethoscope recordings, two more measurements were performed: temperature and respiratory rate (Figure 4.8, centre and right). A standard digital thermometer was used to measure the temperature under the child's armpit and this was manually entered into the mobile application by the user. Doctors, nurses and most community health workers were already familiar with counting respiratory rate based on chest movements. To simplify and standardise this measurement, a timer of 60 seconds was added to the mobile application, and the user could either manually count the breaths and input then after the 60 seconds, or simply tap the screen of the mobile phone for every breath. The entry range for both temperature and respiratory rate was restricted to minimise errors and eliminate non-physiological data.

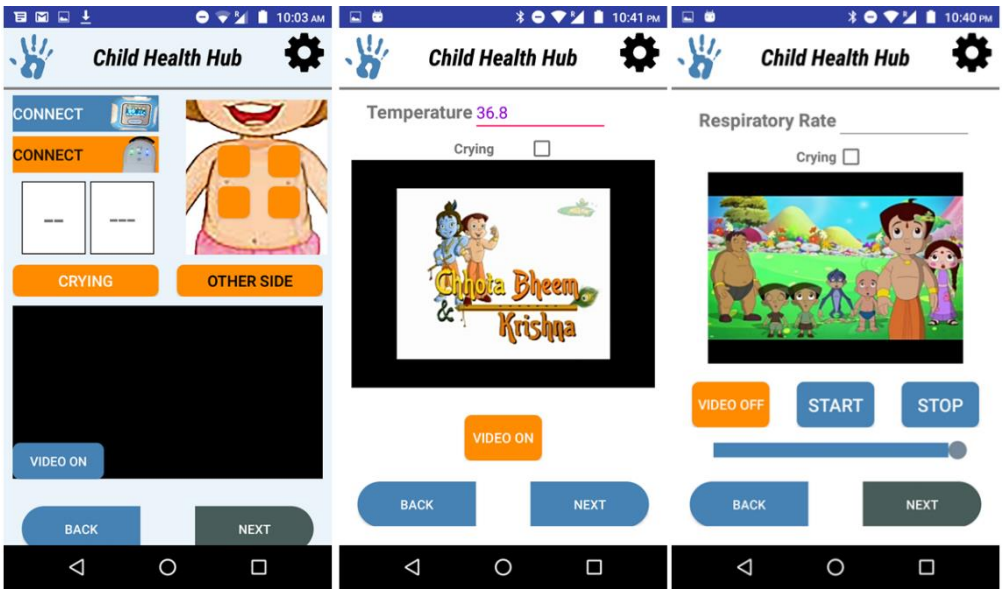


Figure 4.8: Measurement screens for stethoscope & pulse oximeter (left), temperature (middle) and respiratory rate (right).

- **Further tests:** For data collected at the hospital, the mobile application was able to capture and store data on further tests required to reach a final diagnosis (Figure 4.9). In particular, the

hospital doctors provided data on blood tests, X-ray readings, and lung sounds they heard during auscultation.

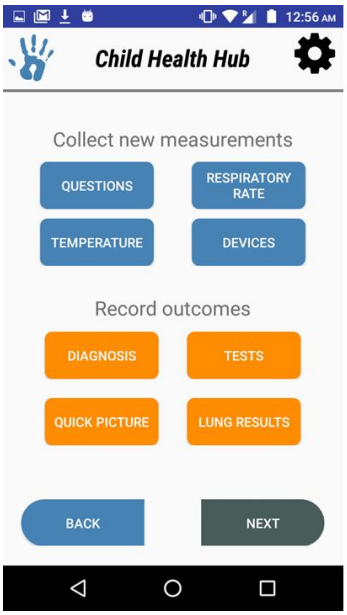


Figure 4.9: Further tests screen of the mobile application.

- **Diagnosis:** Diagnosis was recorded in the mobile application by both the community doctors and the hospital doctors. The specific gold standards followed to reach diagnosis were described in Section 4.3.1.3.
- **Referral:** There were two types of referral possible in the study: (1) referral by the community health worker to the community doctor; and (2) referral by the community doctor to the hospital. After taking all the measurements, the community health worker was asked to indicate in the mobile application if they would like to refer the child to the community doctor. However, the IMCI guidelines were implemented and if the health worker advised no referral but IMCI indicated that referral was necessary, a message was displayed on the screen, prompting the health worker to refer. Disagreement in the opposite scenario (i.e. health worker advises referral, IMCI doesn't) was not reflected in the mobile application to mitigate interference with normal clinical processes that a community health worker would follow.
- **Existing cases:** There were a number of scenarios where an existing case would need to interact with the mobile application again; a child referred from CHW to community doctor; a child referred from community doctor to hospital; or a child admitted to hospital for suspected

pneumonia that is being continually followed by the doctor. For all of those situations, the user ID was used to ensure that any new data were captured against the existing record, and so that existing data could be reviewed for the purposes of follow-up doctor examination.

- **Backend infrastructure -** All data was collected against an SQLite database. Initially, an AWS server connection was implemented to enable instantaneous synchronisation of new data with a cloud database. However, testing revealed that the mobile data connection in the communities (located in slums) was insufficient to support this. Consequently, all new data captured was stored on the mobile phone and uploaded at the end of each day.

4.3.3. Training and implementation

First, a training session was conducted with the hospital nurse and doctor, where they were taught how to use the mobile application and conducted a few trial sessions in the hospital with incoming patients. Another training session was then conducted with the two community doctors, to both familiarise them with the mobile application and to ensure that the same gold standards for diagnosis were followed. Finally, three training sessions, each half a day in length, were conducted with the five community health workers involved in the study. It should be noted that the community health workers participating in the study were previously employed by Apnalaya in the delivery of a range of projects (education, employment, health). As such, they had minimal previous health training and experience. The first training session focused on the use of the mobile application, without the devices, and generated user feedback that was used to iterate the design of the mobile application. The second training session was conducted with the hospital doctor (paediatrician) in attendance, who was responsible for providing clinical training (explaining the significance of different symptoms) and teaching common techniques for keeping young children calm when conducting an examination. During this session, special attention was given to the process of respiratory rate counting to ensure a uniform standard was followed by all five community health workers. The community health workers were then instructed to spend a week collecting data on their own. These data were analysed to identify common errors and data inaccuracies as well as any user-specific issues. Finally, a third training session was conducted in the field, again with the hospital doctor in attendance, where the problems identified during the trial week were addressed.

For example, some users were not applying sufficient pressure when recording data with the digital stethoscope, leading to increased amounts of background noise.

Beyond this training, a data review process was put in place to ensure consistency in data quality. Namely, data from both locations (hospital and community) were reviewed on a weekly basis to identify any gaps or issues with recording quality. If systematic issues were observed, the study manager for each site (the doctor at the hospital and a programme manager in the community) was responsible for providing additional instructions to the relevant study staff to eliminate issues.

4.4. Data summary – Indian (Mumbai) dataset

The collected dataset contained data on 1181 cases – 285 hospital cases and 896 community cases - with a range of conditions, collected across the two locations (hospital and community). A CONSORT study flow chart, including a breakdown of the incidence of disease categories, is presented in Table 4.4. Clinically confirmed diagnosis was available for 542 cases, with: 109 pneumonia cases (19.9%), 286 cases of other respiratory diseases (52.8%), 97 cases of non-respiratory diseases (17.9%) and 50 healthy controls (9.2%). It should be noted here that the majority of 'other respiratory diseases' in the community were upper respiratory tract infection cases, whereas in the hospital there was a variety of more severe respiratory conditions, including bronchiolitis, hyper-reactive airway disease and TB. The remaining 608 cases did not exhibit any of the IMCI danger signs and were therefore not referred to the doctor.

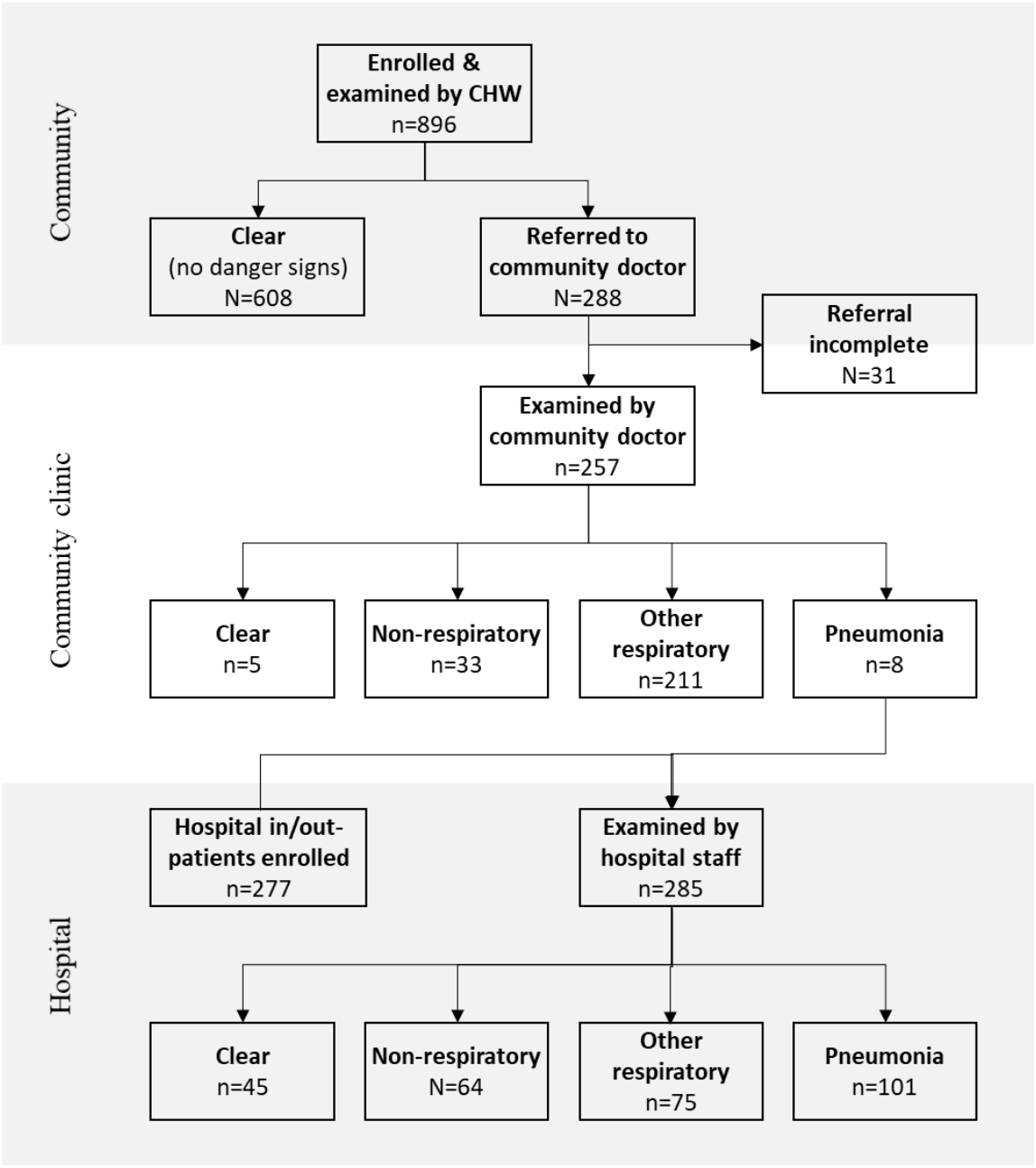


Table 4.4: Incidence of conditions across the two study sites. Unknown cases in the community are those which presented with no danger signs and were therefore not referred to the community doctor.

4.4.1. Clinical characteristics

A full list of the clinical characteristics recorded and their incidence across the dataset is included in Table 4.5. Children from the community where the study took place are typically taken to the hospital when their symptoms have aggravated. Consequently, it is expected that the community and hospital

parts of the dataset will differ in disease incidence rates as well as symptom severity within diseases. Differences in the distribution of clinical characteristics across community and hospital cases are visualised in Figure 4.10. Differences between the community and hospital cohorts were found in terms of age ($p=0.003$), respiratory rate ($p=1.3x10^{-10}$), heart rate ($p=3.7x10^{-10}$), temperature ($p=2.6x10^{-23}$) and oxygen saturation ($p=0.013$).

4.4.2. Lung sounds

The dataset contained 11,331 recordings from 1,181 participants. Many recordings contained background noise that typically stemmed from people talking in the background or the child crying during the examination. The digital stethoscope performed no automated filtering so many recordings also contained heart sounds (heart beats). Figure 4.11 includes three examples of recordings (8 seconds each), with the signal plotted in both the time and frequency domain.

The doctor at the hospital recorded in the mobile application any sounds they heard whilst examining the child. Figure 4.12 contains a summary of the different lung sounds reported as a percentage of all cases in a given disease category. From this it can be seen that 72.5% of pneumonia cases were reported to have crackles, compared to only 2% of cases with other respiratory conditions. Wheeze, on the other hand, was reported in 10.1% of pneumonia cases and in 17.4% of cases with other respiratory conditions. The interpretation of lung signals can be a fairly subjective process, dependent on the clinician's expertise and professional experience. Hence, inter- and intra-user variability is known to be common. To explore this further, seven clinical experts were involved in annotating the lung recordings acquired in the Indian (Mumbai) study and details of this process are described below.

Indian dataset (Mumbai)	**Controls**			**Pneumonia**		
	N = 804			*N = 314*		
Hospital	Mean	Median	Std	Mean	Median	Std
Gender [%, female]		42.4			38.6	
Age [months]	21.9	25	17.2	21.9	17	15.8
Respiratory rate [breaths/min]	48.2	38	9.4	48.2	46	9
Heart rate [beats/min]	137.7	125.6	21	137.7	137.6	17.4
Oxygen saturation [%]	96.9	98.2	3.4	96.9	97.7	3.3
Temperature [°C]	36.8	36.1	0.7	36.8	36.4	1.1
History of fever (%)		66.3			98.0	
Duration [days]	7.3	4.0	3.8	7.3	5.0	4.9
Malnutrition (%)		10.3			16.3	
Z-score	-	-	-	-	-	-
Unable to eat/breastfeed (%)		92.0			83.7	
Lethargic (%)		13.1			17.3	
Chest indrawing (%)		32.6			84.7	
Cough (%)		53.1			93.9	
Duration [days]	5.3	4.0	4.7	5.3	5.5	5.5
Difficulty breathing (%)		53.1			93.9	
Dehydration (%)		4.6			8.2	
Stridor (%)		-			-	
Grunting (%)		-			-	
Crackles (%)		2.3			73.3	
Wheeze (%)		18.6			7.9	
Bronchial breathing (%)		-			-	
Head nodding (%)		-			-	
Cyanosis (%)		-			-	
Pallor (%)		-			-	
Blood culture (%)		-			-	
Haemoglobin [g/dL]	9.4	10.2	2.3	9.4	9.5	1.9
White blood cell count [billion/litre]	1.70E+04	1.30E+04	1.00E+04	1.70E+04	1.30E+04	2.10E+04
Neutrophils [billion/litre]	-	-	-	-	-	-
Lymphocytes [billion/litre]	-	-	-	-	-	-
Monocytes [billion/litre]	-	-	-	-	-	-
Platelets [billion/litre]	3.40E+05	3.70E+05	2.00E+05	3.40E+05	3.30E+05	1.80E+05
Chest X-rays						
Normal (%)		19.8			37.6	
Hyperinflation (%)		8.5			1.0	
Pleural fluid (%)		5.1			5.0	
Lobar changes (%)		2.8			12.9	
Minor patchy changes (%)		0.6			3.0	
Major patchy changes (%)		0.6			3.0	
Positive X-rays [%]		-			-	
Other symptoms						
Diarrhoea (%)		18.9			13.3	
Duration [days]	6.3	3	2.2	6.3	3	9.9
Blood in stool (%)		1.7			1.1	
Rash (%)		2.9			13.3	
Runny nose (%)		33.7			56.1	
Red eys (%)		4.0			4.1	
Ear pain (%)		4.0			6.1	
Ear discharge (%)		4.6			6.1	

Table 4.5: Summary of the clinical characteristics in the Indian dataset (Mumbai study).

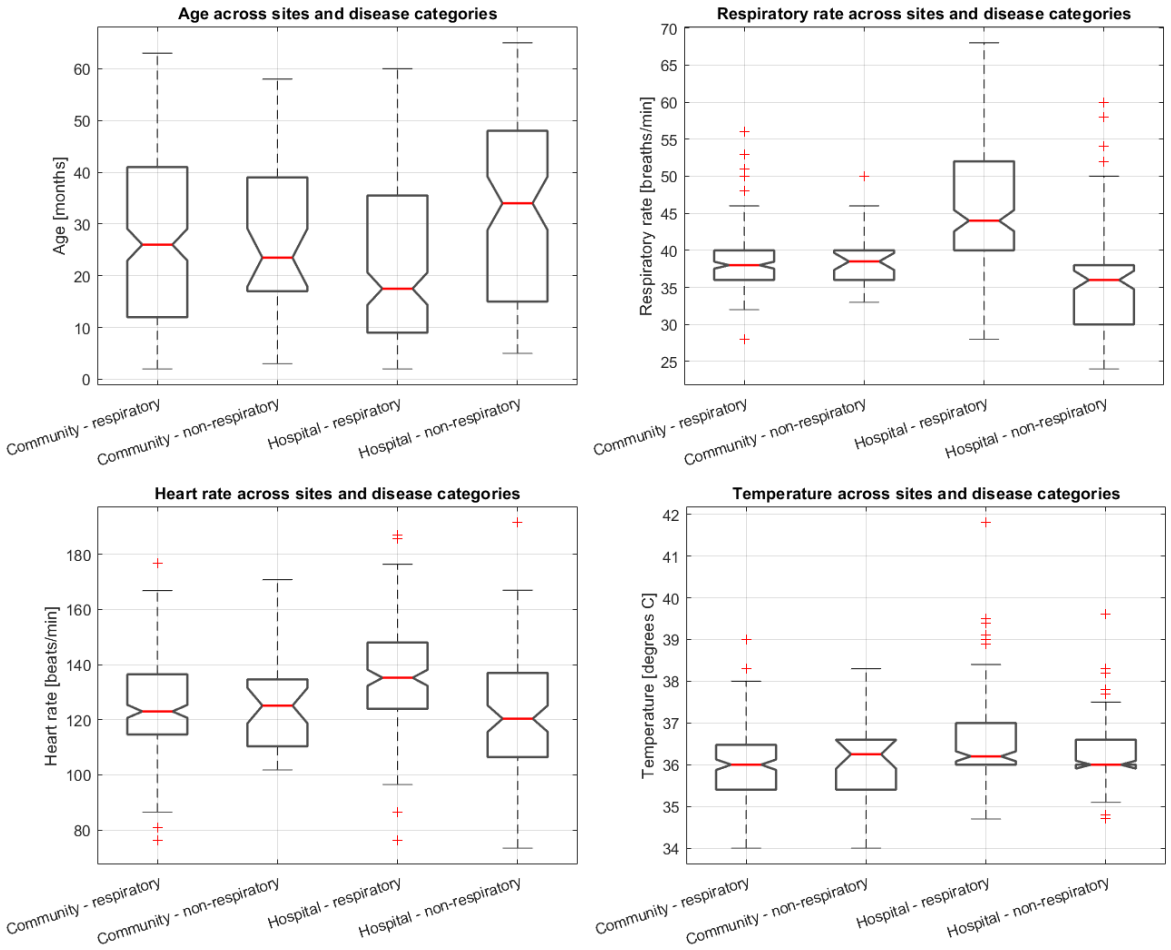


Figure 4.10: Boxplots of age, respiratory rate, heart rate and temperature across respiratory and non-respiratory cases collected at the hospital and in the community.

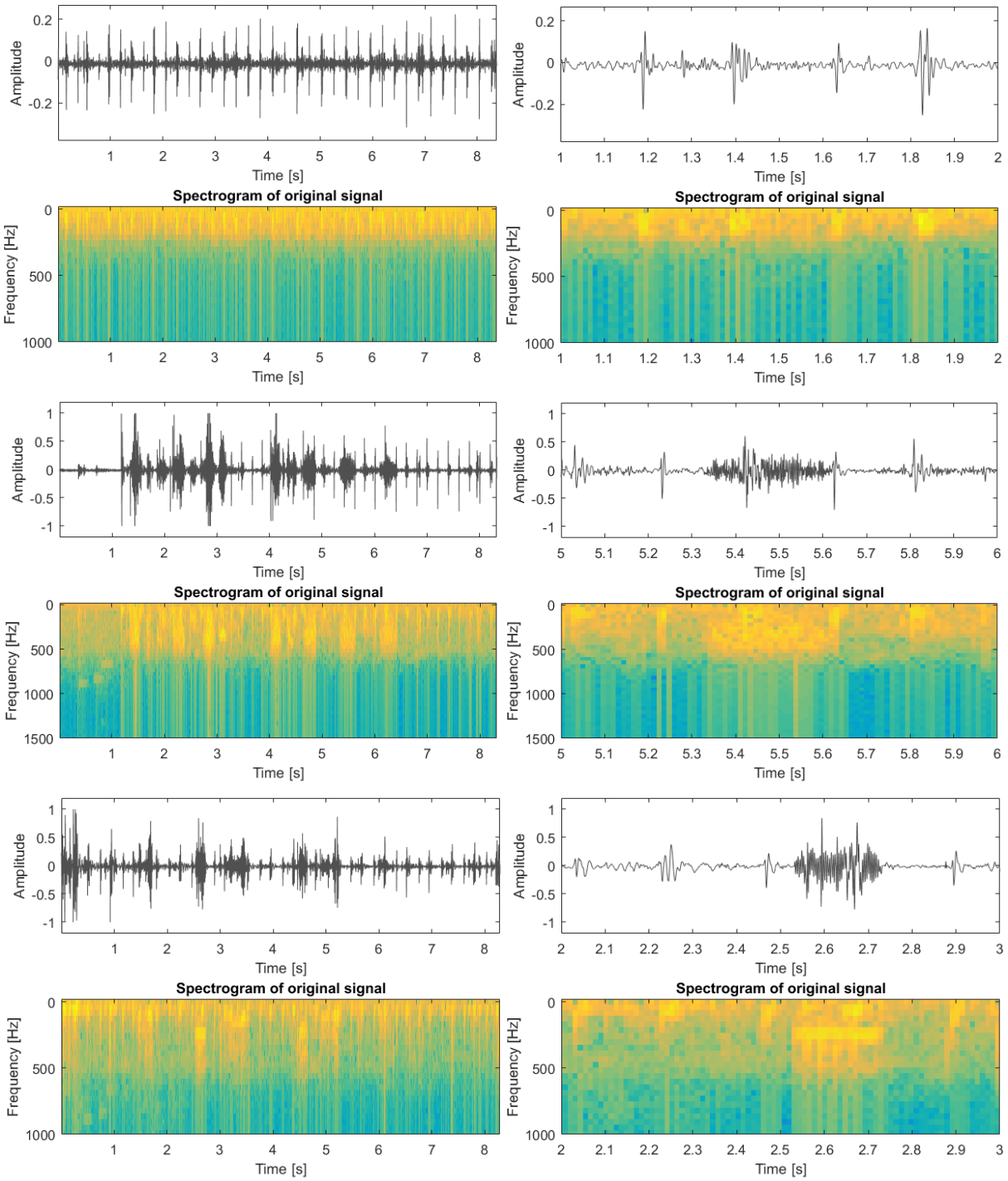


Figure 4.11: Examples of lung recordings in the Indian (Mumbai) dataset. The signal is plotted in the time domain and in the frequency domain (immediately below each time domain plot). Zoomed-in sections of the signal are included in the right column.

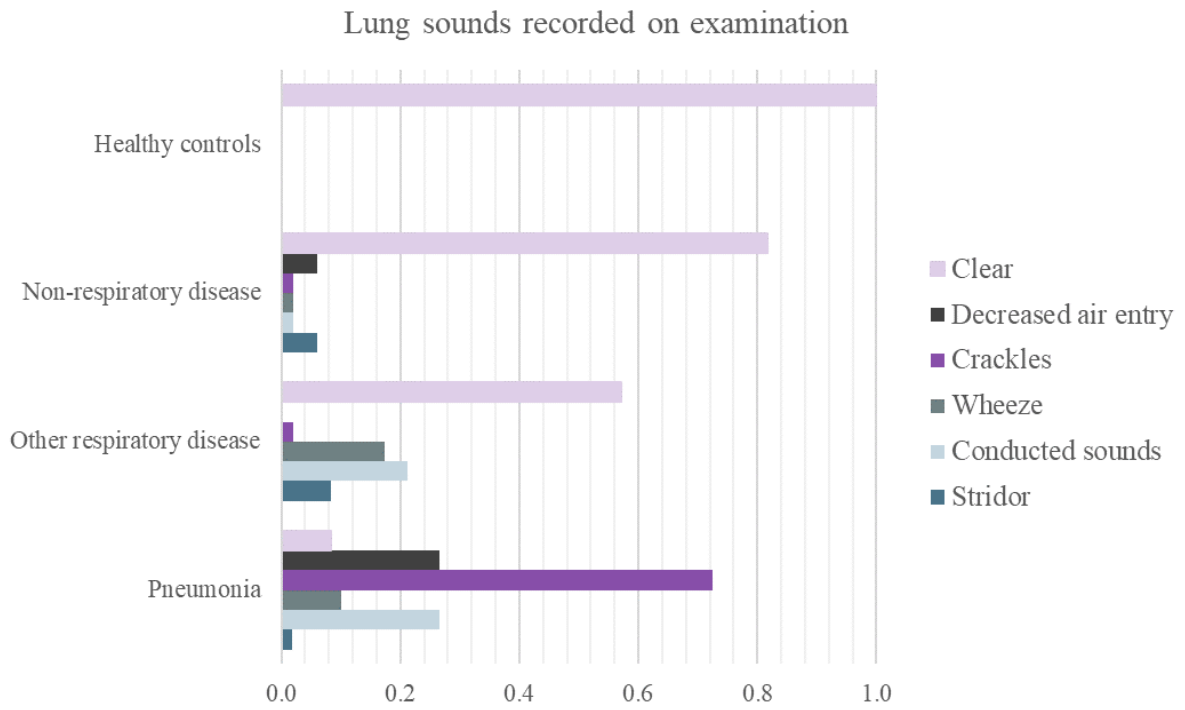


Figure 4.12: Lung sounds recorded by the doctor upon examining the child during the study in Mumbai.

4.4.2.1. Lung sound annotation process

Seven clinicians – 4 paediatric pulmonologists from the John Radcliffe Hospital (Oxford, UK) and 3 paediatricians from the Lokmanya Tilak Municipal General Hospital (Mumbai, India) were involved in annotating the lung recordings collected in this study. Throughout this thesis, annotators from the John Radcliffe Hospital are referred to as JR 1, JR 2, JR 3, JR 4, and annotators from the Lokmanya Tilak Municipal General Hospital in Mumbai (Sion Hospital) are referred to Sion 1, Sion 2, Sion 3 and Sion 4, where Sion 4 was the doctor responsible for the data acquisition. A set of 700 recordings (from 343 participants) were labelled by all 7 annotators to be used as a baseline to compare performance across annotators. The remaining samples were each annotated by at least 3 annotators. Each annotator was instructed to listen to a recording and assign the following labels: (1) crackles; (2) wheeze; (3) bronchial breathing; (4) stridor; (5) normal; (6) can't interpret. Assigning multiple labels to a recording was possible. Each annotator was presented with a few recordings repeatedly, without being notified, in order to assess intra-annotator variability. Annotation was done entirely independently and annotators did not consult each other on the labelling of individual recordings.

Labelling was done on the basis of audio, i.e. the annotators were not presented with signal visualisation. Additionally, any labels were assigned to the whole recordings, without specifying the time point at which the sound occurred (i.e. not the type of annotation that the Peru data had). Hence, if a signal was assigned a label of 'crackles' that meant that the lung sound occurs somewhere in that recording. This annotation approach was selected as a result of a number of consultations with clinical experts who deemed this process to be closest to their clinical practice. Namely, auscultation in clinical practice does not typically involve the doctor having access to a signal visual and does not require them to localise the exact time stamp of specific lung sounds.

In addition to the 7 annotators, the annotation of the doctor who performed data collection in the hospital was also available. However, their labelling was only done on a 'per-patient-basis', i.e. deciding whether a participant has a given pathological lung sound, rather than on the basis of individual recordings.

4.5. Conclusions

This chapter presented the details of a field study conducted in India in order to capture data that best reflect the challenges of pneumonia diagnosis in low-resource settings and hence provide an opportunity to develop machine learning tools for evidence-based community diagnosis. The dataset collected included a wide range of diseases as well as different types of data – clinical characteristics and signals. Consequently, the data were used in Chapter 7 for the development of generalisable models for automated diagnosis and in Chapter 8 for the development of techniques for lung sound analysis. The latter is compared against the labelling performed by clinical annotators – results are reported in Chapter 8.3.

This page is intentionally left blank

5. Analytical Tools for Pneumonia Diagnostics

This work involves the use of a number of different analytical tools, ranging from statistical tests to signal processing and machine learning techniques. This Chapter introduces the theoretical framework behind these techniques which were used in two main types of analysis: (1) combining a range of characteristics/features to derive an overall prediction of outcome (e.g. diagnosing pneumonia on the basis of multiple symptoms); (2) extracting a range of characteristics from a time series in order to evaluate association between the time series and an outcome (e.g. analysing lung signals to identify pathologies indicative of pneumonia). Consequently, this Chapter introduces methods for: feature extraction (from time-series/signals); feature selection; classification; visualisation.

5.1. Simple statistical analysis

When analysing the relevance of individual symptoms to a diagnosis, a number of statistical tests designed to quantify the association between a feature and an outcome are the simplest starting point. Similarly, statistical tests can be used to evaluate differences between two features (e.g. comparing the distribution of symptom values between populations).

- **Mann-Whitney**: a nonparametric test for statistical comparison of two populations, used when the dependent variable is either ordinal or continuous; there is no requirement for the population values to be normally distributed.
- **Chi-Squared**: an independence test used to evaluate the likelihood of a significant association between two categorical variables. The null hypothesis assumes that there is no association.
- **Two-sample Kolmogorov-Smirnov test**: a nonparametric test for goodness of fit, used to evaluate if two samples come from the same distribution.

Whilst such techniques provide preliminary indication of feature/symptom relevance, they are insufficient for the prediction of an outcome/diagnosis from a multitude of such inputs. Hence, we will next explore more advanced methods for extracting features and then combining them to classify disease.

5.2. Feature extraction for lung sound analysis

Medical data are often captured in the form of time series (e.g. digital stethoscope recordings) which need to be analysed for clinically relevant information to be extracted and linked to overall health outcomes. Hence, this section presents a number of techniques that can be used for the automated analysis of lung signals; these techniques are put into practice in Chapter 8.

5.2.1. Annotator performance evaluation

Before, we introduce techniques for future extraction, we should explore methods for establishing a clinical gold standard for the interpretation of lung signals, against which such techniques can be applied. Lung signals can be labelled by a group of clinical experts to try to establish such a gold standard. To analyse the annotators' labels, a number of statistical tools are of key importance and are introduced here and applied to the Indian (Mumbai) data in Chapter 4.

5.2.1.1. Interobserver variability

In order to evaluate the reliability of the gold standard, interobserver variability should be assessed. Two popular ways for quantifying this have been described in the literature and are presented below.

- **Feliss' kappa (Fleiss κ)**: Kappa statistics measure the observed level of agreement between observers/annotators and corrects for agreement that would be expected by chance. Unlike other kappa metrics, such as Cohen's kappa which are typically used for two annotators, Fleiss κ can be used for multiple annotators and is defined as:

$$\kappa = \frac{p_o - p_c}{1 - p_c}$$

  Where p_o quantifies the observed agreement and p_c denotes the estimated chance agreement. Full derivation of κ is included in the Appendix. The derived measure is standardised, ensuring clear interpretation and enabling comparison across studies. Namely, the kappa values are typically interpreted as follows: 0–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.80–1.0 almost perfect agreement (Landis and Koch 1977).

- **Specific agreement (SA)**: SA is a common index of the reliability of categorical measurements. For multiple annotators, SA is the probability of a randomly chosen annotator assigning an item

to a given category given that another randomly chosen annotator has also assigned that item to that category. For multiple annotators and multiple categories, SA is defined as:

$$SA_k = \frac{\sum_{i=1}^{n'} r_{ik}(r_{ik} - 1)}{\sum_{i=1}^{n'} r_{ik}(r_i - 1)}$$

Where n' is the number of items labelled by two or more annotators, r_{ik} denotes the number of annotators that assigned item i to category k, and r_i is the number of annotators that assigned i to any category. Since it quantifies a probability, SA can be interpreted intuitively.

5.2.1.2. Majority voting

The performance of individual annotators can be compared against a majority vote of all annotators, where the majority vote for the i^{th} case can be defined in the following way:

$$v_{ci} = \frac{1}{n_i} \sum_{j=1}^{J} a_{ijc}$$

where n_i is the total number of annotations that were given to the i^{th} case; j denotes the annotator index j = (1, …, J); and a_{ijc} is the individual label assigned by the j^{th} annotator to the i^{th} sample for lung class c = (1, …, C), such that:

$$a_{ijc} = \begin{cases} 1 & \text{if the } j^{th} \text{ annotator has labelled the } i^{th} \text{ case to belong to class c} \\ 0 & \text{if the } j^{th} \text{ annotator has labelled the } i^{th} \text{ case not to belong to class c} \end{cases}$$

The majority vote can then be converted into a binary by assuming a 0.5 threshold, i.e. at least 50% of the annotators should have assigned a given label to a given case, or it can be used as a continuous value, indicating a degree of agreement.

5.2.2. Signal processing for feature extraction

Signal processing techniques can be applied to time-series data to analyse their content and to extract secondary characteristics that quantify different aspects of the time-series and could be indicative of overall patient outcomes. A literature review of such techniques in the context of pneumonia was presented in Chapter 2; based on this, high performing techniques were identified and are introduced in more detail next.

5.2.2.1. Signal windowing and feature extraction

The process of feature extraction transforms originally high dimensional vectors into lower dimensional vectors and hence can be considered as a data reduction process designed to capture the essential characteristics of the analysed signal. To do this, the signal is often segmented into frames using short and overlapping window functions, as illustrated in Figure 5.1. We will next review a range of feature extraction techniques that have been seen to deliver promising results in the context of lung sound analysis, as first highlighted in Chapter 2.

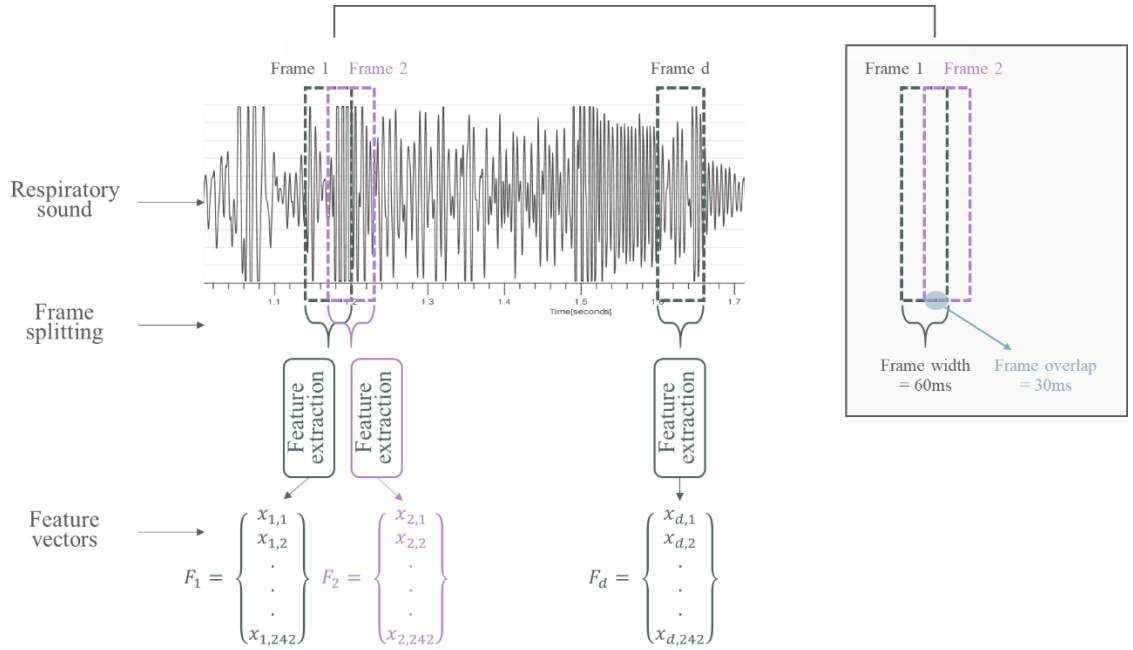


Figure 5.1: Feature extraction process – the signal is broken down into overlapping frames and a set of feature extraction techniques is applied to each frame.

5.2.2.2. Autoregressive modelling

Autoregressive modelling has been explored as a source of signal features for lung sound classification in a few papers (Dorantes-Mendez et al. 2008, Martinez-Hernandez et al. 2005). An AR framework applied to stethoscope signals makes a couple of key assumptions: (1) a linear filter, fed with a white noise signal, describes the process of lung signal generation; (2) the output characteristics can be derived from past and current values of the output. Iyer et al. proposed a physiological justification to the above-stated assumptions, rationalising the choice of AR-models for characterisation of lung sounds (Iyer, Ramamoorthy, and Ploysongsang 1989). Namely, breath sounds originate inside the airways of the lung and can be thought of as white noise sequences as they are produced by turbulent flow in numerous

airway pockets. The parenchyma and chest wall represent an acoustic transmission filter which generates lung sounds as an output. Moreover, wheeze and rhonchus are said to be produced by periodic oscillations of the air and be best modelled by sequences of periodic impulses. Crackles, on the other hand, are believed to be generated by bubbling resulting from passing of air through extraneous liquid or sudden opening/closing or airways and are better modelled as random intermitted impulses.

In the AR model, each sample $x(\mathrm{k})$ is defined as a linear combination of the preceding samples and an error term:

$$x(\mathrm{k}) = -\sum_{i=1}^{\mathrm{P}} a_{\mathrm{i}} x(\mathrm{k}-1) + e(\mathrm{k}), \quad \mathrm{k} = 0,1,2,\ldots$$

where a_{i} are the AR coefficients, P is the order of the AR model and $e(\mathrm{k})$ is a white noise sequence. The AR coefficients are optimised by minimising the variance in $e(\mathrm{k})$ and can be used as features for the classification of lung sounds.

5.2.2.3. Empirical Mode Decomposition

The basis of Empirical Mode Decomposition (EMD) is representing the data through intrinsic mode functions (IMFs). IMFs are oscillatory waves of local zero mean obtained through a sifting process. One of the main advantages of EMD is that it preserves the physical properties of the signal, with IMFs constituting a complete "orthogonal, local and adaptive decomposition" (Huang 1998). Therefore, a complex signal $s(\mathrm{t})$ can be expressed through the following expansion including IMFs:

$$s(\mathrm{t}) = \sum_{\mathrm{k}=1}^{\mathrm{N}} IMF_{\mathrm{k}}(\mathrm{t}) + r_{\mathrm{N}}(\mathrm{t})$$

Where N denotes the total number of *IMF*s and $r_{\mathrm{N}}(\mathrm{t})$ denotes a residual signal. Sifting is conducted as follows:

- All extrema of the signal, $s(\mathrm{t})$, are identified;
- Upper and lower envelopes are generated via cubic spline interpolation of local maxima and minima respectively;
- A local mean, $m_1(\mathrm{t})$, is calculated by averaging the two envelopes;

- Calculate a potential IMF, $h_1(t) = s(t) - m_1(t)$. Check that h_1 satisfies the following conditions to be an IMF: (a) the number of extrema and zero crossings should not differ by more than 1; (b) the mean value of the extrema envelopes of the IMF should be zero at any point.
- For $h_1(t)$ an IMF, the residue $r(t) = s(t) - h_1(t)$; replace $s(t) = r(t)$ and repeat from (1);
- For $h_1(t)$ not an IMF, $s(t) = h_1(t)$ and repeat from (1).
- End sifting when r(t) is no longer oscillatory.

5.2.2.4. Statistical features

A range of statistical features have been suggested as a means of quantifying the morphological complexities of the lung sound signals, including kurtosis, skewness, lacunarity and sample entropy (Mondal 2014).

Kurtosis, the fourth moment of a distribution, describes the shape of the probability density function in terms of flatness or peakedness (Kim 2004):

$$Kurtosis = \frac{E[(y(n) - \mu)^4]}{(E[(y(n) - \mu)^2])^2} - 3$$

where for y(n) – the nth sample value of the lung signal, $\mu = E\big(y(n)\big)$ is the mean of the distribution, and $\sigma = \sqrt{E[(y(n) - \mu)^2]}$ is the standard deviation of the distribution, with $E(.)$ denoting the expectation operator.

Skewness describes the asymmetry of the distribution with respect to its centre point (Kim 2004) and is defined as the ratio of the third order moment and the cube of the standard deviation:

$$Skewness = \frac{E[(y(n) - \mu)^3]}{(E[(y(n) - \mu)^2])^{1.5}}$$

for $E(.)$ denoting the expectation operator.

Entropy has been used as a signal characteristic in a range of applications, including EEG modelling (Phung 2014), lung sound analysis (Mondal 2014), and heart murmur classification (Ahlstrom et al. 2006). In all these applications, entropy is used as a proxy for the "level of chaos", or the degree of complexity, in the signal. In the context of lung sounds, adventitious lung sounds have been reported to

be associated with higher complexity of the signal and hence higher entropy. In this thesis, Shannon entropy, H(X), is used, where:

$$\mathrm{H}(X) = -\mathrm{c}\sum_{i=0}^{\mathrm{m}} \mathrm{p}(x_i)\ln \mathrm{p}(x_i)$$

$X = \{x_1, x_2, \ldots, x_m\}$ - a set of random variables; c – a positive constant acting as a measuring unit and $\mathrm{p}(x_i)$ – probability of $x_i \in X$ satisfying $\sum_{\mathrm{i}=0}^{\mathrm{m}} \mathrm{p}(x_i) = 1$.

Fractal analysis is sometimes employed to compute lacunarity, recording differences in texture or information on heterogeneity (Allain and Cloitre 1991, Hadjileontiadis 2009). In particular, lacunarity features can describe the size of holes in one-dimensional signals or two-dimensional images, capturing the distribution of shapes and gaps within a set. Low lacunarity indicates homogeneity and transitional invariance, whereas high lacunarity indicates multiple gaps distributed across the set (Du 2002). There are several algorithms for computation of lacunarity (Gefen 1983, Plotnick et al. 1996, Du 2002); in this thesis, the popular and computationally simple method of a "gliding box" is used. Namely, a box of size r is placed at the start of the signal and gradually moved forward. At each step, the number of occupied sites is counted and recorded (where s denotes the box mass). Hence, over the whole signal, this process generates a frequency distribution of box masses n(s,r) which, when divided by the total number of boxes of size r, N(r), can be converted into a probability distribution Q(s,r). The first moment, μ_Q, and second moment, σ_Q, of the distribution are calculated and lacunarity is derived as $\zeta_Q = \frac{\sigma_Q}{\mu_Q^2}$. The lacunarity is computed for a range of box sizes, r. Lacunarity was first used as a feature in lung sound analysis by Hadjileontiadis to classify adventitious lung sounds (Hadjileontiadis 2009). The texture of normal and adventitious lung sounds is said to be different, with the latter being more heterogeneous and hence returning higher lacunarity values.

5.2.2.5. Mel frequency cepstral coefficients

Mel frequency cepstral coefficients (MFCCs) are a popular feature extraction technique typically employed in speech signal processing. A couple of research papers have reported on the use of MFCC for lung sound analysis, highlighting promising results (Bahoura 2009). Computation of MFCC for an audio signal involves 5 main steps:

- **Divide the signal into frames**: As most audio signals are non-stationary, breaking them up into short frames allows for the modelling of approximately stationary segments.
- **Compute the Discrete Fourier Transform (DFT) for each frame**: The Fourier transform derives the spectral content of each frame measured in Hertz.
- **Convert frequencies to Mel-scale by applying the Mel-filter bank**: The Mel-scale is a perceptual scale of pitch, modelling pitch as perceived by humans. The Mel-filter bank converts the Hertz values into Mel-scale.
- **Compute the log amplitude of the Mel-scaled spectrum**: The energy of the different frequency bins is evaluated via the power spectral density estimation derived by applying the log to the Mel-scaled spectrum.
- **Compute the Discrete Cosine Transform of the log amplitude of the Mel-scaled spectrum**: Since phase information was lost in step (2) when the Discrete Fourier Transform is applied, the Discrete Cosine Transform is used instead of the Inverse Fourier Transform to obtain the original signal. The resultant cepstral representation of the signal, the MFCCs, contains information about the rate of change at different spectrum bands.

MFCCs contain information regarding the resonances/peak energies of a sound and are indirectly linked to the impulse response of the system which generated the sound. Ellington et al. employed MFCC in their analysis of lung sounds suggesting that the chest is a solid system whose impulse response can be modelled via MFCCs (Ellington et al. 2014). In this scenario, different chest formations with their various chambers, through which the sound travels before reaching the chest wall, are likely to generate differences in the MFCC coefficients.

5.2.2.6. Wavelet Packet Transform

The Wavelet Packet Transform (WPT) can be used to perform multi-resolution decomposition of a signal. WPT requires the implementation of a pair of Finite Impulse Response (FIR) filters: low-pass filter (H) and high pass-filer (G). The WPT is derived through iteration of each branch at any level of the Wavelet Packet Decomposition (WPD) tree, as visualised in Figure 5.2. WPT decomposes an input signal, $x(n)$ into 2^j subbands, where j is the corresponding level of the WPT. Each subband corresponds

to a wavelet coefficient set: $w_k^j[m]=WPT\{x(n), j\}$, where $n = 1, \ldots, N$ and k denotes the subband, where m denotes the number of coefficients in that subband.

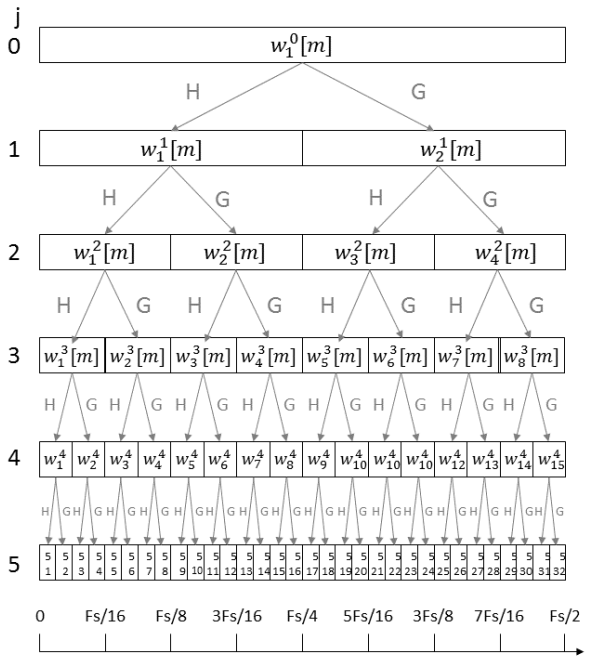


Figure 5.2: Wavelet packet decomposition tree, with j denoting the levels of decomposition, G and H denoting the low-pass and high-pass filters employed, and $w_k^j[m]$ the m^{th} coefficient of the k^{th} subband.

5.2.3. Signal modelling

In the literature review conducted in Chapter 2, we encountered a number of examples where features derived to characterise individual frames in the signal are then fed into a classical classifier to assign a label to each frame (such as crackles or wheeze). Whilst such techniques have been reported to deliver good results on small datasets, often from idealised clinical training databases, they were seen to underperform when applied to real data, rich in noise and physiological ambiguity (e.g. multiple lung sounds occurring in the same signal). An alternative approach to signal modelling aims to capture additional information regarding the transition between individual states (frames) and relies on state-of-the-art techniques for speech analysis as described next.

5.2.3.1. Classical HMM and GMM – HMM

A Hidden Markov Model (HMM) is a statistical Markov model assuming a system/signal can be approximated as a Markov process with hidden states. A Markov process is a stochastic process with the following properties:

(1) A finite number of possible states/outcomes;

(2) At any given point, the outcome depends only on the outcome at the previous point;

(3) The associated probabilities are constant over time.

An HMM, $\lambda = (\mathbf{A}, \mathbf{B}, \Phi)$, includes the following four components:

(1) A set of Q states, $S = \{s_1, s_2, \dots, s_Q\}$, where the state at time t is denoted by q_t;

(2) An initial state distribution of $\Phi = \{\varphi_i\}$ with $\varphi_i = P(q_1 = s_i), 1 \le i \le Q$

(3) A matrix of transition probabilities between states, $\mathbf{A} = \{a_{ij}\}$, where

$$a_{ij} = P(q_{t+1} = s_j | q_t = s_i) \quad \text{for} \quad 1 \le i \le Q$$

(4) A matrix of observation probabilities, $\mathbf{B} = \{b_i(o_t)\}$, where the probability of observing o_t at s_i is denoted by $b_i(o_t)$. A finite mixture represents **B**:

$$b_i(o_t) = \sum_{m=1}^{M} c_{im} \aleph(o_t, \mu_{im}, \mathbf{U}_{im}) \quad \text{for} \quad 1 \le i \le Q$$

where the m^{th} mixture in the s_i state is represented by mixture coefficient c_{im}. $\aleph$ is a density matrix with mean vector μ_{im} and covariance matrix $\mathbf{U}_{im}$. Different options for a choice in $\aleph$ exist such as a Gaussian function, turning the HMM into a GMM-HMM. An introduction to GMMs is included in the Appendix.

In using an HMM model, one should address two main challenges – learning and decoding. In the training stage, a set of model parameters $\lambda^* = (\mathbf{A}^*, \mathbf{B}^*, \Phi^*)$ is optimised to fit a true outcome, **X**, where the forward-backward algorithm is used to calculate $P(\mathbf{X}|\lambda)$ and the Baum-Welch algorithm is used to solve for $\lambda^* = \text{argmax}_\lambda P(\mathbf{X}|\lambda)$ (Rabiner 1989). In the testing stage, given a model λ, a hidden state sequence $(q_1, \dots, q_T)$ that is most likely to have produced the observations in the test data, $O = (o_1, o_2, \dots, o_T)$, is found via the Viterbi algorithm (Rabiner 1989):

$$P(O|\lambda) = \max_{q_1,\dots,q_T} \varphi_{q_1} \prod_{t=2}^{T} p(q_t|q_{t-1}) b_{q_t}(q_t)$$

Hence, for W discrete classes, we can train W different HMMs $\{\lambda_1, \dots \lambda_W\}$ and assign a new sequence O as per:

$$w^* = \arg \max_{1 \le w \le W} P(O|\lambda_w)$$

Where $P(O|\lambda_w)$ is calculated via the Viterbi algorithm.

5.2.3.2. DNN – HMM

There are a number of advantages to using GMMs for modelling the probability distribution of features derived for every state in an HMM. As described before, the EM algorithm and choice of number of components enable the fitting of multidimensional distribution to data, with a required level of accuracy. Previous research has improved evaluation speed as well as the trade-off between overfitting and generalisability (Young 1996). Yet, GMM-HMM models have been reported to underperform for data positioned near or on a nonlinear manifold (Dahl 2012). As an alternative, a hybrid Deep Neural Network – Hidden Markov Model (DNN-HMM) framework has been developed, where the output of the DNN is used to estimate the observation probabilities. DNN-HMMs have been applied to automated speech recognition tasks, delivering superior performance (Yu 2010, Seide 2011, Dahl 2011). A review article from 2012, co-authored by researchers at the University of Toronto, Microsoft Research, Google and IBM Research presented a strong case in support of DNN-HMM for acoustic modelling in speech recognition, including comparative analysis with alternative approaches (Hinton 2012).

The DNN-HMM training procedure is described in (Li 2013) and includes three main steps:

(1) For each class w (w = 1,…,W), a GMM-HMM, λ_w, with Q states is trained using the training data.

(2) The Viterbi algorithm described in 5.2.3.1, each sequence in the training set, $O = (o_1, o_2, \dots, o_T)$, is used to derive a corresponding optimal state sequence for each λ_w, $(q_1^w, \dots, q_T^w)$. A label, $L_i(i \in (1, \dots, W \times Q))$, is assigned to each state q_T^w.

(3) A DNN is trained using the training data, alongside their labelled state sequences, where the outputs are the posterior probabilities of the $W \times Q$ entities. Training of the DNN can be done

using back propagation and either unsupervised pre-training or discriminative pre-training (Li 2013). An introduction to DNN and the two training approaches is included in Appendix 1.

The test procedure includes three main steps:

(1) Posterior probabilities $\{p(L_i|o_t)\}_{i=1,\dots,W\times Q}$ are calculated by applying the DNN to the test sequence O_{test}. Then a state-label mapping table is used to map each L_i to the k^{th} state of the c^{th} model and derive the posterior probability $p(q_t = S_k^c|o_t)$ from $p(L_i|o_t)$.

(2) The likelihood probability $p(o_t|q_t)$ is calculated using Bayes' theorem:

$$p(o_t|q_t) = \frac{p(q_t|o_t)p(o_t)}{p(q_t)}$$

where $p(q_t)$ is calculated from incidences of each state in the training set and $p(q_t)$ is assumed to be constant, given the independence of each feature vector.

(3) Finally, the Viterbi algorithm, introduced in 5.2.3.1, is used to calculate the likelihood probability $p(O|\lambda_w)$, where $p(o_t|q_t)$ is used to substitute $b_{q_t}(o_t)$.

Only one research study applying DNN-HMM to lung sound data was found in the literature, where DNN-HMM was used to model adventitious vs normal lung sounds (Li 2017). The authors reported good results, 95.6% accuracy, for a reasonably sized dataset, recordings from 144 patients. However, no breakdown of performance into sensitivity and specificity was shared and the relevance of the predicted states to overall diagnosis was not investigated.

5.3. Disease classification

Once we have extracted features from clinical signals, we need to develop a model that combines them to reach an overall patient diagnosis. Supervised learning has found a wide range of applications in healthcare, where the aim of the analysis becomes to train a model/ learner on the basis of a wide range of medical symptoms and a given outcome (e.g. diagnosis of a disease). Here, we review a range of techniques that can be used to: (1) identify the most appropriate features for the classification problem; (2) develop a classification model, (3) assess its performance on unseen data in a systematic manner to evaluate generalisability, and (4) visualise multidimensional data in two dimensions to build an intuition

of the underlying structure of the data. These techniques are employed in Chapters 6, 7 & 8 to fuse a number of symptoms into a diagnosis of pneumonia.

5.3.1. Feature selection techniques

The curse of dimensionality has been shown to affect even powerful classifiers such as random forests and support vector machines (Liu 2005, Guyon 2003, Tsanas 2012). Reducing the number of features can lead to a more accurate model. However, an exhaustive search over all possible feature combinations to determine the optimal feature subset is computationally intractable. Instead dimensionality reduction can be achieved either by feature transformation (transforming a feature set to fit a lower dimensional space) or feature selection (identifying a subset of the features in the original set). Next, we review the most common feature selection techniques which have been shown to successfully quantify the predictive ability of features towards the outcome. A brief description of each technique is given, where $\mathbf{X}$ is a dataset of m features and n cases, i and j denote the feature and case indices respectively, and $\mathbf{y}$ is the outcome.

- **Correlation** (maximum relevance on the basis of the linear (Pearson) correlation coefficient):

$$\mathrm{corr}(\mathbf{x}_i, \mathbf{y}) = \mathrm{cov}(\mathbf{x}_i, \mathbf{y}) / \sqrt{\mathrm{var}(\mathbf{x}_i) * \mathrm{var}(\mathbf{y})}$$

- **Maximum relevance minimum redundancy (mRMR)**: the technique first originated from the derivation of mutual information between the feature and the outcome to quantify relevance:

$$\mathbf{I}(\mathbf{x}_i, \mathbf{y}) = \iint \mathrm{p}(\mathbf{x}_i, \mathbf{y}) * \log \frac{\mathrm{p}(\mathbf{x}_i, \mathbf{y})}{\mathrm{p}(\mathbf{x}_i) * \mathrm{p}(\mathbf{y})} \mathrm{d}\mathbf{x}_\mathbf{i} \mathrm{d}\mathbf{y}$$

  Next, the concept of redundancy was added by computing the mutual information between pairs of features. Therefore, mRMR looks for a compromise between relevance and redundancy:

$$\mathrm{mRMR} = \max_{i \in \mathrm{Q-S}} [\mathbf{I}(\mathbf{x}_i, \mathbf{y}) - \frac{1}{|\mathrm{S}|} \sum_{s \in \mathrm{S}} \mathbf{I}(\mathbf{x}_i, \mathbf{x}_s)]$$

  where Q contains the indices of all features from the original feature space, $\mathbf{x}_\mathbf{s}$ is a feature selected from the subspace of selected features S and |S| is its cardinality. In this study, a computationally fast version of mRMR that uses Spearman coefficients rather than mutual information (Tsanas 2013) was used.

- **Relief**: This is a feature weighting algorithm that makes use of the concept of *Nearest Hit (NH)* and *Nearest Miss (NM)*, where the weight associated with each feature (b_i) is derived from:

$$b_i = \frac{1}{q}\sum_{i=1}^{q}\left(-\frac{1}{|NH(\mathbf{x}_j)|} * \sum_{\mathbf{x}_k \in NH(\mathbf{x}_j)}\left\|x_{j,i} - x_{k,i}\right\| + \right.$$

$$\left.\sum_{y_i \neq y_j}\frac{1}{|NM(\mathbf{x}_j)|} * \frac{p(y_l)}{1-p(y_i)}\sum_{\mathbf{x}_k \in NM(\mathbf{x}_j)}\left\|x_{j,i} - x_{k,i}\right\|\right)$$

where q represents the number of randomly samples cases ($q = n$ for an exhaustive search), |.| denotes the size of NH and NM and ||.|| is a distance metric (Euclidian distance).

- **Gram–Schmidt orthogonalization (GSO)**: This technique uses orthogonal decomposition of features to evaluate and select them independently. To do this, the feature which best explains the outcome is selected by maximising:

$$cos^2(\mathbf{x}_i, \mathbf{y}) = \frac{(\mathbf{x}_i \cdot \mathbf{y})^2}{||\mathbf{x}_i||^2||\mathbf{y}||^2}$$

Subsequently, overlapping information is discarded from both the outcome and the remaining features by projecting them onto the null subspace; hence: $\mathbf{x}_{i,new} = \mathbf{x}_i - proj_{\mathbf{w1}}(\mathbf{x}_i)$, where $\mathbf{w_1}$ is the space spanned by **y** and $proj(\mathbf{x_i})$ (Stoppiglia 2003).

- **Least angle shrinkage and selection operator (LASSO)**: This technique uses the $L_1 -$ norm as a sparsity promoting function to both minimise the prediction error and reduce the number of features, in a classical regression setting (Tibshirani 2011). This is achieved through:

$$\mathbf{b}_{LASSO} = arg\min_{\mathrm{b}}\sum_{j=1}^{n}(y_j - \sum_{i=1}^{m}x_{ji}b_i)^2 + \lambda\sum_{i=1}^{m}|b_i|$$

where $\mathbf{b} = (b_1, \ldots, b_m)$ represents the ordinary least squares parameters and λ is a regularisation parameter controlling shrinkage.

- **Elastic net (EN)**: Building upon LASSO, EN can identify groups of variables and not select one feature arbitrarily from a set of similar features. The L_2-norm penalty is added to group variables, coupled with a second regularisation parameter (λ_2) (Zou 2005):

$$\mathbf{b}_{EN} = arg\min_{\mathrm{b}}\sum_{j=1}^{n}(\mathrm{y}_j - \sum_{i=1}^{m}x_{ji}b_i)^2 + \lambda_2\sum_{i=1}^{m}|b_i|^2 + \lambda_1\sum_{i=1}^{m}|b_i|$$

- **Sparse linear discriminant analysis (sLDA)**: Linear Discriminant Analysis identifies a lowdimensional space that maximises the separation of the classes when the data are projected ($\mathbf{X}_{new} = \mathbf{w}^T * \mathbf{X}$). Consequently, the technique looks for an optimal **w** that maximises $J(\mathbf{w}) = \frac{(\widetilde{m_1} - \widetilde{m_2})^2}{\widetilde{s_1}^2 + \widetilde{s_2}^2}$, where $\widetilde{m_1}$ and $\widetilde{m_2}$ are the multidimensional sample means of projected data, and $\widetilde{s_1}$ and $\widetilde{s_2}$ are the scatter parameters for projected samples, in the two classes for a given choice of **w**. A sparseness criterion using the elastic net penalty, as defined by Clemmensen et al., was added (Clemmensen 2011).

5.3.2. Classification techniques

Historically, logistic regression (LR) has been the most popular classification technique in the medical domain, including assessment of pneumonia (as highlighted in Chapter 2). Other, more state-of-the-art classifiers, that have been used for a range of healthcare problems are support vector machines (SVMs) and random forest (RFs). The application of all three techniques to various medical diagnostics problems has been well-documented in the literature (Vapnik 1999, Burges 1998) and a brief overview of the underlying principles is given here:

- **LR**: A widely used linear classifier, LR derives outcome predictions via the sigmoid function:

$$p(\mathbf{x}_j, a, \mathbf{b}) = 1/(1 + e^{-(a+\mathbf{b}\cdot\mathbf{x}_j)})$$

  where $a + \mathbf{b} * \mathbf{x}_j = 0$ defines the decision boundary separating the classes. The a and **b** parameters were optimised in this study through a validation approach described in the main body of the paper.

- **SVM**: The principle behind SVM is the identification of a hyperplane (f(**X**)) that creates boundaries in the feature space to differentiate the classes (Joachims 1998). In the case of linearly separable data, the hyperplane is given by:

$$f(\mathbf{X}) = \sum_{i=1}^{m} b_i \mathbf{x}_i + a = 0$$

  In the case of non-linearly separable problems, a kernel function is applied to map the data to a high-dimensional feature space before constructing the linear decision function. The LIBSVM implementation (Chang 2011) was used to optimise a Gaussian radial basis function kernel,

determining the optimal values of the kernel width γ and the penalty parameter C though a grid search of possible values.

- **RF**: Random forests is a non-parametric technique formed from multiple simple base learners (i.e. trees). The method of Classification and Regression Tree (CART) partitions the feature space into sub-regions by identifying splits (”nodes”), consequently learning an increasingly detailed mapping between the data and the outcome (Hastie 2009). Majority voting from multiple base learners is used to construct ensembles, diluting any bias in the classification of samples (Breiman 2001). The number of trees and the number of features for splitting were optimised though a grid search of possible values - [50,150,250,350,450,550] and [1,...,f] respectively, where f is the number of features and the latter range was incremented by 1.

5.3.3. Performance evaluation

Once a model is trained, its predictions can be compared against the true outcome using a number of different metrics listed below, where TP = number of true positive values, TN = number of true negatives, FP = number of false positives and FN = number of false negatives:

- Sensitivity = TP/(TP +FN)
- Specificity = TN/(TN+FP)
- Balanced accuracy = (sensitivity + specificity)/2
- Area under the curve (AUC) = the area under the receiver operating curve (ROC)
- Matthew’s correlation coefficient $MCC = \frac{\text{TP x TN} - \text{FP x FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$

5.3.4. Performance generalisation

Once a classification model has been trained, it is important to evaluate its performance on unseen data, i.e. its *generalisation performance*. This can be done internally (on a fraction of the dataset that was left out from the training procedure) or externally (on a completely different dataset). Internal validation is typically performed by techniques such as cross-validation or bootstrapping. Throughout this thesis, we use cross-validation as a well-established re-sampling technique and more details are introduced next.

The cross-validation procedure has k-folds that determine how many different parts the dataset is split into. For each one of k iterations, one of the k parts of the data is set aside for testing and the remainder

is used for the training of the model. The model parameters are optimised on the training set and performance is evaluated on the unused test set. The process is typically repeated multiple times, where the data split is randomly permuted to avoid bias and to obtain statistical confidence in the model performance. Different loss functions or performance metrics can be used to optimise the model within the training set. Additionally, in smaller datasets, leave-one-out cross validation is performed, where a single sample is set aside for testing and a model is developed on the remainder of the data. Again, this is performed on an iterative basis until the whole dataset is covered.

5.3.5. Visualisation

A dimensionality reduction technique called t-stochastic neighbourhood embedding (t-SNE) has been used for a number of challenges requiring low-dimensionality representation of high dimensional datasets (van der Maaten and Hinton 2008). Medical data, such as the pneumonia data used in this study, normally occupy nonlinear manifolds; consequently, linear methods such as principal component analysis (PCA) (Hotelling 1933) and multidimensionality scaling (MDS) (Torgerson 1952) may be insufficient as they mainly preserve the separation between dissimilar data entries within a low-dimensional space, at the expense of closure information concerning similar 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 (van der Maaten and Hinton 2008). The original t-SNE implementation in MATLAB was used throughout this thesis (van der Maaten and Hinton 2008).

5.4. Conclusions

This chapter provided an overview of the machine learning techniques relevant to the analyses conducted in future chapters. These techniques will be referenced further when their application to specific pneumonia challenges are described.

This page is intentionally left blank

6. The Tri-fold Diagnostic Challenge of Childhood Pneumonia

6.1. Introduction

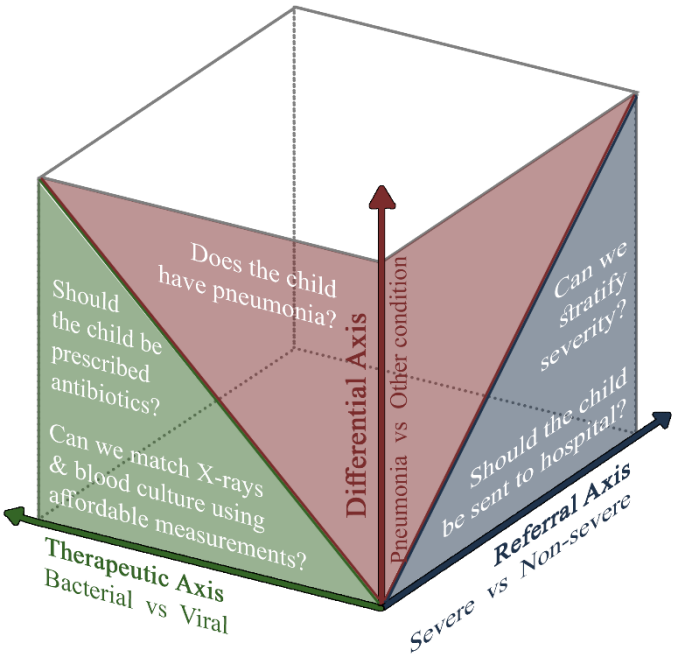


Figure 6.1: The tri-fold diagnostic challenge of childhood pneumonia, split into three axes of analysis: (1) differential axis (identification of pneumonia from other conditions); (2) referral axis (severity determination to inform need for referral to hospital); (3) therapeutic axis (aetiology determination to inform appropriate treatment).

Chapter 1 introduced the challenge of childhood pneumonia, highlighting the devastating mortality burden in low-resource settings, driven by lack of tools for early and accurate diagnosis as well as evidence-based evaluation of disease severity to inform appropriate case management. This chapter investigates the potential role of machine learning in addressing these challenges. A suite of machine learning techniques is proposed to deliver solutions for the three-fold diagnostic challenge of childhood pneumonia – identification of disease, severity determination and aetiology determination (Figure 6.1). The development of these techniques has three main prerogatives: (1) improve diagnostic accuracy and reproducibility, using WHO's IMCI guidelines for community-based diagnosis as a baseline; (2) prioritise the used of quantifiable information that can be captured reliably with affordable point-of-care tools in order for the algorithms to be applicable to low-resource settings; (3) explore the potential contribution of new biomarkers, fused with conventional clinical characteristics, as a tool for improved

diagnostic accuracy as well as evidence-based assessment of essential disease components such as severity and aetiology.

6.2. Methods

The analysis in this chapter includes four main components, laid out in Figure 6.2 and introduced in detail in the subsequent sections. Each of the three diagnostic challenges was addressed separately, following the same analytical pipeline: (1) pre-process data to address sparsity; (2) identify a minimal and most informative set of features using feature selection for the specific classification challenge; and (3) train and test machine learning models for classification of cases according to the specific clinical classes. The theoretical basis behind the techniques employed in this analysis was introduced in Chapter 5 and the subsections below include details on the exact implementation of these techniques to the challenges at hand.

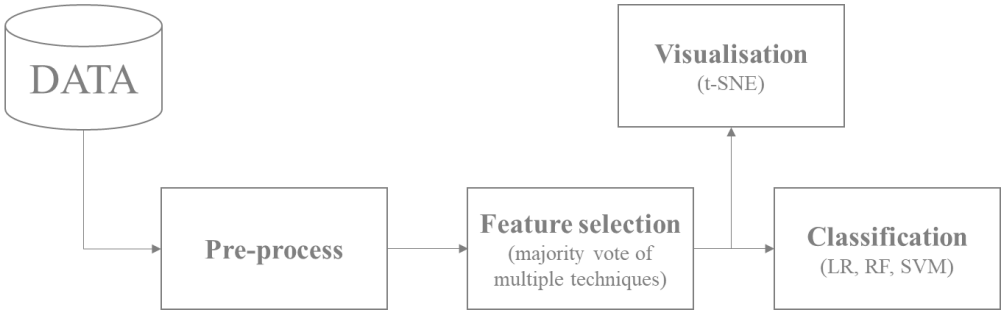


Figure 6.2: Workflow diagram for the analytical steps included in this chapter.

6.2.1. Data

The Gambian dataset, introduced in Chapter 3, was used for analysis in this chapter. For each one of the three classification problems, a different subsection of the data was used:

- **Pneumonia identification**: full Gambian dataset (1581 cases, 57 features).
- **Severity determination**: first severity dataset contained all pneumonia cases and corresponding clinical features (780 cases, 57 features); second severity dataset contained the pneumonia cases who had biomarker information in addition to their clinical features (180 cases, 61 features).
- **Aetiology determination**: pneumonia cases who had X-ray and blood test information (84 cases, 61 features).

6.2.2. Pre-processing

Pre-processing of the original Gambian data was performed. The substantial number of missing values (up to 42% for some of the features) was addressed through imputation: features and cases containing less than 85% of the total number of entries were removed; the remaining missing values were imputed using feature median values (feature mean imputation was also tested but was observed to deliver equivalent results). For features with high magnitude outliers the median is known to be a better imputation factor. Following standard statistical machine learning methodology, we extended the dimensionality of the original design matrix. Specifically, we introduced additional vectors for each feature that mirrored imputations—these vectors contained 'ones' where imputation was done and 'zeros' otherwise (Donders et al. 2006). We will refer to these vectors as 'ghost vectors' throughout this chapter. Several clinical features were excluded from the dataset as they are used in the clinical 'gold standard' and cannot be captured in a quantifiable way: 'lower chest wall indrawing', 'cyanosis', 'able to eat/breastfeed', 'cough (heard)'. In addition to imputation, all features were scaled to a similar range. For continuous valued features, this was achieved by subtracting their minimum value and dividing by their range of values (normalisation). This approach was selected over standardisation (z-score) because of the use of tSNE for dimensionality reduction; the latter uses Euclidian distances between datapoints and standardisation of features can cause features on larger scales to outweigh the rest. For discrete features, a vector of ones was added and the resulting sum was divided by the maximum values for the feature plus one, avoiding multiple zero entries. Gaussianizing of data via the Box–Cox transformation was also attempted but had no considerable effect on performance (<1% difference).

6.2.3. Feature selection

Feature selection in this study was driven by considerations related to data limitations and diagnostic application.

- **Data considerations:** An exhaustive search over all possible feature combinations to determine the optimal feature subset is computationally intractable; hence seven feature selection approaches (correlation, mRMR, Relief, GSO, LASSO, EN and sLDA), identified in Chapter 5 as suitable for medical data, were used as the basis of a majority voting approach.

- **Diagnostic application considerations:** Point-of-care diagnostics would realistically afford a limited number of features. Therefore, a feature selection approach that took into account cost, acquisition time and quantifiability of measurements was used.

A majority voting approach that consolidates results from all techniques was developed in order to dilute the bias of individual feature selection techniques and to obtain a more objective selection of features. Using 10-fold cross-validation and 50 repetitions, feature selection was performed on nine tenths of the data via each one of the seven techniques. The frequency with which each feature occupied each rank was calculated across folds and results were averaged across the number of repetitions. Additionally, normalized scores were derived to reflect the overall frequency of ranks. This approach is summarized in the pseudo-code included in Figure 6.3. Next, two inclusion conditions were applied. First, a feature should appear in the top 10 of at least three 'fundamentally different' techniques; the pairs (correlation and mRMR) and (LASSO and EN) share conceptually similar theoretical bases and are not 'fundamentally different'. Second, the feature should be measurable in a point-of-care setting in an affordable way. The scores associated with selected features were used to derive a feature ranking. The SpaSM toolbox was used to perform LASSO, EN and sLDA in MATLAB (Sjostrand); the remaining feature selection algorithms were implemented from scratch.

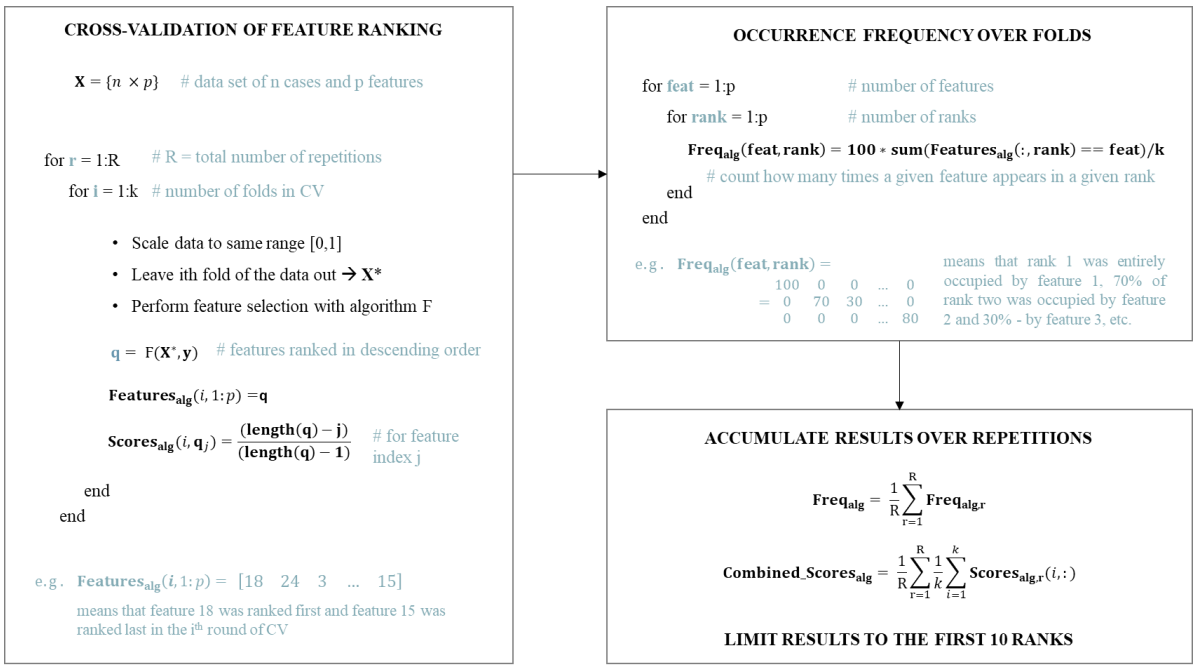


Figure 6.3: Majority voting method for feature selection; 10-fold CV with 50 repetitions are used for statistical confidence.

6.2.4. Classification

The performance of the three classifiers, identified in Chapter 5 as common and appropriate for the analysis of medical data, was compared: LR, SVM and RF. The generalization of each machine learning algorithm (i.e. its expected performance on unseen data) was assessed using either four-fold cross-validation (for the bigger data subsets) or leave-one-out (for the smaller data subsets). For each training set, pre-processing as described in 6.2.2 was performed; these operations were then applied to the test set, using characteristic parameters derived from the training set. Additionally, internal five-fold cross-validation was performed to optimize classifier parameters in each training set, based on AUC. Using the feature ranking derived in 6.2.3, an increasing number of features were gradually fed into the classifier, with a separate algorithm trained for each feature set. The following performance metrics were recorded from the test set: sensitivity, specificity, AUC, balanced accuracy and MCC. MCC and balanced accuracy were used to account for the disproportionality of outcomes, particularly evident in the aetiology dataset.

In a thorough validation approach, the steps above were repeated 20 times to offset any bias in the split of the data into train and test, and assess performance variance (which ideally will be low in order to have confidence in the reported errors). Algorithm performance in this chapter is reported in terms of mean values and variance across the repetitions. The performance of the three classifiers was assessed, identifying both the optimal number of features as well as the best performing classifier.

6.2.5. Visualisation

Childhood pneumonia presents through a complex inter-action of symptoms. Therefore, visualisation of the identified diagnostic features was necessary to link this data-driven approach back to the clinical rationale. The dimensionality reduction technique introduced in Chapter 5, t-SNE, was used to generate two-dimensional representations of the feature sets and explore the distribution of individual features across the data structure.

6.3. Results

Results for each of the three diagnostic challenges (identification, severity and aetiology) are presented separately.

6.3.1. Identifying pneumonia

Applying the feature selection methodology outlined in 6.2.3, the following feature subset was selected: respiratory rate (RR), heart rate (HR), temperature (T), malnutrition (WHZ) and oxygen saturation (Osat), listed in descending order of importance. The full feature selection outcomes can be found in Table 6.1. RR, HR, T and Osat had 13 missing values each, WHZ had 22. Values were imputed using the procedure described in 6.2.2, leading to the creation of five ghost vectors. None of the ghost vectors were ranked amongst the Top 10 on either of the feature selection techniques, but relevant ghost vectors (corresponding to the selected features) were kept as part of the feature set in order to test their effect on classification.

Applying t-SNE to the five clinical characteristics, two clear clusters were seen to emerge in the 2-dimensional representation of the data; when compared to the outcomes these were shown to be consistent with the separation between 'pneumonia' and 'no pneumonia' classes (Figure 6.4). With the exception of a few cases, the t-SNE outcome indicated that separation between the two classes should be possible on the basis of these five features. Additionally, the distribution of each individual feature, or clinical characteristic, across the 2-dimensional representation of the data was investigated (Figure 6.4). From this, pneumonia can be confirmed in cases with: RR > 65, HR > 175, T > 38.5, Osat < 87, WHZ <-5 and WHZ > 1.8. Moreover, correlation coefficients calculated for each of the features and included in Figure 6.4 indicate significant correlation between the outcome (pneumonia vs no pneumonia) and four or the features – RR, HR, T, Osat. However, the t-SNE generated structure of the data contains some uncertainty - there are multiple cases where this type of simple thresholding is insufficient to separate the two classes. A closer look at the feature distribution visualised in Figure 6.4 highlights that pneumonia cases can be present with RR<65, HR<175 and Osat>87; similarly, some no pneumonia cases present with elevated HR or RR. Extreme cases of malnutrition (WHZ≤-4) were seen to be concentrated in the pneumonia group but more moderate values (-4<WHZ<1) were equally distributed across the two classes. This preliminary analysis indicated that more advanced machine learning techniques might be needed to separate the two classes.

	Identify pneumonia	Relevance	mRMR	Relief	Gram-Schmidt	Lasso	Elastic net	sLDA	Quantifiable	Point-of-care	Affordable	RANK
Vital signs	**Temperature**	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	3
	HR	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	2
	RR	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	1
	Osat	N	Y	Y	N	N	N	Y	Y	Y	Y	5
Quantifiable 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	4
	Age	N	N	Y	N	N	N	Y	Y	Y	Y	6
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	

Table 6.1: A list of features (listed vertically) against feature selection techniques (horizontally) and three additional selection conditions of quantitative, potentially measurable in point-of-care setting and affordable. Yes (Y) / No (N) denote if a feature selection technique has selected a specific feature and if a selection condition is satisfied for a feature. The selected features for identification of pneumonia in descending order of their scores are: Respiratory Rate (RR), Heart Rate (HR), Temperature (T), Malnutrition (measured via the WHO Z-score - WHZ) and Oxygen saturation (Osat).

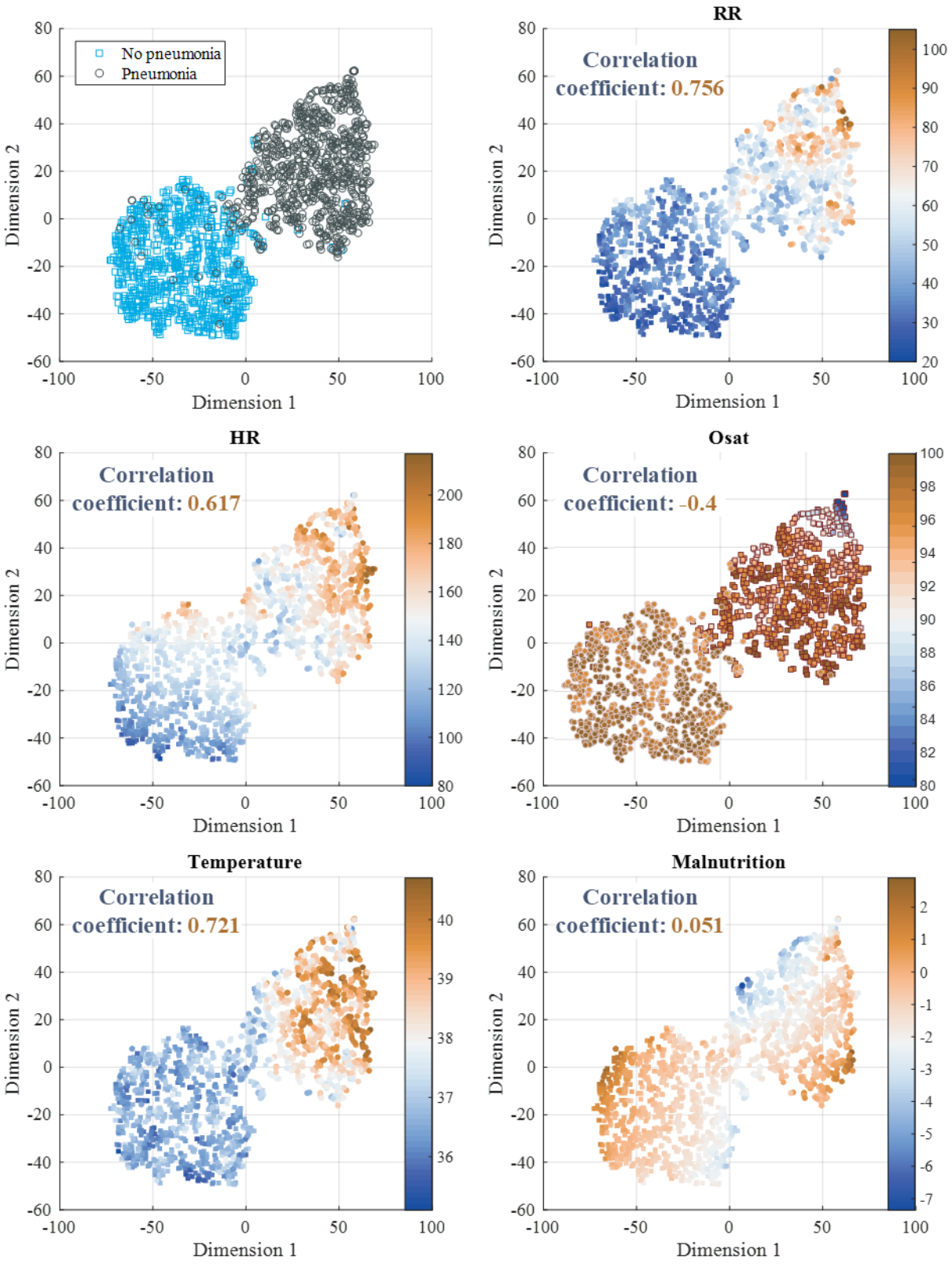


Figure 6.4: Distribution of different features across pneumonia (circle) cases and no pneumonia (square). T-SNE has been applied to all five clinical characteristics in order to reduce dimensionality to two dimensions; the top left graph illustrates affiliation of individual entries with the two classes. The colour scale in the remaining graphs represents the corresponding feature value per entry. Correlation coefficients for each feature and the outcome are also included in the plots.

Different loss function approaches (balanced accuracy, MCC) were explored for the three classifiers tested (LR, SVM, RF). The MCC was observed to be most favourable, optimizing performance in the test set as a result of fine-tuning of model parameters in the training/validation set. Specifically, for SVM, a Gaussian radial basis function kernel, with a kernel width, g, of 0.1 and a cost parameter of 1000 were found to deliver the best performance; for RF, 750 decision trees and searching over two variables at each tree node were the optimal hyperparameters.

The three classification techniques exhibited comparable performance, with somewhat more favourable sensitivity and specificity obtained with RF (Figure 6.5). LR was seen to sometimes outperform SVM but exhibited larger variance. The addition of the ghost vectors had a very limited effect on classification performance, with changes in all metrics limited to +/-1%. Taking a closer look at the RF algorithm, four features (RR, HR, Osat and T), with the addition of age, delivered maximal results: **98.2% (95%CI 97.9–98.8%) sensitivity; 97.6% (95% CI 97.1–98.0%) specificity; 99.7% (95% CI 99.3–99.8%) AUC; 95.9% (95% CI 95.3–96.5%) MCC**. From the cases that were falsely classified as controls, 33% were severe pneumonia. Reducing the number of features to three (RR, HR, Osat), worsened performance marginally: 98.2% (95% CI 97.8–98.6%) sensitivity; 97.1% (95% CI 96.8–97.5%) specificity; 99.6% (95% CI 99.5–99.6%) AUC; 95.2% (95% CI 94.9–96.1%) MCC. Somewhat surprisingly, the addition of malnutrition did not improve results, consistent with the mostly equal distribution of WHZ values across the two classes in the two-dimensional representation of the data (Figure 6.4).

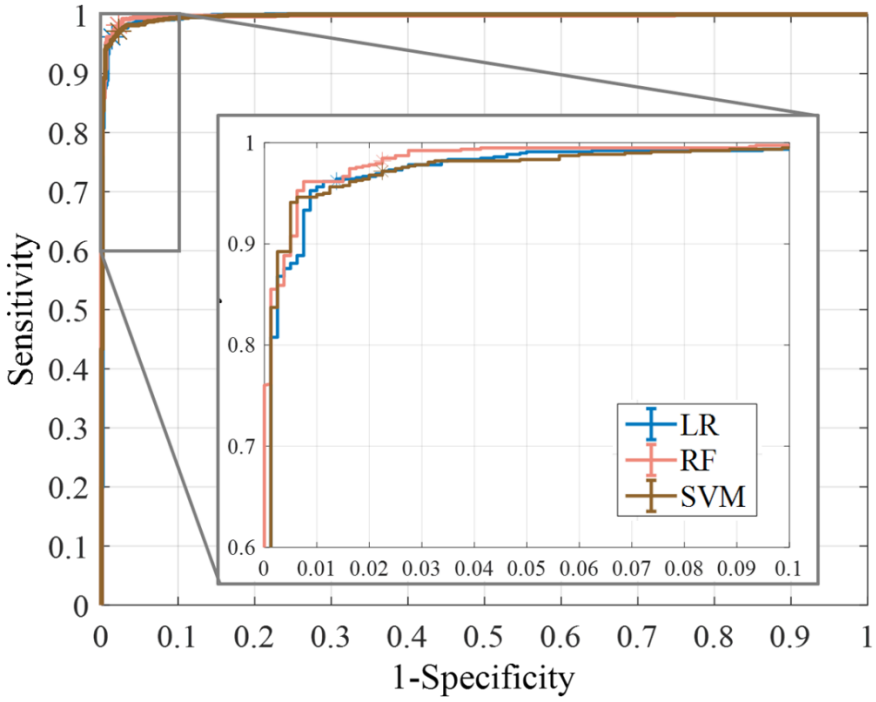


Figure 6.5: Receiver operating characteristic (ROC) curve for the three classifiers applied to the problem of identifying pneumonia; the location of the test results on the ROC curve as per the MCC optimization are denoted via an 'asterisk'.

6.3.2. Determining severity

The aim of this part of the analysis was to investigate whether it is possible to predict severity using quantifiable features rather than the observational features included in the 'gold standard' IMCI guidelines, such as 'chest wall indrawing'. Additionally, it was also investigated whether biomarkers could be fused with vital signs to improve severity prediction. For the purposes of an early warning algorithm, severity was divided into 'non-severe' and 'severe', with the latter combining both severe and very severe pneumonia cases.

In the first severity dataset, selected features were: RR, Osat, crackles, grunting, HR, with full list of feature selection outcomes found in Table 6.2. In the second severity dataset, selected features were: RR, grunting, crackles, CRP, HR, Osat, CD163, Hap, Lcn2, with full list of feature selection outcomes found in Table 6.3. In the first severity dataset, all five features had less than three missing values each, introducing five ghost vectors. In the second severity dataset, the clinical signs had no missing values and the proportion of missing values among the biomarkers were less than 8% (after exclusion of 19 cases that were missing seven out of nine features), introducing four ghost vectors.

First, the mixture of vital signs and lung sounds, listed for the first severity dataset, was used to classify severity (Figure 6.6). The following hyperparameters were found to lead to best performance: a Gaussian radial basis function kernel, with a kernel with g of 0.01 and a cost parameter of 10 000 for SVM; 750 decision trees and searching over two variables at each tree node for RF. MCC proved to be the most suitable optimization metric. Some differences were observed between the classifiers: RF and SVM performed comparably, whereas LR achieved better sensitivity at the cost of specificity (Figure 6.7). The addition of the ghost vectors had very limited effect on classification, with changes in all metrics limited to +0.8%. With five features (and age): LR delivered sensitivity of 84.6% (95% CI 83.6–85.2%); specificity of 68.3% (95% CI 67.5–69.3%); AUC of 83.7% (95% CI 83.2–83.8%); MCC of 53.6% (95%CI 52.1–55.2%); RF delivered sensitivity of **71.8% (95%CI 70.0%–76.6%); specificity of 81.8% (95% CI 73.3–83.5%); AUC of 83.5% (95% CI 82.5–83.8%); MCC of 52.9% (95% CI 49.6–55.2%)**.

The five clinical features listed above could be extracted via two measurements, keeping cost down, while also ensuring feasibility. Biomarkers could be used to improve performance but are associated with additional cost (both for the development of the assay, if it does not exist, as well as the cost of individual probes). Consequently, when introducing biomarkers to the feature set, keeping their number as low as possible (one or two) and identifying optimal combinations were priorities (Table 6.4). CRP and Hap added to the three vital signs and two lung sounds, were seen to deliver best results: sensitivity of 91.4% (95% CI 89.4%–92.3%); specificity of 83.2% (95% CI 82.1–84.2%); AUC of 94.2% (95% CI 93.7– 94.3%); MCC of 73.9% (95% CI 72.8–77.0%). However, with just one biomarker (CRP) and the same list of remaining features, performance was not much worse: **sensitivity of 88.5% (95% CI 87.5%–90.4%); specificity of 82.1% (95% CI 80.0–84.2%); AUC of 92.3% (95% CI 92.0–94.9%); MCC of 71.8% (95% CI 68.9–72.9%)**. The reported results were obtained with RF; with more than five features, LR's specificity fell below 60% and computational time with SVM was very long (17 h for the full validation routine, compared with 2 h with RF).

Taking a closer look at the classifier's predictions, most cases from both classes were classified with high level of certainty. To trace the roots of any uncertainty, the probabilistic outcomes were split into

10 bins and the distribution of individual features in each bin was investigated (Figure 6.8). From this, bins 1 and 2, and, 9 and 10 contained only one misclassified case and displayed a substantial difference between their CRP and RR values. Cases with more moderate CRP and RR values led to more uncertainty and consequently higher misclassification rates. Finally, the IMCI severity guidelines were used for comparison, delivering 79.3% sensitivity and 67.7% specificity (first severity dataset).

	Determine severity (1)	Relevance	mRMR	Relief	Gram-Schmidt	Lasso	Elastic net	sLDA		Quantifiable	Point-of-care	Affordable	RANK
Vital signs	**HR**	Y	N	Y	N	Y	Y	Y		Y	Y	Y	4
	RR	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y	1
	Osat	N	Y	N	N	N	N	Y		Y	Y	Y	
Observational clinical	**Pallor**	Y	Y	N	Y	Y	Y	N		N	Y	Y	
	Sternal indrawing	Y	Y	N	N	N	N	N		N	Y	Y	
	Head nodding	Y	Y	Y	Y	Y	Y	N		N	Y	Y	
	Days unwell	N	N	Y	Y	N	N	Y		N	Y	Y	
	Feeding obstruction	N	N	N	Y	Y	N	N		N	Y	Y	
Quantifiable clinical	**Age**	N	N	N	N	N	N	Y		Y	Y	Y	
	WHZ06	N	N	Y	N	N	N	Y		Y	Y	Y	
	Neutrophils	N	Y	N	N	Y	Y	N		Y	N	N	
Lung sounds	**Grunting**	Y	Y	Y	Y	Y	Y	N		Y	Y	Y	3
	Crackles	Y	Y	Y	Y	Y	Y	N		Y	Y	Y	2
	Bronchial breathing	Y	Y	N	N	N	N	N		Y	Y	Y	

Table 6.2: A list of features (listed vertically) against feature selection techniques (horizontally) and three additional selection conditions of quantitative, potentially measurable in point-of-care setting and affordable. Yes (Y) / No (N) denote if a feature selection technique has selected a specific feature and if a selection condition is satisfied for a feature. The selected features for severity determination in descending order of their scores based on the first severity dataset: RR, Grunting, Crackles, CRP, HR, Osat, CD163, Haptoglobin (haptog), Lipocalin-2 (Lcn2).

Determine severity (2)		Relevance	mRMR	Relief	Gram-Schmidt	Lasso	Elastic net	sLDA		Quantifiable	Point-of-care	Affordable	RANK
Vital signs	**HR**	Y	Y	Y	N	Y	Y	Y		Y	Y	Y	5
	RR	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y	1
	Osat	N	Y	Y	N	Y	Y	Y		Y	Y	Y	6
Observational clinical	**Yellow fever**	N	N	N	Y	Y	Y	N		N	Y	Y	
Quantifiable clinical	**WHZ06**	N	N	N	N	N	N	Y		Y	Y	Y	
Lung sounds	**Grunting**	Y	Y	Y	Y	Y	Y	N		Y	Y	Y	2
	Bronchial breathing	Y	Y	N	N	N	N	N		Y	Y	Y	
	Crackles	Y	Y	Y	Y	Y	Y	N		Y	Y	Y	3
Biomarkers	**haptoq**	Y	Y	N	Y	Y	Y	N		Y	Y	Y	8
	cd163	Y	Y	N	N	Y	Y	Y		Y	Y	Y	7
	crp	Y	Y	Y	Y	Y	Y	N		Y	Y	Y	4
	lpc	Y	N	Y	N	N	N	Y		Y	Y	Y	9

Table 6.3: A list of features (listed vertically) against feature selection techniques (horizontally) and three additional selection conditions of quantitative, potentially measurable in point-of-care setting and affordable. Yes (Y) / No (N) denote if a feature selection technique has selected a specific feature and if a selection condition is satisfied for a feature. The selected features for severity determination in descending order of their scores based on the second severity dataset: RR, Grunting, Crackles, CRP, HR, Osat, CD163, Haptoglobin (haptog), Lipocalin-2 (Lcn2).

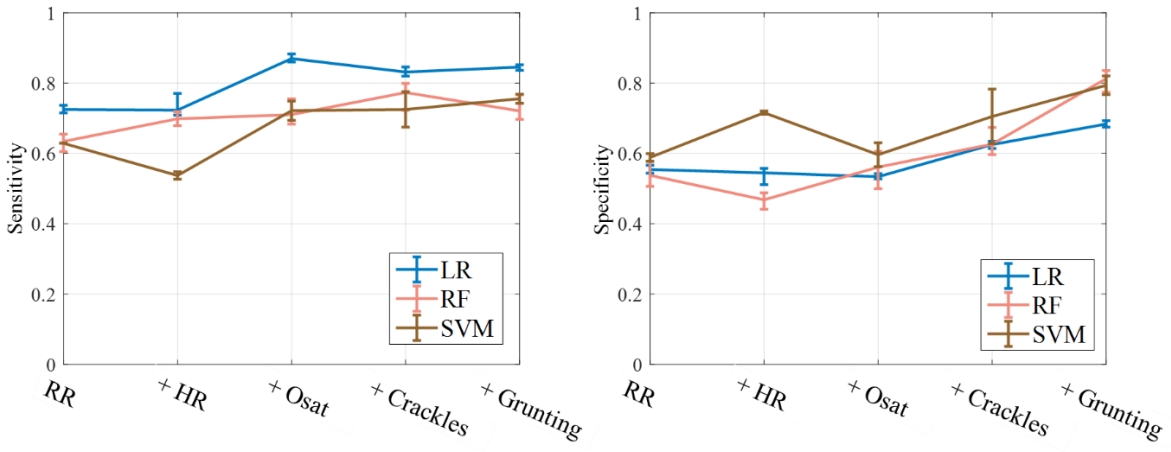


Figure 6.6: Classifiers' performance in determining severity, reported as sensitivity (left) and specificity (right) for an increasing number of features. Along the x-axis, features are listed in an additive manner, i.e. each x-entry represents classification with that specific feature and all features on the left of it. Additionally, age was added to the classifier throughout. Values represent the mean as well as the minimum and maximum values attained across the multiple iterations. Features that can be derived from the same measurement/signal have been listed adjacently and surrounded by a dashed box.

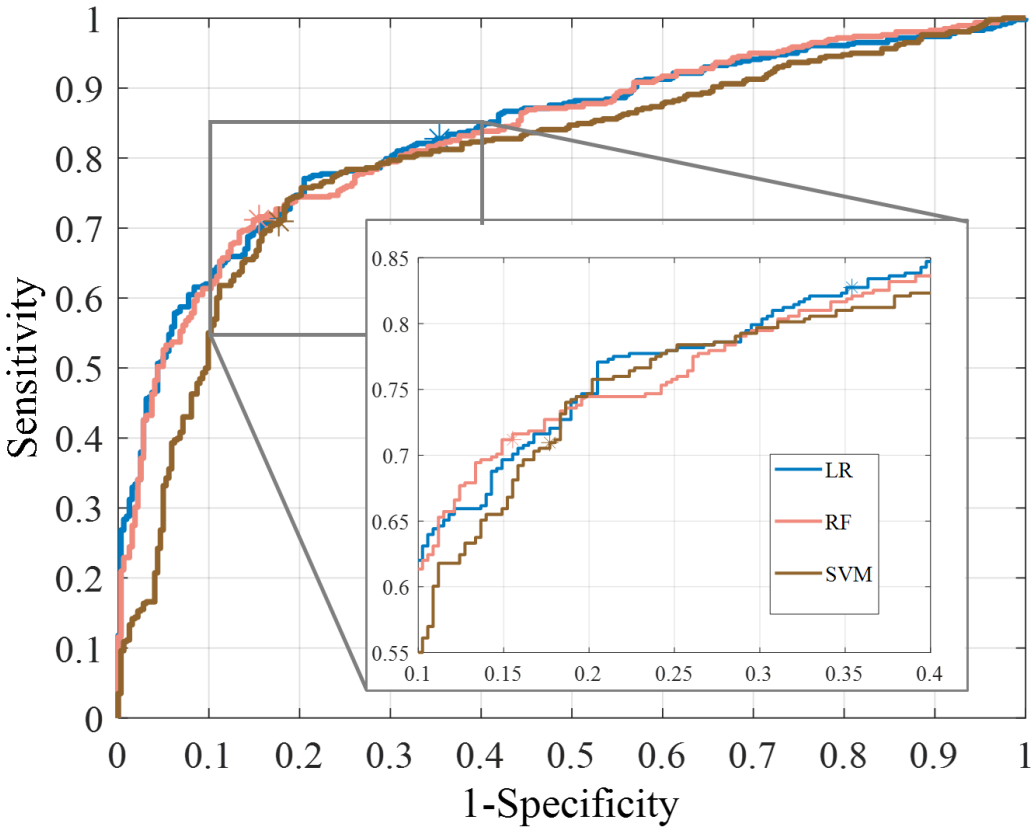


Figure 6.7: Receiver operating characteristic (ROC) curve for the three classifiers applied to the problem of determining severity based on 5 clinical characteristics – RR, HR, Osat, crackles, grunting; the location of the test results on the ROC curve as per the MCC optimization are denoted via an asterisk.

Number of features	Feature list	Sensitivity	Specificity
5a	RR, HR, Osat, crackles, grunting	0.85	0.68
5b	RR, HR, Osat, crackles, grunting	0.72	0.82
6b	RR, HR, Osat, crackles, grunting + CRP	0.89	0.82
6b	RR, HR, Osat, crackles, grunting + Lcn2	0.87	0.80
6b	RR, HR, Osat, crackles, grunting + Hap	0.88	0.79
6b	RR, HR, Osat, crackles, grunting + CD163	0.88	0.77
7b	RR, HR, Osat, crackles, grunting + CRP + Lcn2	0.91	0.78
7b	RR, HR, Osat, crackles, grunting + CRP + Hap	0.91	0.83
7b	RR, HR, Osat, crackles, grunting + Lcn2 + Hap	0.88	0.76

Table 6.4: Sensitivity and specificity of severity determination. First, biomarkers are added individually, followed by combinations of pairs of the top three performing biomarkers. With five features, LR outperformed RF's sensitivity; with more features, RF delivered both greater sensitivity and specificity.

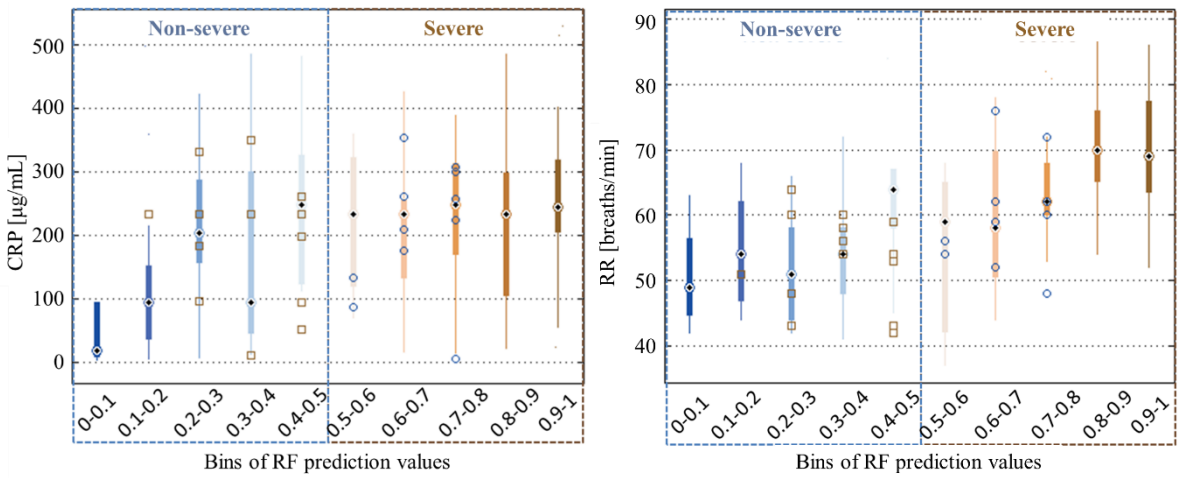


Figure 6.8: [LEFT] Distribution of CRP values across 10 probabilistic groups/bins. Along the x-axis, the range of RF probabilistic predictions was divided into 10 bins, where bin 1 contains cases assigned probabilities between [0,0.1], i.e. 90–100% certainty of non-severe pneumonia, and bin 10 contains cases assigned probabilities between [0.9,1], i.e. 90–100% certainty of severe pneumonia. The number of predicted cases in bins 1–10 were: 29, 21, 14, 23, 12, 7, 19, 23, 23, 28. In each bin, the feature distribution of correctly classified cases is visualized via a boxplot. In each box, the central dot represents the median, the edges are the 25th and the 75th percentiles, and the thin lines extend to the most extreme data points. Misclassified cases in each bin are plotted on top of the boxplot with squares denoting severe cases and circles denoting non-severe ones. [RIGHT] Distribution of RR values across ten probabilistic groups/bins. The number of predicted cases in bins 1-10 were: 29,21,14,23,12,7,19,23,23,28.

6.3.3. Determining aetiology

This part of the analysis explored whether a combination of vital signs and biomarkers could be used to predict aetiology (bacterial versus viral pneumonia), providing a potential alternative in settings where X-rays and blood culture laboratories are unavailable. The feature selection identified six features: Lcn2, Hap, RR, CRP, HR, CD163 (Table 6.5). Missing values were imputed using the median value of features, where imputation was under 10% in all concerned features, leading to the creation of four ghost vectors.

Optimization of hyperparameters was performed via MCC, selecting a g of 0.01 and a cost parameter of 1000; and 750 decision trees and searching over three variables at each tree node for RF. Substantial differences were observed in the performance of the three classifiers (Figure 6.9, left)—SVM was seen to underperform despite efforts to fine-tune the two relevant hyperparameters. RF was seen to deliver better sensitivity than LR for a comparable specificity. Similar to the severity part of the study, biomarkers were added to the feature set, aiming to minimize the number required to achieve satisfactory classification results (Table 6.6). The addition of Lcn2 to RR, HR and Osat was seen to deliver good performance: **sensitivity of 81.8% (95% CI 81.8–81.8%); specificity of 90.6% (95% CI 89.1–**

92.2%); AUC of 91.6% (95% CI 89.6–92.8%); MCC of 70.5% (95% CI 68.1–73.0%). Adding a second biomarker did not improve results. The addition of ghost vectors led to marginal changes in performance (+0.7%).

Similar to the analysis for severity, probabilistic predictions were split into bins and the distribution of individual features across these bins was studied (Figure 6.9, right). Cases with Lcn2 values below 200 ng ml21, were classified as viral with a high degree of certainty. However, three bacterial cases were seen to have Lcn2 values in that range and some viral cases also presented with elevated Lcn2. Additionally, the relationship between severity and aetiology of pneumonia is not straight-forward. The available dataset contained a mixture of cases spread across the severity and aetiology categories: from the bacterial cases, 10 were severe and 12 non-severe; from the viral cases, 12 were severe and 52 non-severe. Three of the severe cases (13.6%) were misclassified in terms of their aetiology. Based on the IMCI guidelines, all 86 cases should have been administered antibiotics. Hence, according to the IMCI guidelines, 100% of the viral cases would have been misclassified according to their aetiology.

	Determine aetiology	Relevance	mRMR	Relief	Gram-Schmidt	Lasso	Elastic net	sLDA		Quantifiable	Point-of-care	Affordable	RANK
Vital signs	**HR**	N	N	Y	Y	N	N	Y		Y	Y	Y	5
	RR	Y	Y	N	Y	Y	Y	N		Y	Y	Y	3
	Osat	N	N	N	N	N	N	Y		Y	Y	Y	
Observational clinical	**hepc**	Y	Y	Y	Y	N	Y	N		N	Y	Y	
Quantifiable clinical	**WCB**	Y	Y	Y	N	N	N	N		Y	N	N	
	WHZ	N	N	N	Y	N	Y	N		Y	Y	Y	
	Neutrophils	N	Y	N	Y	N	Y	Y		Y	N	N	
Lung sounds	**Grunting**	Y	N	Y	N	N	N	N		Y	Y	Y	
	Crackles	N	N	N	Y	N	N	N		Y	Y	Y	
Biomarkers	**lpc**	Y	Y	Y	Y	N	Y	N		Y	Y	Y	1
	haptoq	Y	Y	Y	Y	N	Y	N		Y	Y	Y	2
	crp	Y	Y	Y	N	N	N	N		Y	Y	Y	4
	cd163	Y	N	Y	N	N	N	Y		Y	Y	Y	6

Table 6.5: A list of features (listed vertically) against feature selection techniques (horizontally) and three additional selection conditions of quantitative, potentially measurable in point-of-care setting and affordable. Yes

(Y) / No (N) denote if a feature selection technique has selected a specific feature and if a selection condition is satisfied for a feature. The selected features for aetiology determination in descending order of their scores: Lipocalin-2 (Lcn2), Haptoglobin (haptog), RR, CRP, CD163.

Number of features	Feature list	Sensitivity	Specificity
3	RR, HR, Osat	0.5	0.86
4	RR, HR, Osat + CRP	0.64	0.88
4	RR, HR, Osat + Lcn2	0.82	0.91
4	RR, HR, Osat +Hap	0.64	0.86
4	RR, HR, Osat + CD163	0.46	0.91
5	RR, HR, Osat + Lcn2 + CRP	0.82	0.88
5	RR, HR, Osat + Lcn2 + Hap	0.82	0.91
5	RR, HR, Osat + Lcn2 + CD163	0.82	0.89

Table 6.6: Sensitivity and specificity of RF for prediction of aetiology. The best performing biomarker was identified first; combinations with a second biomarker were explored with no substantial change in performance.

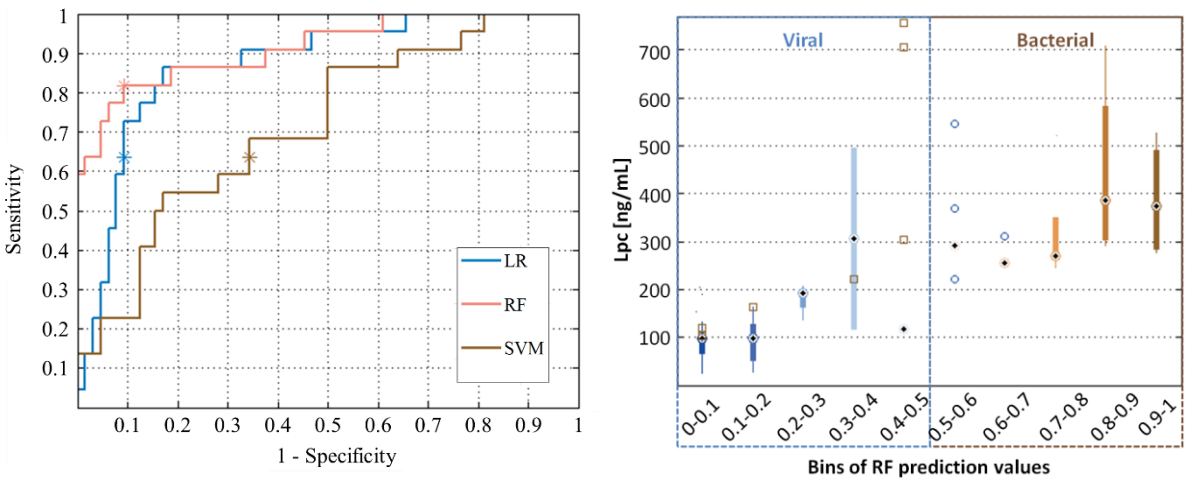


Figure 6.9: [LEFT] ROC curve for the three classifiers applied to the problem of aetiology determination; the location of the test results on the ROC curve as per the MCC optimization are denoted via an 'asterisk'. [RIGHT] Distribution of Lcn2 values across 10 probabilistic bins, presented as in Figure 6.8. The number of predicted cases in bins 1–10 were: 42, 14, 4, 3, 4, 4, 2, 5, 4, 7. Misclassified cases in each bin are plotted on top of the boxplot with squares denoting bacterial and circles denoting viral cases.

6.4. Discussion & conclusions

This chapter investigated the use of a suite of machine learning techniques to derive clinical insights regarding a holistic diagnostic assessment, including identification, severity and aetiology determination. The proposed approach can: identify pneumonia with 98.2% sensitivity and 97.6% specificity via five features (age, RR, HR, Osat, T); determine severity with 88.5% sensitivity and 82.1% specificity via seven features (age, RR, HR, Osat, crackles, grunting, CRP); and determine aetiology with 81.8% sensitivity and 90.6% specificity via four features (RR, HR, Osat, Lcn2). These findings are

very clinically relevant as they suggest that machine learning could be used to strengthen the capacity of health system in low-resource settings to deal with childhood pneumonia, despite sparsity of advanced equipment and highly trained clinicians.

The purpose of the analysis was not to simply automate the guidelines but rather to offer a more accurate and reproducible alternative. To achieve this, some of the features highlighted by the feature selection techniques as predictive ('cough', 'days unwell', 'head nodding') were excluded because they could not be measured unambiguously. Both 'cough' and 'days unwell' rely on a parent's careful observations. Similarly, features containing a huge degree of variability between providers of health (e.g. head nodding) were also excluded. However, lung sounds ('crackles', 'grunting', 'bronchial breathing') were kept in the feature set as recent technological advancements have indicated that the acquisition of lung sounds could be automated through appropriate signal processing. The latter is explored in detail in Chapter 2 and a novel framework for automated lung sound analysis is proposed in Chapter 8.

Some of the selected features were grouped together, e.g. RR, HR and Osat in Figure 6.6, regardless of the ranking order derived during feature selection. This was done in order to illustrate that some features could be obtained through the same measurement/signal, reducing complexity of clinical examination and cost. For example, the PPG signal obtained through a pulse oximeter could be used to derive RR, HR and Osat. Similarly, multiple lung sounds (crackles, grunting, bronchial breathing) can be obtained through a single stethoscope measurement. From the classifiers examined, RF was seen to outperform other methods (including LR which is conventionally used in clinical studies), indicating likely nonlinear interaction between the clinical signs measured. Identification of pneumonia was achieved with a high degree of confidence using four clinical features that could be derived from just two measurements— a PPG measurement (delivering RR, HR and Osat) and a temperature measurement.

The severity analysis elucidated a few key findings. First, with three vital signs and two lung sounds, it was possible to determine severity with high specificity (82.2%) but lower sensitivity (72.4%) using an RF algorithm. However, keeping the diagnostic application in mind, low sensitivity would mean severely ill children who should have been referred to hospital get missed. The use of LR had the opposite effect, favouring sensitivity at the expense of specificity (Figure 6.7). Consequently, triaging

of severe cases using the LR algorithm might be more efficient but would also lead to referral of a lot of non-severe cases. Second, the addition of biomarkers was seen to only improve sensitivity, with very limited changes to specificity. With or without biomarker information, fusion of features via machine learning was seen to outperform the IMCI guidelines for severity (as demonstrated in 6.3.2), where the latter uses observational and unquantifiable features.

Finally, the study also suggested an alternative source of aetiology information, which is typically obtained using X-rays and blood culture, by combining a couple of vital signs with a recently proposed biomarker (Lcn2). Moreover, only three out of the 22 severe cases were misclassified (13.6%) in terms of their aetiology. Misclassifications within the aetiology problem would be most detrimental in cases of severe bacterial/viral pneumonia as this could hinder the timely administration of appropriate treatment.

Multiple risk factors associated with pneumonia were discussed in Chapter 1, with malnutrition playing a substantial role in the clinical outcome. In the dataset analysed here, extreme malnutrition scores (from Figure 6.4, scores less than -4 and more than 1) were seen to be related to the presence of pneumonia, confirming the status of malnutrition as a high-risk factor. However, moderate malnutrition scores were seen to be equally distributed across both pneumonia and controls. Additionally, malnutrition showed limited significance to prediction of severity or aetiology. Nevertheless, it is expected that malnutrition would be more relevant as a predictor of survival but the dataset available did not contain information on such outcomes.

Finally, the dataset analysed in this study contained some limitations. Specifically, the control cases were generally quite healthy (apart from some odd cases of elevated HR and low Osat). In a realistic clinical setting, children will also present with various other diseases; therefore, a reliable evidence-based machine learning algorithm should ideally be trained to differentiate childhood pneumonia from other conditions that might appear similar (e.g. malaria or tuberculosis). The clinical study described in Chapter 3 was designed and conducted to collect such a dataset and Chapter 7 assesses the transferability of learnings attained in this Chapter to this richer and more realistic dataset.

This page is intentionally left blank.

7. Transferability of machine learning models for childhood pneumonia diagnosis

7.1. Introduction

Chapter 6 presented analysis of the Gambian dataset and proposed a combination of feature selection and classification techniques for multi-faceted diagnosis of childhood pneumonia, including identification, severity and aetiology determination. Whilst this approach generated very promising results, it leaves a number of questions unanswered. Namely:

- How well do such machine learning techniques, trained on one dataset, generalise to other datasets?
- How well do such machine learning techniques generalise across different populations?
- Are country-specific models required to account for potential population differences or should a global diagnostic model be built on the basis of multiple datasets?
- How well do such machine learning approaches perform in less idealistic datasets which include a lot of other symptomatically similar diseases?

This chapter addresses these questions by performing a series of analyses on the three datasets introduced in Chapters 3 & 4 and exploring the generalisability of machine learning approaches across datasets. Throughout, model *generalisability* is defined as the performance of a model developed on one dataset and applied to an independent second dataset; this is also known as external validation. First, statistical analysis is performed to quantify differences in the manifestations of pneumonia across datasets. Second, the performance of models developed on one dataset and applied to another is quantified. Finally, a global model, combining all three datasets is developed and performance is evaluated.

7.2. Pneumonia across datasets

Chapters 3 & 4 introduced three pneumonia datasets contained clinical characteristics – Gambian dataset, Indian (Bangalore) dataset and Indian (Mumbai) dataset. Since, data were collected as part of separate studies, in different settings, differences in symptomatic manifestation of pneumonia are possible. For example, in the Gambian dataset, 58.6% of pneumonia cases are severe, whereas in the Indian (Bangalore) dataset the proportion is much higher – 71%. Symptomatic differences are summarised in Table 7.1 and the differences across datasets are examined via the two-sample Kolmogorov-Smirnov test/Fisher test. For continuous clinical characteristics, such as respiratory rate, the two-sample Kolmogorov-Smirnov test returns a decision for the null hypothesis that samples from two different datasets are from the same continuous distribution. The null hypothesis was rejected for all continuous variables, indicating that samples come from different distributions. For categorical variables, the Fisher test returns a decision on the null hypothesis that there are no non-random associations between two categorical variables (i.e. a categorical clinical characteristic and the dataset source). The null hypothesis was rejected for all categorical variables in Table 7.1, implying that the dataset source bears some associations with the clinical characteristics.

Additionally, using the Chi-Squared and Mann-Whitney tests, the association of different clinical characteristics and the presence of pneumonia, in each dataset, was quantified. The following characteristics were found to be statistically significant in the Gambian dataset: **respiratory rate, heart rate, oxygen saturation, temperature, malnutrition, as well as white blood cells, neutrophils, lymphocytes and monocytes**. In the Indian (Bangalore) dataset, the statistically significant characteristics were: **respiratory rate, oxygen saturation, temperature, crackles and wheezing, history of fever, malnutrition, able to eat, lethargic and chest indrawing**. Some of the vital signs, such as heart rate and respiratory rate, appear to be less significant than in the Gambia dataset. This is likely to be attributed to the different disease constituents in the two datasets – the control cases in the Gambian dataset are mostly healthy, whereas the control cases in the Indian (Bangalore) dataset have other respiratory diseases. Consequently, generic vital signs, like heart rate and respiratory rate, are likely to be informative when differentiating between sick and healthy children but less informative for subtle differences between different respiratory diseases. Finally, for the Indian (Mumbai) dataset, the

statistically significant features were: **age, respiratory rate, heart rate, oxygen saturation, temperature, crackles and wheezing, history of fever, malnutrition, able to eat, lethargic and chest indrawing, cough, difficulty breathing**.

Taking a closer look, Figure 7.1 contains boxplots for the main clinical characteristics across the datasets. Differences in the distribution of respiratory rate, heart rate, temperature and oxygen saturation are apparent even on visual inspection. Some of these differences might be driven by differences in the age distribution (Indian (Mumbai) population is older than the other two) since vital signs such as respiratory rate and heart rate are age-dependent. The Indian (Mumbai) dataset was seen to have lower average and range than the other two datasets. Taking a closer look, the last graph (bottom-right) of Figure 7.1 includes the age stratified boxplots of respiratory rate values in the India (Mumbai) dataset with corresponding ranges of normality marked via blue boxes. From this, it is the youngest children (<13 months) that appear to have respiratory rate values within the clinical bounds of normality, despite having radiologically confirmed pneumonia. Age-stratification of respiratory and heart rates was performed to explore the effects of age on distribution of the two vital signs; median and standard deviation values are listed in Table 7.2. Even accounting for age, differences in respiratory rate between the Indian (Mumbai) dataset and the other two are substantial. Similarly, age-stratified heart rate values in the Gambian dataset appear to be substantially different from those in the other two datasets. For temperature, differences might be driven by different methods of measurement – temperature was measured orally in the Gambian and Indian (Bangalore) datasets, and under the armpit in the Indian (Mumbai) dataset. Additionally, due to the different setups in the studies, it is likely that children in some populations were given antibiotics/ symptomatic medication for temperature control, before being taken to hospital. The incidence of malnutrition was found to be substantially different across datasets (24.2% in the Gambian dataset, 62.3% in the Indian (Bangalore) dataset, and 16.5% in the Indian (Mumbai) dataset). As discussed in Chapter 1, it has been previously documented that manifestations of pneumonia can be different in children who are heavily malnourished.

To provide a more visual overview of the overall symptomatic differences across datasets, Figure 7.2 includes a radar chart. This plot illustrates the high level of symptomatic variability in pneumonia cases

from different studies and suggests that generalisability of diagnostic models from one dataset to another might be challenging. Taking this investigation a step further, t-SNE, as introduced in Chapter 5, was used to perform dimensionality reduction on the combined three datasets and explore symptomatic differences in the reduced two-dimensional space. Namely, Figure 7.3 and Figure 7.4 visualise the two-dimensional structure of the data, as well as the distribution of individual clinical characteristics across that structure. From this, five distinct clusters of pneumonia cases emerge, as numbered in Figure 7.3; some of these contain cases from a specific dataset (i.e. clusters 2-5 are primarily formed of Gambia data cases) and one of the clusters, 1, contains a mixture of cases from all three datasets. The means and standard deviation of each feature across the five clusters were calculated (Table 7.3); the two-sample Kolmogorov-Smirnov test and Fisher test were used to test examine distribution differences across the clusters (Table 7.4). Cluster 1 was found to differ from the rest mainly in HR, Osat, age, T, malnutrition and crackles, whereas differences between the remaining four clusters were primarily driven by differences in lung sounds and malnutrition. These symptomatic differences are consistent with the reported differences in severity.

Table 7.1: Clinical characteristics in pneumonia cases across datasets; the statistical significance of differences is quantified through Two-sample Kolmogorov-Smirnov test/Fisher test.

Pneumonia	**1. Gambian dataset**			**2. Indian data (Bangalore)**			**3. Indian data (Mumbai)**			**Two-sample Kolmogorov-Smirnov test / Fisher test**		
	N = 777			*N = 314*			*N = 109*			*p-values*		
	Mean	Median	Std	Mean	Median	Std	Mean	Median	Std	1-2	1-3	2-3
Gender [%, female]		-			35.7			39.4				
Age [months]	17.2	15	11.6	16.8	10	20	22.2	17	15.7	**<0.001**	**<0.001**	**<0.001**
Respiratory rate [breaths/min]	61.4	60	12	59.5	58.5	11.5	47.9	46	8.9	**<0.001**	**<0.001**	**<0.001**
Heart rate [beats/min]	157.3	157	18.9	138	137.5	25.2	137.4	137.4	17.9	**<0.001**	**<0.001**	**<0.001**
Oxygen saturation [%]	98.4	96	32.7	93.5	95.5	5.4	96.8	97.7	3.3	**<0.001**	**<0.001**	**<0.001**
Temperature [°C]	38.7	38.6	3.3	37.9	37.9	1	36.8	36.4	1	**<0.001**	**<0.001**	**<0.001**
History of fever (%)		-			94.3			96.1		-	-	-
Duration [days]	-	-	-	-	-	-	7.2	5.0	4.9	-	-	-
Malnutrition (%)		24.2			62.3			16.5		**<0.001**	**<0.001**	**<0.001**
Z-score	-1.4	-1.4	1.3	-	-	-	-	-	-	-	-	-
Unable to eat/breastfeed (%)		18.1			29.9			22.9		**<0.001**	**<0.001**	**<0.001**
Lethargic (%)		17.1			31.5			18.4		**<0.001**	**0.004**	**<0.001**
Chest indrawing (%)		3.3			76.8			82.5		**<0.001**	**<0.001**	**<0.001**
Cough (%)		86.4			96.8			93.2		**<0.001**	**<0.001**	**<0.001**
Duration [days]	-	-	-	-	-	-	7.5	5.0	5.5	-	-	-
Difficulty breathing (%)		-			88.5			82.5		-	-	**<0.001**
Dehydration (%)		4.0			-			69.2		-	**0.001**	-
Stridor (%)		0.1			-			-		-	-	-
Grunting (%)		21.0			-			-		-	-	-
Crackles (%)		39.1			76.1			72.5		**<0.001**	**<0.001**	**<0.001**
Wheeze (%)		0.4			37.9			10.1		**<0.001**	**<0.001**	**<0.001**
Bronchial breathing (%)		10.8			0.0			-		-	-	-
Head nodding (%)		4.1			-			-		-	-	-
Cyanosis (%)		0.3			-			-		-	-	-
Pallor (%)		19.9			-			-		-	-	-
Blood culture (%)		12.5			-			-		-	-	-
Haemoglobin [g/dL]	8.6	9.3	3.4		-		9.4	9.5	1.9	-	**<0.001**	-
White blood cell count [billion/litre]	15.6	13.8	11.3		-		17	13	21	-	**<0.001**	-
Neutrophils [billion/litre]	47.1	52.1	25.9		-		-	-	-	-	-	-
Lymphocytes [billion/litre]	30.1	31	19.9		-		-	-	-	-	-	-
Monocytes [billion/litre]	5.8	6.4	3.3		-		-	-	-	-	-	-
Platelets [billion/litre]	-	-	-				3.4	3.3	1.8	-	-	-
Chest X-rays												
Normal (%)		-			18.2			34.9		-	-	-
Hyperinflation (%)		-			7.0			0.9		-	-	-
Pleural fluid (%)		-			42.0			4.6		-	-	-
Lobar changes (%)		-			19.1			11.9		-	-	-
Minor patchy changes (%)		-			10.2			2.8		-	-	-
Major patchy changes (%)		-			6.7			2.8		-	-	-
Positive X-rays [%]		23.6			-			-		-	-	-
Other symptoms												
Diarrhoea (%)		-			-			14.6		-	-	-
Duration [days]		-			-		5.9	3	9.2	-	-	-
Blood in stool (%)		-			-			1.9		-	-	-
Rash (%)		-			-			13.6		-	-	-
Runny nose (%)		-			-			58.3		-	-	-
Red eys (%)		-			-			3.9		-	-	-
Ear pain (%)		-			-			5.8		-	-	-
Ear discharge (%)		-			-			5.8		-	-	-

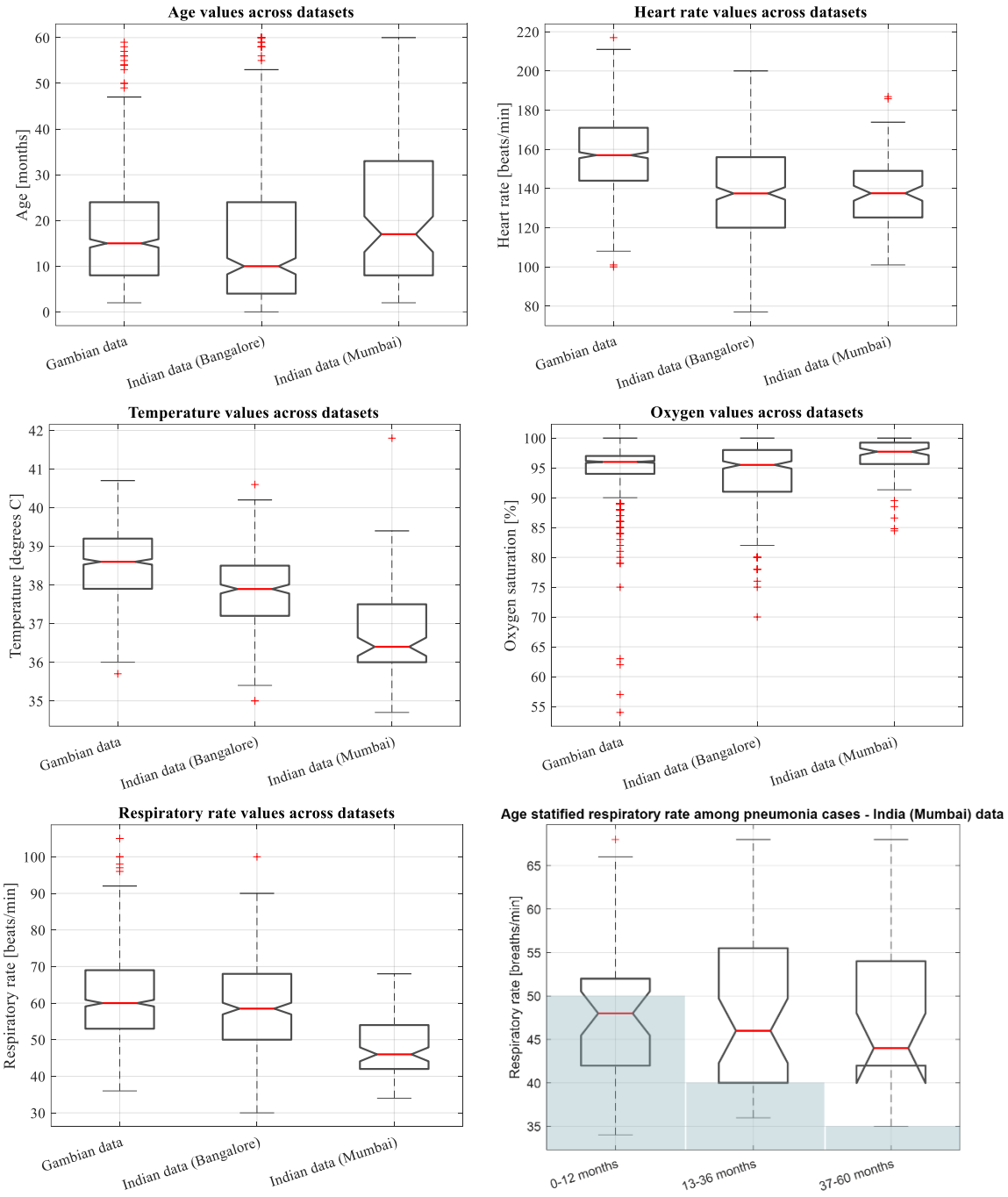


Figure 7.1: Boxplots of age, heart rate and temperature and oxygen saturation, respiratory rate in pneumonia cases across datasets. The bottom right plot includes age-stratified boxplot of respiratory rate values in the India (Mumbai) dataset; the blue blocks mark the clinically defined bounds of normality of respiratory rate for each age group.

Age	Normal respiratory rate	Median respiratory rate			Standard deviation in respiratory rate		
[months]	*[breaths/min]*	*[breaths/min]*			*[breaths/min]*		
	Clinical standard	*Gambian*	*Indian (Bangalore)*	*Indian (Mumbai)*	*Gambian*	*Indian (Bangalore)*	*Indian (Mumbai)*
0-12	30-50	64	62	48	11	11	8
13-36	24-40	58	54	46	12	10	10
37-60	22-34	53	50	44	11	11	9

Age	Normal heart rate	Median heart rate			Standard deviation in heart rate		
[months]	*[beats/min]*	*[beats/min]*			*[beats/min]*		
	Clinical standard	*Gambian*	*Indian (Bangalore)*	*Indian (Mumbai)*	*Gambian*	*Indian (Bangalore)*	*Indian (Mumbai)*
0-12	100-160	165	144	140	18	24	12
13-36	90-150	154	130	136	19	25	20
37-60	80-140	145	120	130	15	26	21

Table 7.2: Age-stratified median respiratory and heart rates across the three datasets.

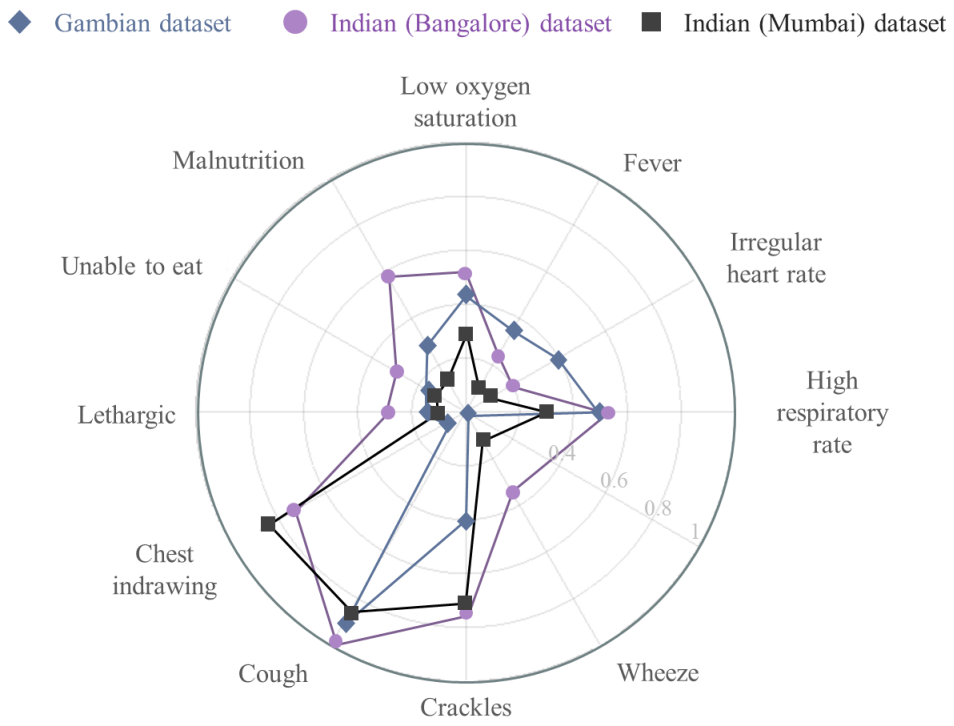


Figure 7.2: Radar chart of incidence of difference clinical characteristics in pneumonia cases from each one of the three datasets. Both irregular heart rate and high respiratory rate were derived using the age stratification introduced in Table 7.2.

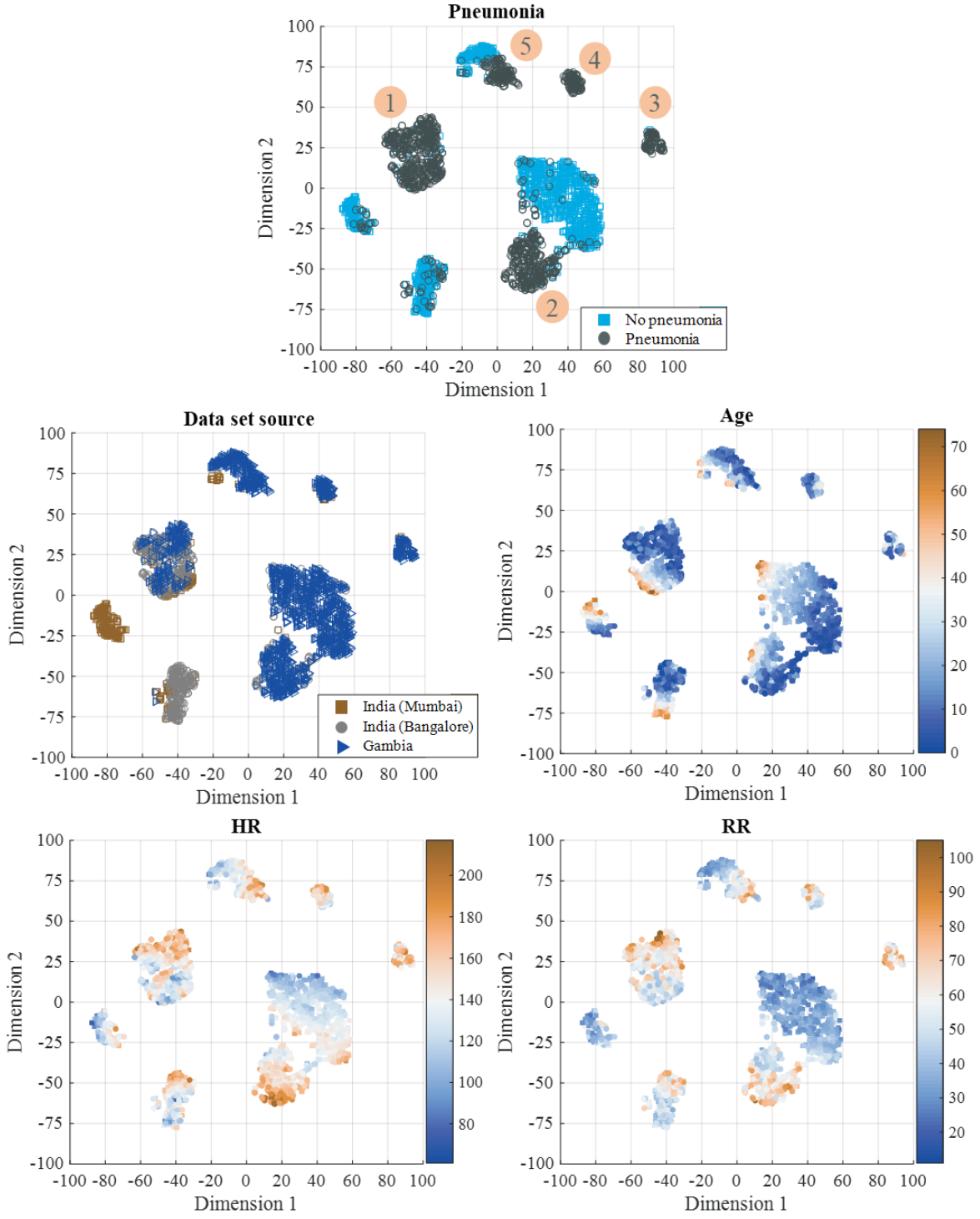


Figure 7.3: Dimensionality reduction applied to the combination of the three datasets using t-SNE. The two-dimensional structure was obtained by applying t-SNE to a feature set of nine characteristics (age, HR, RR, T, Osat, Malnutrition, Crackles, Wheeze, Bronchial breathing). In the different plots, data points are colour-coded according to their corresponding attributes (e.g. class or feature values). Five main pneumonia clusters emerge visually, as numbered in the first plot.

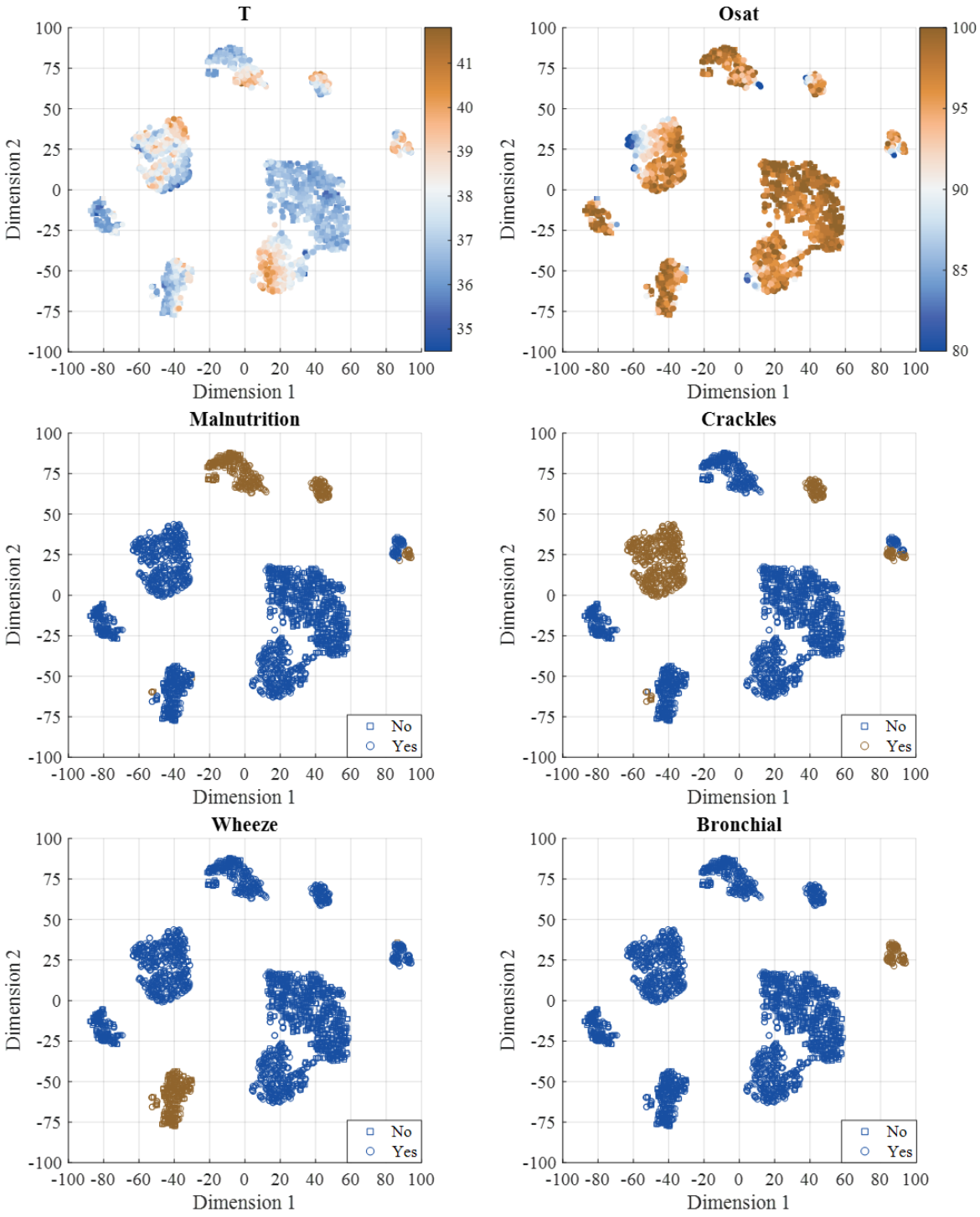


Figure 7.4: (continued from Figure 7.3) Dimensionality reduction applied to the combination of the three datasets using t-SNE. The two-dimensional structure was obtained by applying t-SNE to a feature set of nine characteristics (age, HR, RR, T, Osat, Malnutrition, Crackles, Wheeze, Bronchial breathing). In the different plots, data points are colour-coded according to their corresponding feature values.

MEAN	RR	HR	Osat	T	Malnutrition (%)	Crackles (%)	Wheeze (%)	Bronchial (%)	Age
Pneu 1	59.2	140.6	93.8	37.8	0	1.28E-01	0	0	16.3
Pneu 2	58.3	154.5	96.1	38.6	0	0	0	0	17.4
Pneu 3	56.6	153.0	95.4	38.7	8.34E-04	1.25E-03	4.17E-04	5.84E-03	15.7
Pneu 4	52.5	148.9	95.9	37.8	6.67E-03	6.67E-03	0	0	14.6
Pneu 5	57.3	154.9	96.6	38.6	2.5	0	0	0	17.1

STD	RR	HR	Osat	T	Malnutrition	Crackles	Wheeze	Bronchial	Age
Pneu 1	10.9	23.6	5.1	1.0	0	1.28E-01	0	0	15.7
Pneu 2	10.9	18.5	3.2	1.0	0	0	0	0	12.4
Pneu 3	7.7	17.9	1.5	1.2	8.34E-04	1.25E-03	4.17E-04	5.84E-03	7.4
Pneu 4	7.0	15.8	3.7	1.0	6.67E-03	6.67E-03	0	0	7.7
Pneu 5	10.4	16.9	1.9	1.0	2.50E-02	0	0	0	10.9

Table 7.3: Mean and standard deviation of values (for continuous characteristics) or of percentage incidence (for binary characteristics) across the five pneumonia clusters (rows) identified in Figure 7.3.

	Pneu 1	**Pneu 2**	**Pneu 3**	**Pneu 4**	**Pneu 5**
Pneu 1	-				
Pneu 2	HR, Osat, T -, Age C, -, -	-			
Pneu 3	HR, Osat, T M, - C, W, B	-, -, - M, - C, -, B	-		
Pneu 4	HR, -, - M, Age -, -, -	-, -, - M, - C, -, _	-, -, - M, - C, -, B	-	
Pneu 5	HR, Osat, T M, Age C, -, -	-, -, - M, - C, -, -	-, -, - M, - C, -, B	-, -, T M, - C, -, -	-

Table 7.4: Statistically significant differences in features across the five pneumonia clusters identified in Figure 7.3. The two-sample Kolmogorov-Smirnov test and Fisher test with 5% significance level are used for continuous and categorical data respectively. The initials of a feature are included in the table if differences for the corresponding two clusters were shown to be significant, where C = crackles, W = wheeze, B = bronchial breathing, M = malnutrition and the rest of the feature abbreviations are as defined before. No statistically significant differences in RR between the different clusters were found.

7.3. Methods – diagnostic algorithms for multiple populations

7.3.1. IMCI benchmark

The IMCI guidelines, as defined in Chapter 1, were applied to each dataset to establish a baseline representing the current gold standard for community-based diagnosis.

7.3.2. Model generalisability

To investigate model generalisability across different pneumonia datasets, this chapter employs two approaches, illustrated in Figure 7.5.

Approach 1 applies the analytical pipeline introduced in Chapter 6 to also develop models on the basis of the India (Mumbai) and India (Bangalore) datasets; it then performs external validation by applying these models to each of the two datasets that were not used for their development (Figure 7.5, [LEFT]). Hence, for each dataset, we generate a family of models $\mathbf{\Psi}_{d,i,j,k}$, where d = (1,2,3) is the dataset index, i = (1, …, 20) is the iteration in training, j = (1,…, 4) is the fold of cross-validation and k = (1, …,K) is the number of features used from a dataset with K available clinical characteristics. The internal performance of the models (i.e. when applied to a test set from the same population as the training set) is quantified to asses their ability to classify cases generated by the same population. The external performance of the models (i.e. when applied to a new population) is quantified to assess generalisability across settings. For the latter to be accomplished, both averaging and model selection were explored as potential techniques to derive a single prediction in a new dataset from the family of models $\mathbf{\Psi}_{d,i,j,k}$. The following method was found to deliver best results (highest performance in test set):

(1) For every $\mathbf{\Psi}_{d,i,j,k}$, an RF model trained on a fraction of the dataset $\mathbf{X}_{d,k}$, predictions on the new dataset can be derived: $\mathbf{\Psi}_{d,i,j,k}(X_{b,k}): RD \rightarrow [0,1]$, where $X_{b,k} = \{x_n \in \mathrm{R}^k, n = 1 \ldots N\}$, N is the total number of samples in the new dataset and b is the dataset index such that b = (1,2,3) and $b \neq d$. It should be noted here that the new dataset is normalised using mean and standard deviation values from the training fraction of $\mathbf{X}_{d,k}$.

(2) Next, balanced accuracy was used to eliminate models that underperformed in the validation fraction of $\mathbf{X}_{d,k}$. Namely, for a given number of features, *k*, models which delivered balanced accuracy less than 70% were excluded, shrinking the model set from $\boldsymbol{\Psi}_{d,i,j,k}$ to $\boldsymbol{\Psi}_{d,i\prime,j\prime,k}$.

(3) Predictions derived from the remaining set of models were averaged, before applying a threshold to convert the final outcome into a binary value. This was used to derive performance metrics for the new dataset.

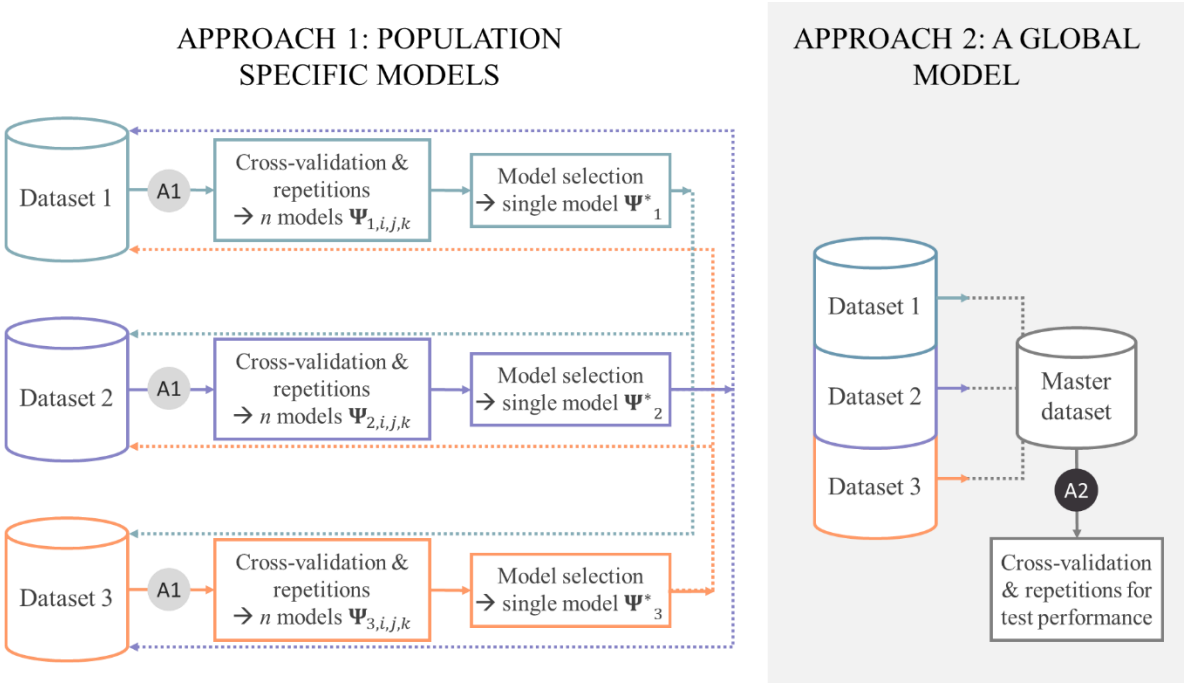


Figure 7.5: Two approaches to designing diagnostic algorithms for childhood pneumonia. [LEFT] Approach 1 – developing population specific models and performing external validation across datasets. Here A1 denotes the analytical approach proposed in Chapter 4; with 20 iterations and four-fold cross-validation, this generates 80 models, $\boldsymbol{\Psi}_{d,i,j,k}$, where $d = (1,2,3)$ is the dataset index, $i = (1, \ldots, 20)$ is the iteration in training, $j = (1,\ldots, 4)$ is the fold of cross-validation and $k = (1, \ldots,K)$ is the number of features used from a dataset with K available clinical characteristics. [RIGHT] Approach 2 – developing a global model on the combined data across datasets. Here A2 denotes the proposed approach 2.

Approach 2 explores the development of a '*global model*' for childhood pneumonia diagnosis (Figure 7.5, [RIGHT]). As highlighted in 7.2, there are significant differences in the symptomatic manifestations of pneumonia across populations. Hence, for a model to be generalisable, it should capture this variability. Consequently, the three datasets were combined into one dataset (referred to as 'master dataset') and diagnostic models, following the RF analytical pipeline introduced in Chapter 6, were developed and validated on this large dataset (referred to as 'master dataset' in this chapter). The RF

importance scores were used to rank features in decreasing order of significance and add them sequentially, generating a separate RF model for each feature configuration.

7.4. Results: IMCI benchmark

The IMCI guidelines applied to each of the three datasets, delivered the following results:

- Gambian data: 77.8% sensitivity and 61.4% specificity
- Indian (Bangalore) data: 82.2% sensitivity and 43.1% specificity
- Indian (Mumbai) data: 50.0% sensitivity and 83.3% specificity

7.5. Results: Approach 1 – population specific models

The following section reports on results obtained via approach 1, i.e. applying models developed on one dataset to other datasets. To enable such comparisons, the feature sets of all datasets were limited to the overlap; this list of features was also consistent with the main symptoms used in clinical practice to identify the disease, as introduced in Chapter 1. These features included: RR, HR, Osat, T, cough, ability to eat/breastfeed (feeding), sternal indrawing, malnutrition, crackles, wheezes, bronchial breathing.

Table 7.5 summarises the performance of population specific models, developed on each of the three datasets and evaluated both internally (on a test set from the same population) and externally (on the other two datasets). For each training dataset, the best performing combination of features is reported. From this, a few key takeaways were derived:

- The Gambian dataset delivered very high internal performance – **98.2% sensitivity and 97.6% specificity** - (highest amongst the three datasets) but very poor generalisability to the other two datasets (especially the Indian (Bangalore) dataset). In particular, specificity in external validation was very low – 50.7% for the Indian (Mumbai) data and 3.7% for the Indian (Bangalore) data.
- The Indian (Bangalore) dataset delivered good performance during internal validation – **89.1% sensitivity and 77.5% specificity –** with a combination of crackles, wheeze and temperature. Generalisability to the other two datasets was limited, with high sensitivity (85.9% for the

Gambian data and 88.5% for the Indian (Mumbai) data) but poor specificity (46.3% for the Gambian data and 52.1% for the Indian (Mumbai) data).

- The Indian (Mumbai) dataset, collected during the field study described in Chapter 2, delivered good performance during internal validation - **77.8% sensitivity and 85.9% specificity –** with a combination of crackles, wheezes, T, Osat, HR, RR. Generalisability was reasonably good for the Gambian data (**60.8% sensitivity and 98.2% specificity**) and very good for the Indian (Bangalore) data (**85.6% sensitivity and 71.6% specificity**). In comparison, classification of pneumonia in the Indian (Mumbai) dataset with vital signs alone delivered 85.9% sensitivity and 69.7% specificity.

Datasets		**Model features***	**Accuracy**		**Sensitivity**		**Specificity**	
Train	Test		**Mean**	(95% CI)	**Mean**	(95% CI)	**Mean**	(95% CI)
G	G	RR, HR, Osat, T	**0.979**	(0.974-0.983)	**0.982**	(0.979-0.988)	**0.976**	(0.971-0.98)
G	I (M)	RR, HR, Osat, T	**0.7**	(0.678-0.722)	**0.892**	(0.876-0.908)	**0.507**	(0.5-0.508)
G	I (B)	RR, HR, Osat, T	**0.505**	(0.483-0.526)	**0.972**	(0.969-0.975)	**0.037**	(0.033-0.042)
I (M)	I (M)	Crackles, Wheezes, T, Osat, HR, RR	**0.818**	(0.796-0.841)	**0.778**	(0.715-0.84)	**0.859**	(0.812-0.906)
I (M)	G	Crackles, Wheezes, T, Osat, HR, RR	**0.764**	(0.735-0.794)	**0.608**	(0.525-0.691)	**0.982**	(0.968-0.997)
I (M)	I (B)	Crackles, Wheezes, T, Osat, HR, RR	**0.754**	(0.729-0.778)	**0.856**	(0.826-0.887)	**0.716**	(0.647-0.784)
I (B)	I (B)	Crackles, Wheeze, T	**0.833**	(0.823-0.843)	**0.891**	(0.87-0.911)	**0.775**	(0.799-0.831)
I (B)	G	Crackles, Wheeze, T	**0.627**	(0.578-0.676)	**0.859**	(0.796-0.922)	**0.463**	(0.823-0.843)
I (B)	I (M)	Crackles, Wheeze, T	**0.625**	(0.56-0.69)	**0.885**	(0.847-0.924)	**0.521**	(0.394-0.648)

Table 7.5: RF model performance for classification of pneumonia versus non-pneumonia for a range of training and testing datasets combinations. The best performing feature combination for each training dataset is reported, where age is included in all features sets. The following abbreviations for the datasets are used: G = Gambian dataset; I(M) = Indian (Mumbai) dataset; I(B) = Indian (Bangalore) dataset. * All models included age as a feature.

7.6. Results: Approach 2 – a global model

This section reports on the performance of approach 2, i.e. developing a global model for pneumonia diagnosis, using the combination of the three datasets (the 'master dataset'). Optmising MCC across five-fold internal cross validation, RF with 750 trees, searching over two variables at each tree node, was trained to classify pneumonia versus all other cases, both healthy controls and other diseases, in the master dataset. Results are presented in Figure 7.6 in terms of sensitivity, specificity and ROC curves for an increasing number of features. Optimal performance was achieved with all 7 features plus age (RR, T, HR, Osat, wheeze, crackles and other lung sounds): sensitivity of **96.3% (95% CI 95.7 – 97.0%) and specificity of 92.9% (95% CI 91.9 – 93.8%)**.

Figure 7.7 takes a closer look at what disease category the misclassified cases belong to. The majority of misclassification stemmed from the India (Bangalore) and India (Mumbia) datasets, with the Gambian part of the data only generating 4 false negatives and no false positive values. A large number of the misclassifications stemmed from the wheeze disease cases in the Indian (Bangalore) dataset being mistakenly labelled as pneumonia. This, combined with other respiratory disease cases (e.g. upper respiratory tract infection, bronchitis) accounted for 69% of the false positives in the classification. From the false negative values – 18 originated from the Indian (Mumbai) dataset (17.8% of all pneumonia cases in that dataset) and 30 (9.5% of all pneumonia cases in that dataset) from the Indian (Bangalore) dataset. Hence, this approach resulted in improvement in classification performance, when compared to the population specific models presented in 7.4.

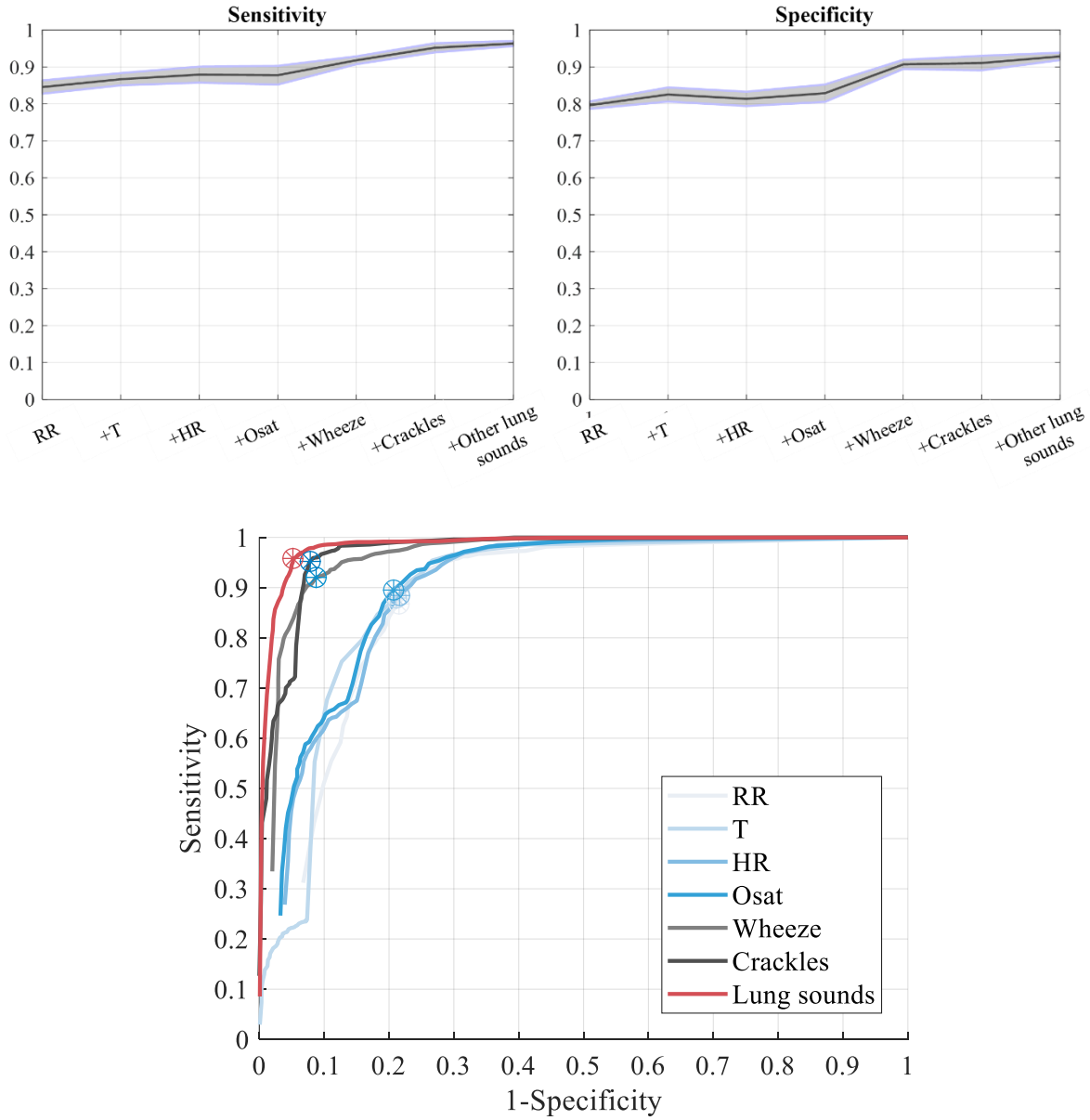


Figure 7.6: Performance of RF classifier differentiating between pneumonia and all other cases in the test set of the global dataset. Results are presented in terms of sensitivity & specificity [TOP] and ROC curves [BOTTOM] for an increasing number of features. Note that in both graphs each consecutive feature has been added to the preceding ones and all feature combinations include age.

Where does the model go wrong in pneumonia versus the rest?												
	1 features*				5 features*				7 features*			
	India (Mumbai)	India (Bangalore)	Gambia	Total	India (Mumbai)	India (Bangalore)	Gambia	Total	India (Mumbai)	India (Bangalore)	Gambia	Total
N=	*277*	*546*	*1567*	*2390*	*277*	*546*	*1567*	*2390*	*277*	*546*	*1567*	*2390*
False Negatives												
Total	**40**	**51**	**60**	**151**	**28**	**68**	**3**	**99**	**18**	**30**	**4**	**52**
Pneumonia	40	51	60	151	28	68	3	99	18	30	4	52
False Positives												
Total	**47**	**183**	**30**	**260**	**58**	**67**	**0**	**125**	**22**	**49**	**0**	**71**
Other respiratory	33	173	0	206	28	57	0	85	11	38	0	49
Wheeze disease	4	173	-	177	0	57	-	57	0	38	-	38
Hyper-active airway disease	8	-	-	8	3	-	-	3	2	-	-	2
Upper respiratory tract infection	15	-	-	15	21	-	-	21	6	-	-	6
Whooping cough	1	-	-	1	1	-	-	1	1	-	-	1
Bronchitis	5	-	-	5	3	-	-	3	2	-	-	2
Non-respiratory	11	10	0	21	15	10	0	25	10	11	0	21
Acute febrile illness	3	-	-	3	5	-	-	5	1	-	-	1
Diarrhoea	7	-	-	7	8	-	-	8	4	-	-	4
Pleural effusion	-	-	-	0	-	-	-	0	3	-	-	3
Epilepsy	-	-	-	0	-	-	-	0	-	-	-	0
Acute malnutrition	-	-	-	0	-	-	-	0	-	-	-	0
Urinary tract infection	-	-	-	0	1	-	-	1	-	-	-	0
Others	1	10	-	11	1	10	-	11	2	11	-	13
Healthy	3	0	30	33	15	-	0	15	1	-	-	1

**Across 4 x 597 cases classified*

Figure 7.7: A breakdown of false negative and false positives values generated by RF classification of pneumonia in the master dataset; results include variable number of features – 1, 5 and 7 (presented in the following order: RR, T, HR, Osat, wheeze, crackles, other lung sounds).

7.7. Discussion & Conclusions

In this chapter, the generalisability of machine learning models for pneumonia diagnosis across study populations is explored. The statistical analysis of the three pneumonia datasets performed in 7.2 highlighted some statistically significant differences in clinical characteristics across datasets. These differences stem from a combination of factors. One, different pneumonia severity levels in the different populations, mean that the symptomatic manifestations of the disease would vary, e.g. clinical characteristics such as oxygen saturation are lower in more severe cases. Two, natural physiological differences across populations in different geographies; and three, differences in study protocols and equipment. Specifically, the Indian (Mumbai) dataset was seen to contain lower respiratory rate values, compared to the other two datasets, with the bottom age bracket (0-12 months) containing a substantial number of pneumonia cases with no tachypnoea. This is consistent with other studies that have previously reported on the lower sensitivity of diagnostic value of tachypnoea, depending on age, malnutrition and duration of illness (Palafox et al. 2000). Duration of illness is likely to be an important contributor towards differences between the three datasets, where community-identified cases in the India (Mumbai) dataset are often in an earlier stage of the disease when tachypnoea is less common than in later stages of the disease (cases at hospital).

Population specific models

The machine learning approach for pneumonia classification proposed in Chapter 6 was found to successfully model all three datasets individually, delivering high outcomes during internal validation: 97.9% (95% CI 97.4-98.3%) accuracy for the Gambian data; 83.3% (95% CI 82.3-84.3%) accuracy for the Indian (Bangalore) data; and 81.8% (95% CI 79.6-84.1%) accuracy for the Indian (Mumbai) data. However, population models were found to differ in the feature set delivering optimal performance as well as in their generalisability when applied to other datasets.

Namely, models developed on the Gambian data were seen to select conventional vital signs as optimal features (RR, HR, T, Osat); since the non-pneumonia class in the dataset is primarily healthy, a combination of conventional vital signs is mostly a sufficient differentiator. In contrast, the Indian (Bangalore) dataset favoured the lung sound features (crackles and wheezes), in addition to T, as the non-pneumonia class in the dataset included mostly other respiratory diseases, necessitating the use of more specific features such as lung sounds to make the differentiation. The Indian (Mumbai) data contained a range of respiratory, non-respiratory and healthy cases in its non-pneumonia class and hence feature selection favoured a mixture of conventional vitals signs and lung sounds (RR, T, HR, Osat, wheezes, crackles and other vital signs).

Generalisability of models developed on the Gambian data to the other two datasets was quite poor, especially in terms of specificity. This is to be expected since the Gambian data contain almost no instances of other diseases; hence, when presented with other respiratory or non-respiratory diseases in the two Indian datasets, it classifies these as more similar to pneumonia than the healthy controls. This was particularly true with the Indian (Bangalore) data which contained primarily respiratory cases, similar to pneumonia on the basis on conventional vital signs, resulting in 3.7% specificity and 97.2% sensitivity. At the other extreme of the disease spectrum, models developed on the Indian (Bangalore) data generalised poorly to datasets containing healthy controls (Gambian data) or a wider range of non-respiratory diseases (Indian (Mumbai) data). Models developed on the Indian (Mumbai) data were seen to generalise best because the data were more encompassing, capturing a wide range of other diseases as well as healthy controls. The t-SNE visualisation performed in Figure 7.3 provided some initial

indication of this generalisability since the Indian (Mumbai) data points were found to be scattered amongst clusters formed by the other two datasets. More specifically, many of the pneumonia data points from the two Indian datasets were found to aggregate in the same cluster within the two-dimensional structure (marked as cluster 1 in Figure 7.3). Consistent with this visual observation, generalisability of the Indian (Mumbai) models to the Indian (Bangalore) data was better than to the Gambian data: 85.6% sensitivity and 71.6% specificity achieved on the Indian (Bangalore) data versus 60.8% sensitivity and 98.2% specificity on the Gambian data. This was also supported by the statistical analysis conducted on the symptomatic manifestations of pneumonia across dataset (Figure 7.1 and Table 7.2), where differences in the distribution of clinical features across pneumonia cases from the Gambia and the India (Mumbai) data were seen to be greater than between the two Indian datasets. Specifically, the Gambian pneumonia cases were seen to have significantly higher HR, even accounting for age stratification (Table 7.4).

Global model

The alternative approach to generalisability, designing a diagnostic model on the combined master dataset, was seen to deliver superior results: **96.3% (95% CI 95.7 – 97.0%) and specificity of 92.9% (95% CI 91.9 – 93.8%)** when validated on an independent test set, across four-fold cross validation and 20 repetitions. Similar to models developed on the Indian (Mumbai) data, the feature selection favoured the use of a mixture of features, including conventional vital signs as well as lung sounds. A deeper dive into the source of any remaining misclassification (Table 7.5) revealed that the majority of misclassification is driven by symptomatically similar respiratory diseases in the two Indian datasets. These findings support two main conclusions. First, a large global dataset, capturing heterogeneity within pneumonia as well as other diseases, is required to design models generalisable across populations and the approach and results presented here give confidence in the high performance of such approach. Two, more specific symptoms are required to resolve any remaining misclassification between pneumonia and other symptomatically similar respiratory conditions. Therefore, acoustic lung signals were explored in more detail in an effort to identify respiratory-specific clinical features that generalise across datasets and Chapter 8 presents findings in this area.

This page is intentionally left blank.

8. Automated lung sound analysis for childhood pneumonia

8.1. Introduction

Chapter 2 reviewed existing approaches to lung sound analysis, as a potential support tool in the diagnosis of pneumonia. At the end of Chapter 2, four outstanding challenges to the automated analysis of lung sounds were identified and discussed in detail. This chapter proposes a new approach to lung sound analysis for pneumonia diagnosis with two main components: (1) GMM-HMM / DNN-HMM models for state prediction in the lung signals; and (2) an RF model combining features derived from the states to reach a final classification of pneumonia. The methodology employed is presented, followed by results on two big datasets – the Peru dataset and the Indian (Mumbai) dataset collected during the field study. The goal of the proposed analysis is two-fold. First, this work aims to improve diagnostic performance achieved through the use of conventional vital signs, as presented in Chapter 7. As a reminder, classification of pneumonia versus other conditions with vital signs alone in the Indian (Mumbai) dataset delivered **sensitivity of 85.9% and specificity of 69.7%**. Second, this work aims to develop a method for automated lung sound analysis that addresses the shortage of clinical expertise in low-resource settings and hence matches or outperforms the ability of individual experts to interpret a signal. Hence, annotation performed by expert clinicians, as described in Chapter 4, is used as a benchmark.

8.2. Methodology

8.2.1. Annotator performance

Annotator performance on the Indian (Mumbai) dataset is analysed, where inter-annotator agreement is quantified via two statistical metrics – Fleiss κ and SA – as introduced in Chapter 5.

8.2.2. Automated lung sound analysis

This chapter proposes two different methods for automated lung sound analysis, suitable for datasets with different gold standard availability. The different components of the two methods are visualised in Figure 8.1 and the overall purpose of each model is described here:

(1) **Method 1**: For datasets that contain gold standard annotation of individual lung sounds (e.g. the Peru dataset introduced in Chapter 3 had a clinical expert mark the time stamps of different lung pathologies), it is possible to train HMM models for the different lung sounds. Given the imperfection of the gold standard, we are interested in a breakdown of the lung signal contents of each patient (i.e. what percentage of their lung signals were predicted to have crackles etc.), rather than precise overlap between the signal and the predicted lung sound states. The signal contents predictions can be turned into patient features and their contribution towards overall diagnosis can be quantified. The Peru dataset is used to develop and evaluate the performance of this approach.

(2) **Method 2**: For datasets that do not contain precise labels of lung sounds (e.g. the Indian study dataset introduced in Chapter 4), it is possible to train HMM models for the different disease categories that patients fall within (i.e. signals associated with pneumonia, signals associated with non-respiratory disease etc.). The predicted lung signal content for each patient can be converted into features reflecting likelihood of disease association. Combined with other symptoms such as vital signs, these features inform an overall patient diagnostic model. The Indian study dataset is used to develop and evaluate the performance of this approach.

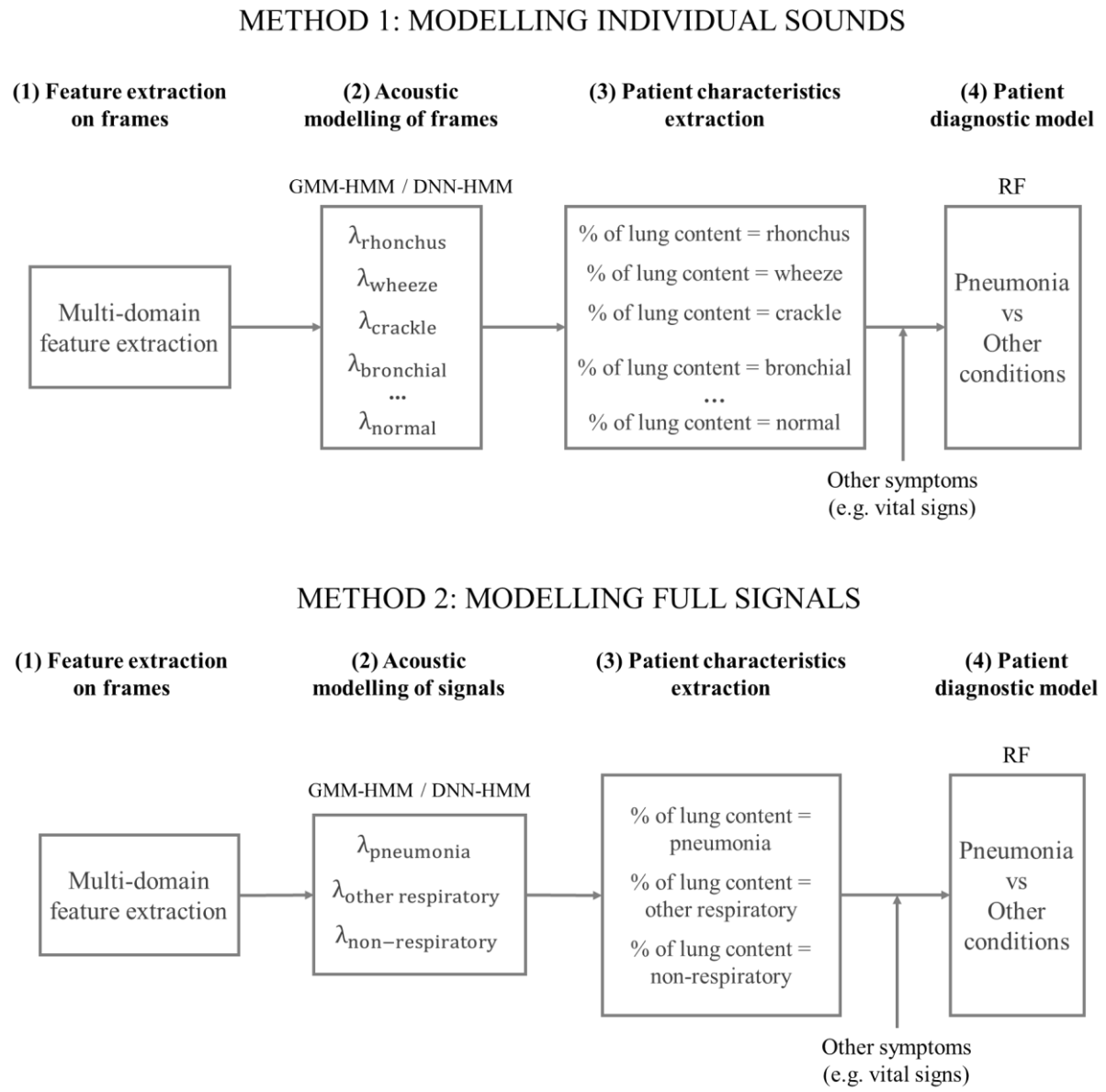


Figure 8.1: Analysis workflow. Two different methods are proposed – method 1 modelling individual lung sounds (if gold standard annotation available) and method 2 modelling full signals to derive acoustic biomarkers directly linked with disease.

8.2.3. Feature extraction

Each signal is broken into frames – 60 milliseconds in length with 30 milliseconds of overlap between subsequent frames. Multi-domain features are extracted on a per-frame basis and each frame resulted in a feature array containing the following categories of features:

$$\mathbf{F}_{i,j,d} = \left\{ \begin{matrix} Kurtosis \\ Skewness \\ Entropy \\ Fractal\ analysis \end{matrix} \middle| \begin{matrix} AR \\ features \end{matrix} \middle| \begin{matrix} EMD \\ features \end{matrix} \middle| \begin{matrix} MFCC \\ features \end{matrix} \middle| \begin{matrix} Wavelet \\ features \end{matrix} \right\}$$

where $\mathbf{F}_{i,j,d}$ is an array of 242 features derived from the d^{th} frame of the i^{th} signal of the j^{th} case/patient. The theoretical framework behind each feature category was presented in chapter 5 and specific details of the implementation are presented next.

Fractal analysis features

Lacunarity was computed on each frame for a variable box size, in the range $\{1, 2, 4, \ldots, 2^p\}$, where p is such that 2^p is smaller or equal to the length of the signal. For the 60ms frames, this equated to 13 boxes per frame. In addition to the lacunarity values, fractal dimensions and fractal abundance were also recorded for each frame.

MFCC features

For each frame, twelve cepstral coefficients were derived via a filter bank with 20 triangular filters, equally spaced between 1Hz and 3000Hz, consistent with optimal performance achieved in previous studies (Mayorga 2010, 2011, Rabiner 1993) In addition to the MFCCs, filter bank energies (FBEs) for each frame are recorded before the discrete cosine transform is applied to compute the cepstral coefficients.

Autoregressive features

For each frame, a 6^{th} order AR model is fitted, deriving intercept (b), model coefficients $\{a\}_{i=1}^{6}$, number of non-zeros coefficients (n), as well as the associated eigenmodes and noise covariance. The choice of mode order (i.e. 6^{th}) was based on the literature review from Chapter 2, where a 6^{th} order model was

seen to be most common for analysis of lung sounds and sufficient to capture the most common lung sounds such as crackles and wheeze.

Empirical mode decomposition

The number of IMFs was calculated for each frame, as well as the dominant frequency of each of the first six IMFs.

Wavelet features

WPT was used to perform multi-resolution decomposition of each frame up to the 5th level, i.e. 63 coefficients per frame. Two different mother wavelets were used - Duabechies-4 and Daubechies-40. The choice of wavelets was based on a morphological resemblance between the Duabechies-4 wavelet and the typical shape of crackles, and that between the Daubechies-40 wavelet and wheeze/rhonchus/bronchial breathing, as illustrated in Figure 8.2. A series of alternative wavelets were explored and these two were found to maximise the differences in distribution between the different lung sound classes in the online lung database.

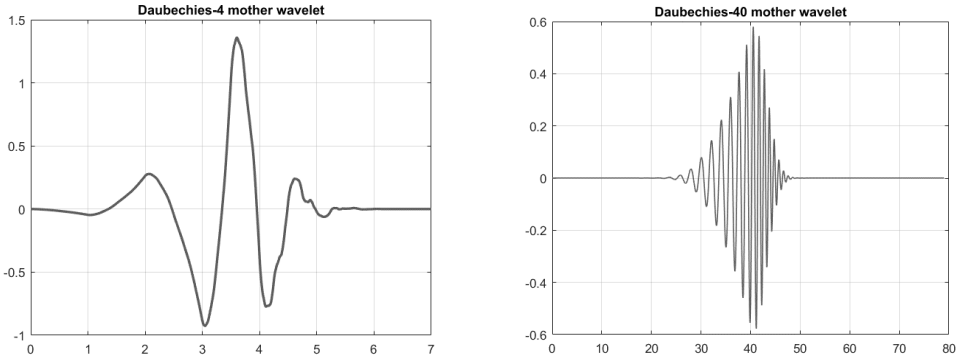


Figure 8.2: Daubechies-4 mother wavelet (left) and Daubechies-40 mother wavelet (right).

Feature trimming

In order to inspect the large number of features (242) generated through this multi-domain approach and reduce their number, class Probability Distribution Functions (PDFs), i.e. PDFs fitted to the data in each class of the classification problem, were used. Kolmogorov-Smirnov test was used to identify features that come from the same distribution and remove them from the feature set.

8.2.4. Acoustic modelling

In Chapter 5, we introduced the theoretical framework of HMMs, as well as the potential of combining HMM with powerful techniques such as GMM and DNN to better model individual states in a signal, described by a large number of features. Building upon this, models 1 & 2 use GMM-HMM and DNN-HMM to model: (1) different lung sounds as separate states in model 1; (2) signals associated with different disease categories in model 2. The annotator's labels of individual lung sounds in the Peru data set and the disease associations of signals in the Indian (Mumbai) dataset were used to train the HMM models.

- **GMM-HMM**: The classic GMM-HMM algorithm, as introduced in Chapter 5 is used, where each state, parameterised by the features described above, is modelled via a separate GMM and the different GMMs are used as input to an HMM sequence of states, where parameter optimisation is performed via the Baum-Welch method. To classify an unknown test sequence, maximum likelihood scores across the different class models are compared.
- **DNN-HMM**: Similar to the speech recognition algorithms described in Chapter 5, a DNN-HMM model is constructed as illustrated in Figure 8.1. The GMM layer is replaced by a DNN layer and the models are trained using: discriminative pretraining for model 1 (where the lung sound labels are used as 8 different states) and unsupervised pre-training for model 2 (where the overall association of a signal with a disease class is targeted and the number of constituent states can be varied); error back propagation for fine-tuning on the entire network.

With both HMM techniques, care is taken to ensure that frames that stem from the same patient cannot feature in both the train and test set, and hence minimise the information leakage identified during the literature review in Chapter 2.

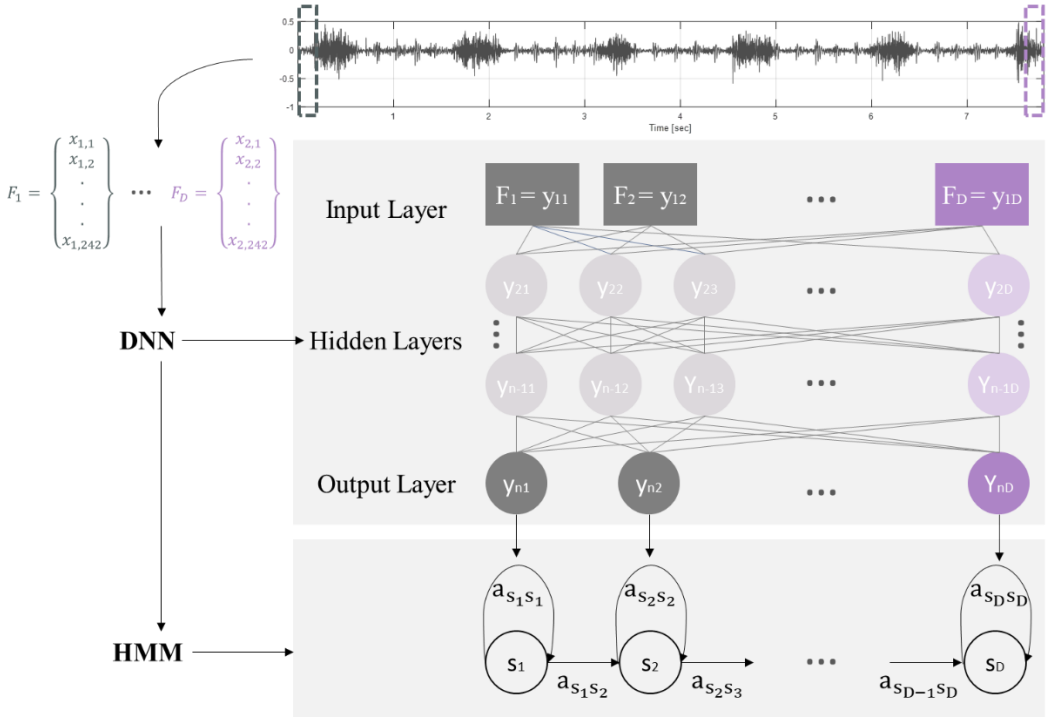


Figure 8.3: DNN-HMM model layout

8.2.5. States path characteristics

Both the GMM-HMM and DNN-HMM methods described above assign a state to each frame in the test set. The state could either be trained on a lung sound basis (crackles vs wheeze vs normal etc.) or on a signal basis (pneumonia signal vs non-respiratory disease signal vs other respiratory disease signal). The frames states are converted to first derive features describing the contents of each signal, i.e. what fraction of the signal contains crackles or what fraction of the signal is labelled as a pneumonia signal. For the Indian dataset, which has multiple signals per patient, these signal features are further aggregated into patient features – what fractions of the patient's lung signals were associated with pneumonia. The resultant patient feature set, for each of the two models, contains the following features:

- **Model 1**: 7 features from each of the HMM models corresponding to the different lung sounds (rhonchus, wheeze, crackles, bronchial breathing, cough, other adventitious sounds, conductive sounds), where each feature quantifies the fraction of patient signal predicted to be affiliated with the given lung sound.
- **Model 2**: 3 features from each of the HMM models corresponding to the different disease classes (pneumonia, non-respiratory, other respiratory), where each feature quantifies the fraction of all patient signals predicted to be affiliated with a given disease class.

8.2.6. RF classification

Leave-one-out (LOO) cross validation was used to develop and optimise an RF model on the HMM-derived features, assess performance on the test set and quantify variability across 100 repetitions. The process employed was equivalent to the one used in Chapter 6 to combine a range of vital signs into pneumonia diagnosis. For the Peru dataset, performance for classification of asthma and bronchitis was also assessed.

8.3. Results: lung sound annotation outcomes

Results on baseline

One of the annotators (Sion 3) only labelled the baseline and hence will only feature in reporting on those 700 recordings. Table 8.1 reports on the constituent lung sounds of the baseline recordings, as labelled by the 7 annotators. Substantial differences were recorded, with incidence of sounds like crackles ranging from 0.4% to 13.4% across annotators. Examining annotation on a patient level made these differences even starker – in Table 8.2 the attribution of lung sounds to patient was visualised, where a patient is assumed to be labelled as having a specific lung sound if at least one of their recordings is labelled to have that lung sound. In this, the incidence of crackles ranges between 1.5% and 40.8%, and that of wheeze between 1.5% and 38.2%.

The Fleiss' kappa coefficient was found to be 'slight' (<0.2) for all pathological sounds and 'fair' (~0.25) for normal sounds (Figure 8.4), indicating poor agreement between the annotators. Similarly, the specific agreement coefficient was found to be very low for pathological sounds (<0.25) and not much higher for normal sounds (Figure 8.5).

Baseline signal annotation

Sound category	Annotator						
	JR 1	*JR 2*	*JR 3*	*JR 4*	*Sion 1*	*Sion 2*	*Sion 3*
Crackles	0.4	10.6	6.1	11.9	13.4	5.1	9.7
Wheeze	0.7	5.3	7.0	1.3	10.6	2.9	7.0
Bronchial breathing	8.1	3.6	2.4	6.1	8.1	16.6	5.3
Stridor	0.9	11.0	2.9	1.1	0.0	3.1	0.0
Normal	58.7	46.0	47.0	72.7	60.1	55.4	64.3
Can’t interpret	31.3	27.3	34.7	8.0	12.7	20.7	16.9

Table 8.1: Summary of incidence of lung sounds in the baseline batch of recordings, reported by each annotator.

Baseline patients annotation

Sound category	Annotator							
	JR 1	*JR 2*	*JR 3*	*JR 4*	*Sion 1*	*Sion 2*	*Sion 3*	*Sion 4*
Crackles	1.5	19.5	9.6	18.4	40.8	10.2	22.4	6.1
Wheeze	1.5	8.5	10.5	2.3	38.2	5.8	14.6	6.1
Bronchial breathing	12.8	6.7	4.7	9.0	17.5	31.5	9.6	7.3
Stridor	1.5	16.3	4.1	1.5	0.0	6.1	0.0	2.9
Normal	74.6	66.5	64.7	81.3	83.1	75.2	82.2	81.3
Can’t interpret	51.6	46.6	49.6	14.6	32.1	36.7	31.8	2.3

Table 8.2: Summary of incidence of lung sounds in the baseline batch of participants, reported by each annotator.

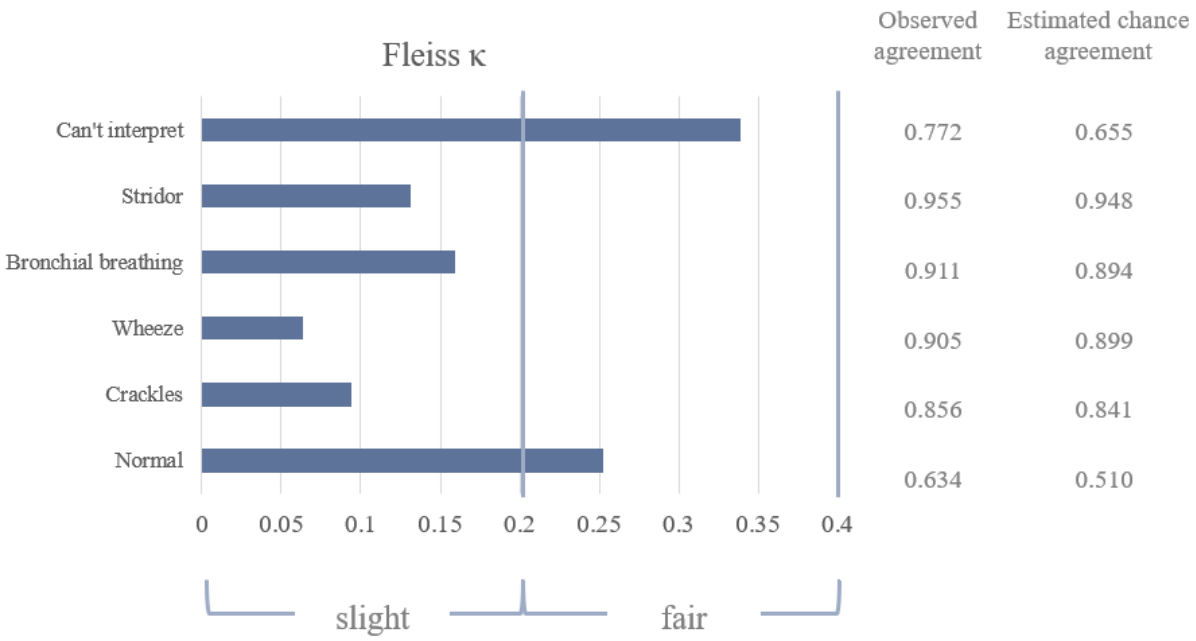


Figure 8.4: Annotator agreement (Fleiss’ kappa) between 7 observers on labelling of lung sounds from the 700 baseline recordings.

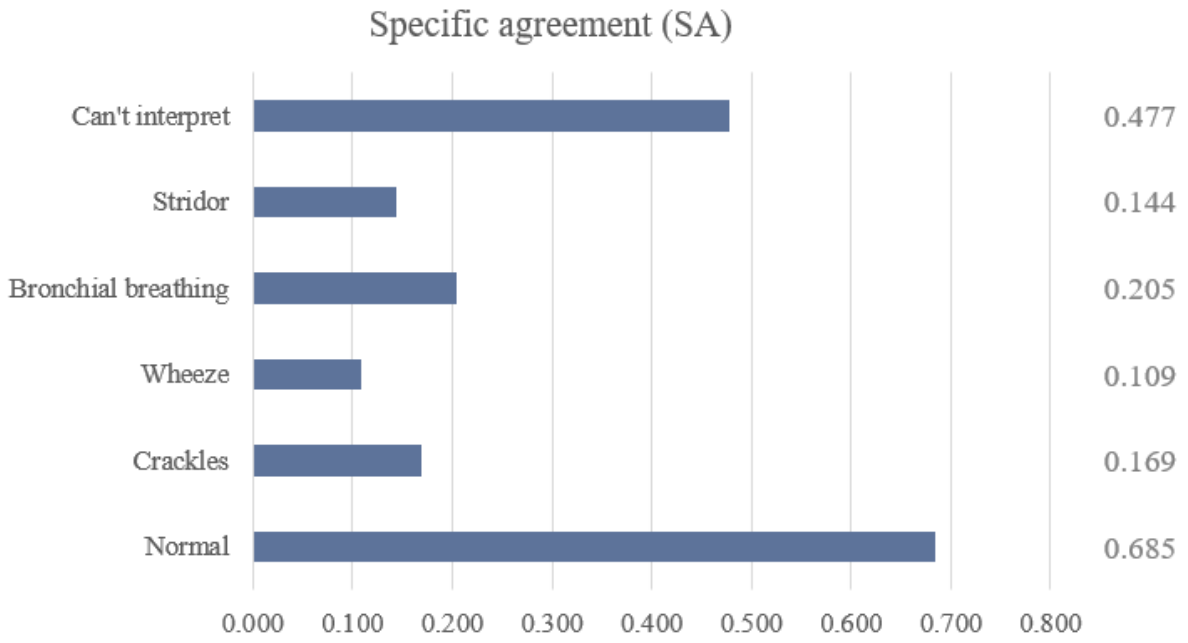


Figure 8.5: Annotator agreement (specific agreement) between 7 observers on labelling of lung sounds from the 700 baseline recordings.

		Annotator						
		1	*2*	*3*	*4*	*5*	*6*	*7*
Sensitivity	**Normal**	0.883	0.771	0.757	0.910	0.905	0.897	0.234
	Crackles	0	1	0.400	0.571	0.800	0.571	0
	Wheeze	0	1	1	0.250	0	0.333	0.333
	Bronchial breathing	0.600	0.000	0.286	0.800	0.667	0.375	0.500
	Stridor	0	NA	NA	NA	0	0	0
	Can't interpret	0.959	0.825	0.673	0.403	0.742	0.661	0.047
Specificity	**Normal**	0.717	0.717	0.776	0.517	0.429	0.444	0.738
	Crackles	0.983	0.827	0.915	0.831	0.601	0.790	0.944
	Wheeze	0.987	0.924	0.901	0.978	0.613	0.854	0.942
	Bronchial breathing	0.890	0.938	0.966	0.932	0.841	0.914	0.941
	Stridor	0.988	0.844	0.968	0.994	1	1	0.976
	Can't interpret	0.625	0.652	0.581	0.947	0.811	0.795	0.980

Table 8.3: Sensitivity and specificity for agreement between individual annotators and majority vote.

Results for the annotation by individual annotators and the majority vote, derived according to the methodology introduced in Chapter 5, are summarised in Table 8.3; most of the annotators are very specific in their annotation of pathological sounds but not very sensitive.

Results on all annotator samples

The breakdown of labels for all signals labelled per annotator is summarised in Table 8.4. Similar to the baseline, it is evident that there is a substantial variability in the incidence of pathological lung sounds as a percentage of labels processed by each annotator. For example, the incidence of crackles as a percentage of all annotated signals ranges between 0.9% (annotator JR 1) and 16.7% (annotator Sion 1). Given the large number of recordings per annotator, between 2100 and 2900, and the randomised process of assigning recordings for labelling, it is reasonable to expect that the incidence of individual lung sounds as a percentage of all signals labelled by an annotator should be comparable across annotators. No improvement was achieved when the labels were consolidated to a participant level (i.e. participants reported to have a given lung sound in any of their recordings); in fact, the differences in incidence rate exacerbated further (Table 8.5).

Finally, the distribution of assigned labels across the health conditions of corresponding cases is visualised in Table 8.6. Namely, participants were divided in four main disease categories – healthy, pneumonia, other respiratory diseases and non-respiratory diseases. Annotators were blinded to the diagnosis at the time of labelling the signals. Hence, Table 8.6 was designed to identify obvious clinical contradictions, such as assigning pathological sounds to children who are either healthy or suffer from non-respiratory conditions. Results were reported as a percentage of all cases in the given disease category. For some annotators (e.g. Sion 1), as many as 39.4% of the signals from healthy patients were labelled as pathological. This analysis indicates that individual labelling is unreliable as a gold standard and hence alternative methods, such as majority voting, are required.

JR 1	Total	%
Number of annotated samples	*2152*	*100.0*
Crackles	19	*0.9*
Wheeze	32	*1.5*
Bronchial breathing	145	*6.7*
Stridor	15	*0.7*
Normal	1214	*56.4*
Can't interpret	727	*33.8*

JR 2	Total	%
Number of annotated samples	*2254*	*100.0*
Crackles	148	*6.6*
Wheeze	98	*4.3*
Bronchial breathing	49	*2.2*
Stridor	134	*5.9*
Normal	975	*43.3*
Can't interpret	850	*37.7*

JR 3	Total	%
Number of annotated samples	*2159*	*100.0*
Crackles	80	*3.7*
Wheeze	142	*6.6*
Bronchial breathing	33	*1.5*
Stridor	37	*1.7*
Normal	1192	*55.2*
Can't interpret	675	*31.3*

JR 4	Total	%
Number of annotated samples	*2292*	*100.0*
Crackles	261	*11.4*
Wheeze	33	*1.4*
Bronchial breathing	79	*3.4*
Stridor	42	*1.8*
Normal	1560	*68.1*
Can't interpret	317	*13.8*

Sion 1	Total	%
Number of annotated samples	*2907*	*100.0*
Crackles	486	*16.7*
Wheeze	495	*17.0*
Bronchial breathing	165	*5.7*
Stridor	0	*0.0*
Normal	1242	*42.7*
Can't interpret	519	*17.9*

Sion 2	Total	%
Number of annotated samples	*2542*	*100.0*
Crackles	155	*6.1*
Wheeze	127	*5.0*
Bronchial breathing	49	*1.9*
Stridor	0	*0.0*
Normal	2047	*80.5*
Can't interpret	164	*6.5*

Table 8.4: Summary of lung sound labels assigned by each annotator, presented as a total number of signals as well as a percentage of all annotated samples.

JR 1	Total	%
Number of annotated patients	*634*	*100.0*
Crackles	11	*1.7*
Wheeze	15	*2.4*
Bronchial breathing	82	*12.9*
Stridor	6	*0.9*
Normal	469	*74.0*
Can't interpret	369	*58.2*

JR 2	Total	%
Number of annotated patients	*657*	*100.0*
Crackles	110	*16.7*
Wheeze	50	*7.6*
Bronchial breathing	36	*5.5*
Stridor	83	*12.6*
Normal	454	*69.1*
Can't interpret	372	*56.6*

JR 3	Total	%
Number of annotated patients	*651*	*100.0*
Crackles	59	*9.1*
Wheeze	68	*10.4*
Bronchial breathing	29	*4.5*
Stridor	21	*3.2*
Normal	464	*71.3*
Can't interpret	355	*54.5*

JR 4	Total	%
Number of annotated patients	*683*	*100.0*
Crackles	158	*23.1*
Wheeze	27	*4.0*
Bronchial breathing	51	*7.5*
Stridor	16	*2.3*
Normal	558	*81.7*
Can't interpret	156	*22.8*

Sion 1	Total	%
Number of annotated patients	*580*	*100.0*
Crackles	261	*45.0*
Wheeze	259	*44.7*
Bronchial breathing	116	*20.0*
Stridor	0	*0.0*
Normal	477	*82.2*
Can't interpret	338	*58.3*

Sion 2	Total	%
Number of annotated patients	*344*	*100.0*
Crackles	35	*10.2*
Wheeze	20	*5.8*
Bronchial breathing	108	*31.4*
Stridor	21	*6.1*
Normal	260	*75.6*
Can't interpret	126	*36.6*

Sion 4	Total	%
Number of annotated patients	*1062*	*100.0*
Crackles	70	*6.6*
Wheeze	58	*5.5*
Bronchial breathing	78	*7.3*
Stridor	28	*2.6*
Normal	874	*82.3*
Can't interpret	31	*2.9*

Table 8.5: Summary of lung sound labels assigned by each annotator to patients, presented as a total number of patients as well as a percentage of all annotated patients.

JR 1

	Healthy control	Pneumonia	Other respiratory	Non-respiratory
Crackles	0.000	0.034	0.015	0.000
Wheeze	0.000	0.046	0.019	0.000
Bronchial	0.100	0.172	0.127	0.070
Stridor	0.000	0.011	0.011	0.028
Normal	0.500	0.437	0.476	0.606
Can't interpret	0.400	0.299	0.352	0.296

JR 2

	Healthy control	Pneumonia	Other respiratory	Non-respiratory
Crackles	0.161	0.157	0.118	0.066
Wheeze	0.032	0.056	0.082	0.053
Bronchial	0.097	0.045	0.032	0.039
Stridor	0.032	0.146	0.100	0.053
Normal	0.355	0.270	0.379	0.461
Can't interpret	0.323	0.326	0.289	0.329

JR 3

	Healthy control	Pneumonia	Other respiratory	Non-respiratory
Crackles	0.038	0.137	0.073	0.045
Wheeze	0.000	0.178	0.113	0.091
Bronchial	0.038	0.014	0.028	0.015
Stridor	0.000	0.068	0.032	0.030
Normal	0.538	0.301	0.474	0.500
Can't interpret	0.385	0.301	0.279	0.318

JR 4

	Healthy control	Pneumonia	Other respiratory	Non-respiratory
Crackles	0.077	0.189	0.190	0.162
Wheeze	0.000	0.056	0.036	0.074
Bronchial	0.077	0.144	0.073	0.029
Stridor	0.000	0.056	0.020	0.029
Normal	0.654	0.422	0.547	0.559
Can't interpret	0.192	0.133	0.134	0.147

Sion 1

	Healthy control	Pneumonia	Other respiratory	Non-respiratory
Crackles	0.212	0.239	0.216	0.194
Wheeze	0.061	0.221	0.225	0.153
Bronchial	0.121	0.168	0.105	0.112
Stridor	0.000	0.000	0.000	0.000
Normal	0.364	0.212	0.301	0.347
Can't interpret	0.242	0.159	0.152	0.194

Sion 2

	Healthy control	Pneumonia	Other respiratory	Non-respiratory
Crackles	0.231	0.154	0.171	0.150
Wheeze	0.154	0.205	0.110	0.217
Bronchial	0.077	0.103	0.079	0.050
Stridor	0.000	0.000	0.000	0.000
Normal	0.385	0.385	0.496	0.417
Can't interpret	0.154	0.154	0.145	0.167

Sion 3

	Healthy control	Pneumonia	Other respiratory	Non-respiratory
Crackles	0.000	0.051	0.053	0.017
Wheeze	0.077	0.026	0.031	0.033
Bronchial	0.038	0.141	0.197	0.083
Stridor	0.000	0.013	0.039	0.033
Normal	0.154	0.269	0.307	0.267
Can't interpret	0.115	0.115	0.136	0.233

Table 8.6: Distribution of annotated lung sounds across disease categories (healthy, pneumonia, other respiratory and non-respiratory). Results are reported as a percentage of all cases in that disease category. Clinical contradictions, such as pathological lung sounds in healthy children or in non-respiratory cases, are highlighted in blue.

8.4. Results: diagnostic value of clinically annotated sounds and signals

Chapter 3 & 4 described in detail the two types of annotation of lung signals: (1) *lung sound labelling* performed by a single clinician on the Peru dataset which resulted in precise delineation of individual sounds; (2) *lung signal labelling* performed by multiple annotators on the Indian dataset which resulted in overall signal content labelling. The diagnostic value features derived from each type of labelling is reported next, both in isolation as well as in combination with vital signs. The attained performance will be used as a benchmark against which the proposed models for automated analysis are compared.

8.4.1. Lung sound annotation (Peru dataset)

The predictive value of the different annotators' labels for identifying one of three respiratory conditions (bronchitis, asthma and pneumonia) was quantified in terms of sensitivity and specificity and listed in Table 8.7. As previously highlighted in Chapter 2, lung sounds are quite specific but not very sensitive in identifying pneumonia. Next, the lung sound labels were converted into features which can be used for the training and testing of a classifier for respiratory disease prediction. A random forests classifier was used and results are summarised at the bottom of Table 8.7. Classification performance for pneumonia delivered **sensitivity of 53.0% and specificity of 72.2%**. This will be used as a benchmark for classification performance of model 1.

	Asthma		**Bronchitis**		**Pneumonia**	
	Sensitivity	*Specificity*	*Sensitivity*	*Specificity*	*Sensitivity*	*Specificity*
Rhonchus	0.243	0.647	0.424	0.730	0.239	0.665
Wheeze	0.014	0.632	0.169	0.730	0.522	0.839
Crackles	0	0.586	0.407	0.791	0.413	0.776
Bronchial	0	0.662	0.322	0.824	0.348	0.820
Cough	0	0.925	0.051	0.953	0.109	0.969
Other adventitious sounds	0	0.940	0	0.946	0.087	0.975
Conductive sounds	0.068	0.932	0.153	0.966	0	0.913
RF classification (all sounds)	0.778	0.661	0.557	0.725	0.530	0.722

Table 8.7: Sensitivity and specificity values for different lung sounds, labelled by the clinical annotator, as predictors of each of the three respiratory conditions. The last row presents classification results for an RF model built with the annotation labels.

8.4.2. Lung signal annotation (India (Mumbai) dataset)

Similarly, the diagnostic value of lung signal labels for identification of pneumonia was explored based on the majority vote of all annotators' labels on a given signal. The process of majority vote derivation was described in detail in Chapter 3. Results are summarised in Table 8.8. As with the lung sound labels in the Peru dataset, the different pathological lung sounds are very specific but not very sensitive in identifying pneumonia. Next, the signal labels were converted into features for classification of pneumonia on a patient level. Since the Indian dataset contained multiple signals for each patient, the incidence of a given lung sound across all signals for that patient was calculated. As vital signs were also available in this dataset, classifier performance with their addition was also explored. Table 8.9 reports results for an RF classifier trained on the described features, as well as on lung sound annotation

features from individual annotators. Best performance was recorded when the majority vote lung features were combined with vital signs – **sensitivity of 66.8% and specificity of 90.3%**. This result will be used as a benchmark for classification performance of model 2.

	Pneumonia		
	Sensitivity	*Specificity*	*Accuracy*
Normal	0.831	0.072	0.452
Crackles	0.675	0.777	0.726
Wheeze	0.545	0.553	0.549
Bronchial	0.532	0.629	0.581
Stridor	0.338	0.903	0.620
Can't interpret	0.143	0.849	0.496

Table 8.8: Sensitivity and specificity values for different lung signal contents, as labelled by the clinical annotators, as predictors of pneumonia.

	Pneumonia					
	Sensitivity		*Specificity*		*Accuracy*	
	Mean	(95% CI)	**Mean**	(95% CI)	**Mean**	(95% CI)
Majority vote of all annotators	**0.553**	(0.539-0.568)	**0.889**	(0.883-0.895)	**0.721**	(0.712-0.730)
Majority vote of all annotators + vital signs	**0.668**	(0.659-0.676)	**0.903**	(0.899-0.907)	**0.785**	(0.781-0.790)
Annotator 1 + vital signs	**0.623**	(0.601-0.646)	**0.866**	(0.857-0.875)	**0.745**	(0.731-0.758)
Annotator 2 + vital signs	**0.507**	(0.490-0.523)	**0.904**	(0.901-0.908)	**0.705**	(0.696-0.715)
Annotator 3 + vital signs	**0.564**	(0.582-0.545)	**0.894**	(0.890-0.897)	**0.729**	(0.720-0.738)
Annotator 4 + vital signs	**0.652**	(0.621-0.683)	**0.838**	(0.835-0.842)	**0.745**	(0.730-0.760)
Annotator 5 + vital signs	**0.597**	(0.586-0.609)	**0.855**	(0.846-0.863)	**0.726**	(0.718-0.734)
Annotator 6 + vital signs	**0.462**	(0.400-0.524)	**0.903**	(0.882-0.924)	**0.683**	(0.661-0.705)
Annotator 7 + vital signs	**0.826**	(0.812-0.840)	**0.965**	(0.962-0.968)	**0.895**	(0.887-0.904)

Table 8.9: Classification performance of an RF classifier including lung signal labels (either majority vote or individual annotators) combined with vital signs. Performance is quantified in terms of sensitivity and specificity, reported as mean (95% CI).

8.5. Results: Pneumonia classification via model 1

Model 1 was developed and tested on the Peru dataset, making use of the delineations of lung sounds within lung signals as performed by a clinical expert. This section reports on model performance.

8.5.1. Multi-domain features

The feature derivation approach described in 8.2 was applied to derive a feature set of 134 features for the Peru data. Differences in the distribution of feature values across lung sound classes were explored and examples are provided in Figure 8.6. From this, some of the features display clear differences in the main parts of the distributions across lung sound classes – e.g. the mean of the wavelet (db4) feature for crackles and bronchial breathing is much higher than that for normal sounds. Many features contained large numbers of outliers and differences were sometimes observed in the ranges that these outliers occupied across lung sound classes. Taking a closer look at the annotator's labelling, it was clear that the delineation of lung sounds was rarely precise, i.e. the annotator marked regions where crackles occurred but these regions often included other sounds too (normal or noise). Hence, when examining feature distributions across lung sound classes, an overlap is to be expected, particularly for the lung sounds that are short in duration and occur more sporadically, like crackles.

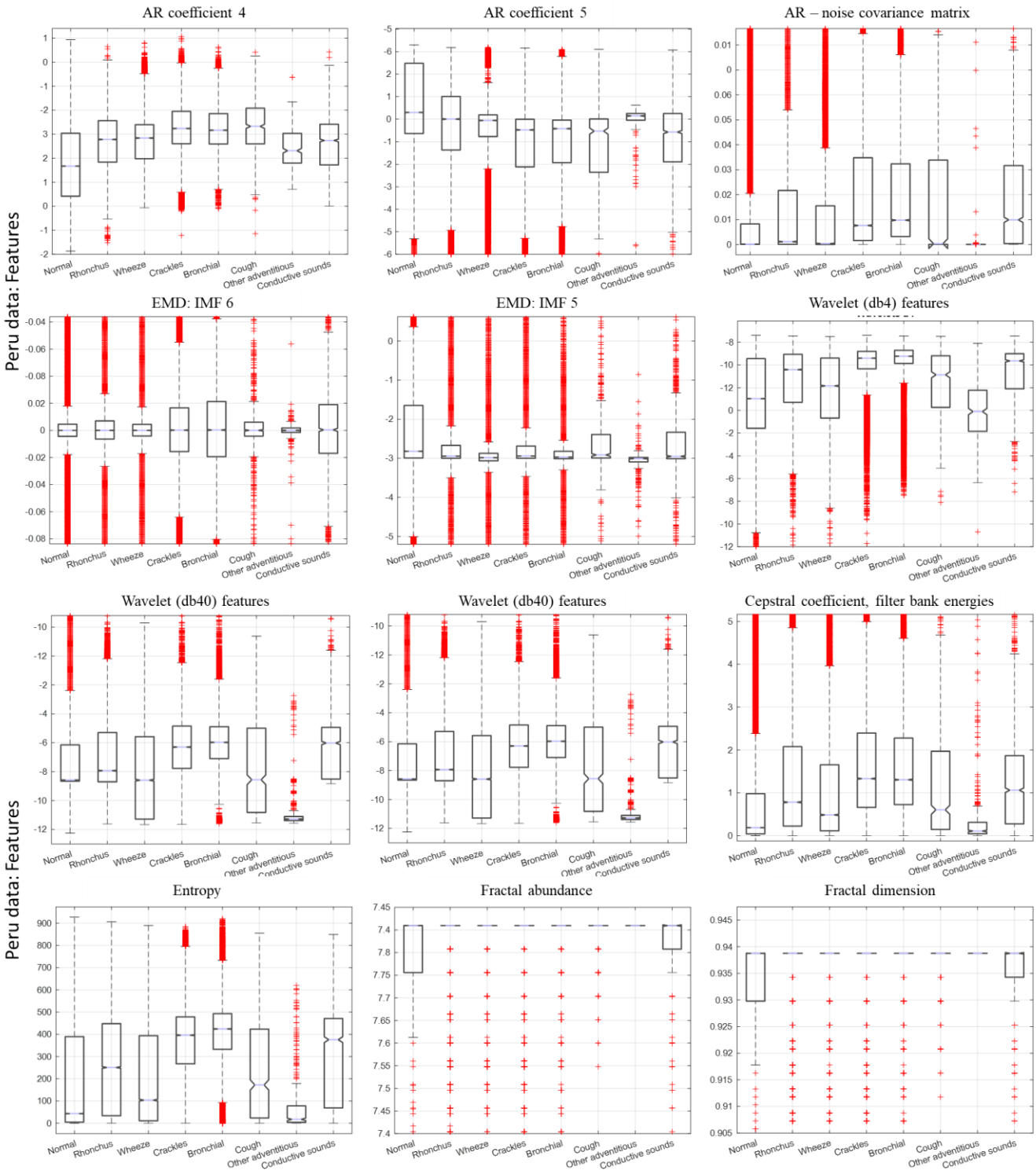


Figure 8.6: Boxplots of some of the features derived for the Peru dataset, divided across lung sound classes. Differences in the main body of the distributions across lung sound classes are observed in AR coefficients, AR noise covariance matrix, mean IMF values, wavelet coefficients, entropy and fractal analysis. The range has been reduced to focus on the main distribution of the boxplots, excluding some of the outliers.

8.5.2. Acoustic modelling of lung sounds

An example of the output of the different HMM models is presented in Figure 8.7, where the original signal and the annotator's labels are included for comparison. As can be seen from this example, the lung sound delineation, as performed by the clinical annotator, was not sound specific but rather labelled approximate regions. Hence, a direct comparison between the HMM output and the labels might be somewhat deceptive, as the algorithm is likely to be more specific. For example, the first crackle region (between 0 and 15 sec) has been marked as one continuous region of crackles by the annotator but is in reality a discontinuous series of crackles. The EM-GMM-HMM model in this region would not generate perfect overlap with the gold standard annotation but is potentially a more realistic representation of the signal.

A more general overview of the HMM output is presented in Table 8.10, where the overlap between annotated lung sounds and HMM predicted states is examined. Results are reported as a fraction of the true outcome (i.e. annotator's labels) rather than absolute values. Additionally, the overlap was calculated on a per-signal basis rather than a per-frame basis, assessing overlap in the reported contents of a given signal, rather than absolute positioning of the lung sound. A threshold, determining the percentage of frames required for an overall signal to be predicted to contain a given lung sound was explored and optimised such that the diagonal in the confusion matrix (i.e. the agreement between annotator's labels and HMM states) is maximised. All three models deliver high sensitivity and identify most of the labels marked by the clinician. They do however deliver poor specificity, predicting additional lung sounds that were not marked by the clinician.

Given the weakness of the gold standard (annotated labels) and the overall goal of diagnosing pneumonia, the HMM model predictions were kept as continuous values (i.e. percentage of the signal predicted to contain certain lung sounds) and their contributions towards an overall diagnostic model for pneumonia diagnosis was examined next.

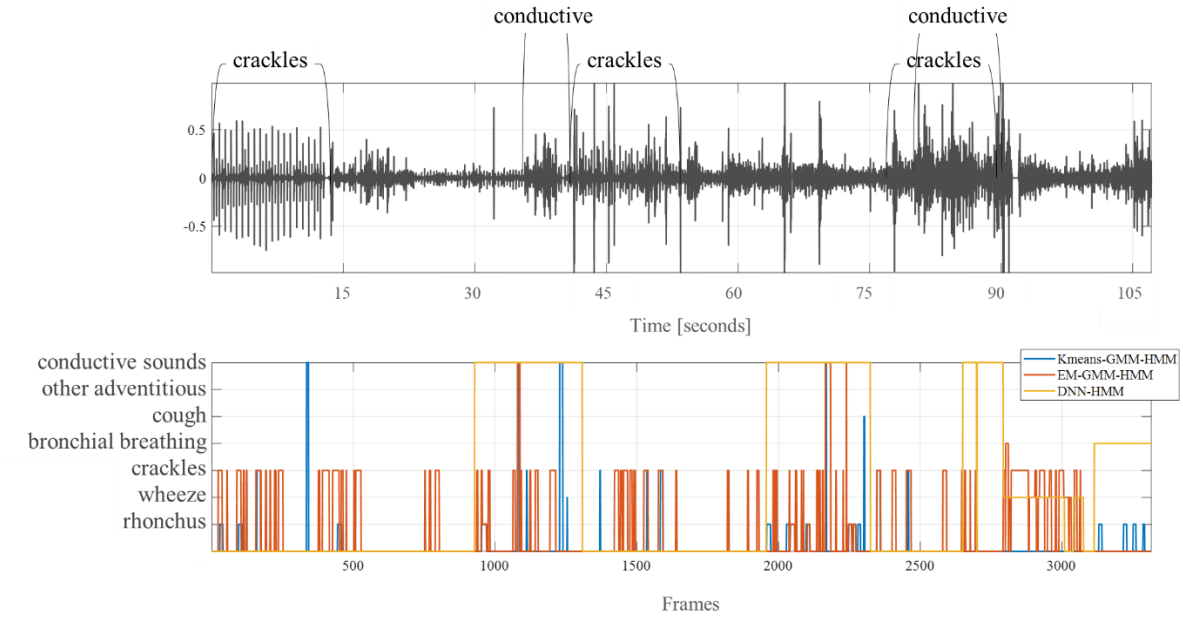


Figure 8.7: Example of the output generated by the different HMM models, compared to the original signal with annotator's labels included above.

Kmeans-GMM-HMM

Annotator labels	*Rhonchus*	*Wheeze*	*Crackles*	*Bronchial*	*Cough*	*Conductive sounds*
Rhonchus	1	0.90	0.54	0.50	0.40	0.10
Wheeze	0.69	1	0.66	0	0.33	0
Crackles	0.23	0.92	1	0.46	0.38	0
Bronchial	0.11	1	0.34	0.60	0.30	0.10
Cough	1.00	1	0.44	0	1	0
Conductive sounds	0.42	0.89	0.23	0.67	0.11	0

EM-GMM-HMM

Annotator labels	*Rhonchus*	*Wheeze*	*Crackles*	*Bronchial*	*Cough*	*Conductive sounds*
Rhonchus	1	0.80	1	0.70	0	0
Wheeze	0.33	1	0	0	0	0
Crackles	0.45	0.92	1	0.69	0.42	0.08
Bronchial	0.66	0.80	0.67	0.60	1	0.10
Cough	0	0.56	0.26	0	0	0
Conductive sounds	1.00	0.89	0.89	0.56	0.60	0

DNN-HMM

Annotator labels	*Rhonchus*	*Wheeze*	*Crackles*	*Bronchial*	*Cough*	*Conductive sounds*
Rhonchus	1	0.10	0.10	0	0	0.80
Wheeze	0	1	0	0	0	0
Crackles	0	0.23	0	0	0	0.85
Bronchial	0	0.30	0	1	0	0.80
Cough	0	1	0	1	1	0.34
Conductive sounds	0	0.11	0	0.89	1	0.78

Table 8.10: Confusion matrix for annotator labelled lung sounds and predicted HMM states, across the three models. The overlap between lung sound classes is reported as percentage of the true outcome (annotated label). Multi-labelling (i.e. multiple lung sounds assigned to the same signal) is possible and the reason why rows do not add up to 1.

8.5.3. RF classification of respiratory conditions

The HMM states, converted into patient features through the approach described in 8.2, were used in the training and test of an RF model for each one of the three respiratory conditions (using a one-versus-all approach) and results are presented next.

8.5.3.1. Classification of pneumonia

The HMM features derived from the two different types of models, EM-GMM-HMM and DNN-HMM, were first used separately. Table 8.11 reposts on classification performance, quantified in terms of sensitivity, specificity and accuracy. From this, RF with the EM-GMM-HMM features was seen to outperform classification with the other two types, delivering **81.7% sensitivity** and **73.7% specificity**. Additionally, the importance scores generated by the RF model were examined and features were grouped according to their lung sound class to examine the significance of different lung sounds, regardless of the model used for derivation, in classifying pneumonia. Results are summarised in Table 8.12, where a combination of wheeze, rhonchus, crackles and bronchial breathing features, derived across HMM models, delivered **84.4% sensitivity** and **74.3% specificity**. Taking a close look, Figure 8.8 includes the confusion matrix and ROC curve generated for the RF model developed on these 12 features (four lung sound features generated from three different HMM models each). Both balanced accuracy and MCC optimisation were performed in the validation set to determine a threshold for the test set and MCC was found to deliver superior results.

	Sensitivity		**Specificity**		**Accuracy**	
	Mean	(95% CI)	**Mean**	(95% CI)	**Mean**	(95% CI)
Benchmark: annotator	**0.530**	(0.512-0.541)	**0.722**	(0.718-0.734)	**0.626**	(0.615-0.31)
EM-GMM-HMM	**0.817**	(0.799-0.835)	**0.737**	(0.723-0.750)	**0.778**	(0.773-0.780)
DNN-HMM	**0.729**	(0.699-0.759)	**0.682**	(0.674-0.691)	**0.706**	(0.693-0.719)

Table 8.11: Sensitivity, specificity and accuracy for RF models based on the three different types of features and developed for the classification of pneumonia. Classification with annotator's labels is included as benchmark comparison.

	Sensitivity		Specificity		Accuracy	
	Mean	(95% CI)	**Mean**	(95% CI)	**Mean**	(95% CI)
Benchmark: annotator	**0.530**	(0.512-0.541)	**0.722**	(0.718-0.734)	**0.626**	(0.615-0.31)
Wheeze & rhonchus	**0.820**	(0.792-0.849)	**0.738**	(0.725-0.750)	**0.779**	(0.766-0.793)
Wheeze & rhonchus & crackles	**0.759**	(0.732-0.787)	**0.757**	(0.737-0.776)	**0.758**	(0.752-0.765)
Wheeze & rhonchus & crackles & bronchial	**0.844**	(0.806-0.882)	**0.743**	(0.730-0.757)	**0.794**	(0.775-0.813)
Wheeze & rhonchus & crackles & bronchial & conductive	**0.820**	(0.796-0.844)	**0.745**	(0.723-0.766)	**0.783**	(0.771-0.794)

Table 8.12: Sensitivity, specificity and accuracy for RF models for pneumonia classification based on lung sound features, derived across the three different types of HMM models (e.g. wheeze from Kmeans-GMM-HMM, wheeze from EM-GMM-HMM and wheeze from DNN-HMM). Classification with annotator's labels is included as benchmark comparison.

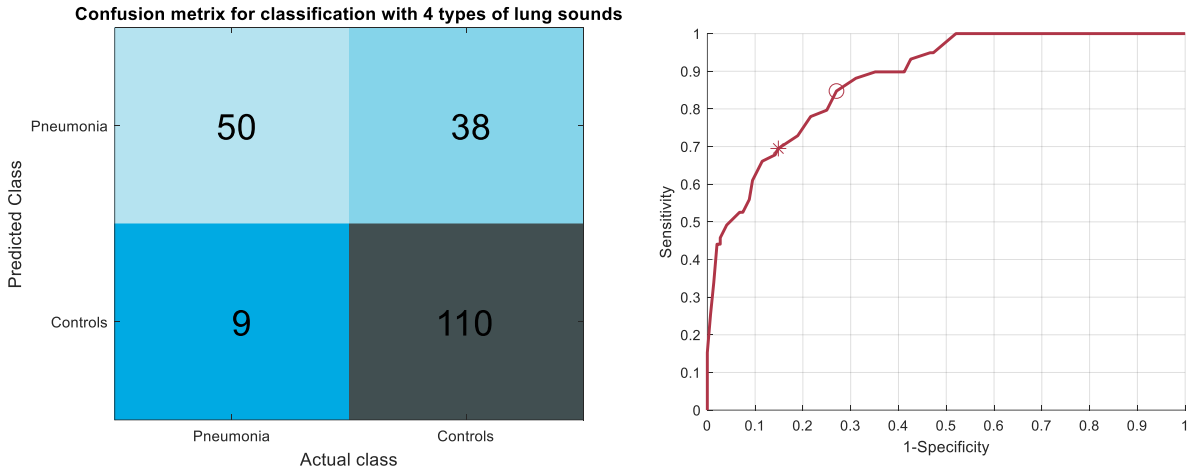


Figure 8.8: Confusion matrix [LEFT] and ROC curve [RIGHT] for RF classification of pneumonia combining 12 features, derived via all three HMM models and features corresponding to four lung sound types – wheeze, rhonchus, crackles and bronchial breathing. The star markers correspond to thresholding optimisation via balanced accuracy and the circles to optimisation via MCC.

8.5.3.2. Classification of other respiratory conditions

Model 1 was also implemented for the classification of asthma and bronchitis, as illustrated in Table 8.13. With lung sound features derived from Kmeans-GMM-HMM, an RF model for asthma achieved **87.0% sensitivity** and **83.2% specificity**. With lung sound features derived from EM-GMM-HMM, an RF model for bronchitis achieved **72.1% sensitivity** and **75.3% specificity**.

		Sensitivity		Specificity		Accuracy	
		Mean	(95% CI)	**Mean**	(95% CI)	**Mean**	(95% CI)
Asthma	**Benchmark: annotator**	**0.778**	-	**0.661**	-	**0.720**	-
	EM-GMM-HMM	**0.913**	(0.880-0.946)	**0.671**	(0.648-0.693)	**0.792**	(0.783-0.801)
	DNN-HMM	**0.896**	(0.851-0.940)	**0.713**	(0.697-0.729)	**0.804**	(0.790-0.819)
Bronchitis	**Benchmark: annotator**	**0.557**	-	**0.725**	-	**0.641**	-
	EM-GMM-HMM	**0.721**	(0.664-0.779)	**0.753**	(0.740-0.766)	**0.737**	(0.712-0.763)
	DNN-HMM	**0.636**	(0.578-0.694)	**0.706**	(0.680-0.732)	**0.671**	(0.638-0.704)

Table 8.13: Sensitivity, specificity and accuracy for RF models based on the three different types of features and developed for the classification of asthma & bronchitis. Classification with annotator's labels is included as benchmark comparison.

8.6. Results: Pneumonia classification via model 2

Model 2 was developed and tested on the Indian (Mumbai) dataset as an alternative for lung sound analysis in the absence of lung sound delineation.

8.6.1. Multi-domain features

As no gold standard information on the specific location of lung sounds in the signal was available, it was impossible to perform feature distribution comparisons across the different lung sound classes. Two alternative feature distribution comparisons were attempted: (1) across signals labelled as containing different lung sounds, with examples included in Figure 8.9 [bottom 2 rows]; (2) across signals of patients suffering from different types of conditions, with examples illustrated in Figure 8.9 [top 2 rows]. For both of these, no quantifiable differences in the main moments of the feature distributions across classes were found. This was expected as most signals, regardless of whether or not the patient has a respiratory condition, contain a large amount of noise or normal sounds which dominate the distribution. Hence, we explored the outliers in the feature distributions, looking for statistically significant differences between signals of normal and abnormal cases. Table 8.14 lists the feature ranges for signals labelled by the annotators as normal; it also reports on the incidence of outliers, beyond that range, across the different classes of lung sounds. Taking into account the typical duration of each of these lung sounds (e.g. 100ms for wheeze), incidence values above that are highlighted. The latter was performed as any outliers with lower incidence are definitely simply noise. Such features were excluded from the feature set, provided they also had equivalent mean and standard deviation values across the lung sound and disease classes.

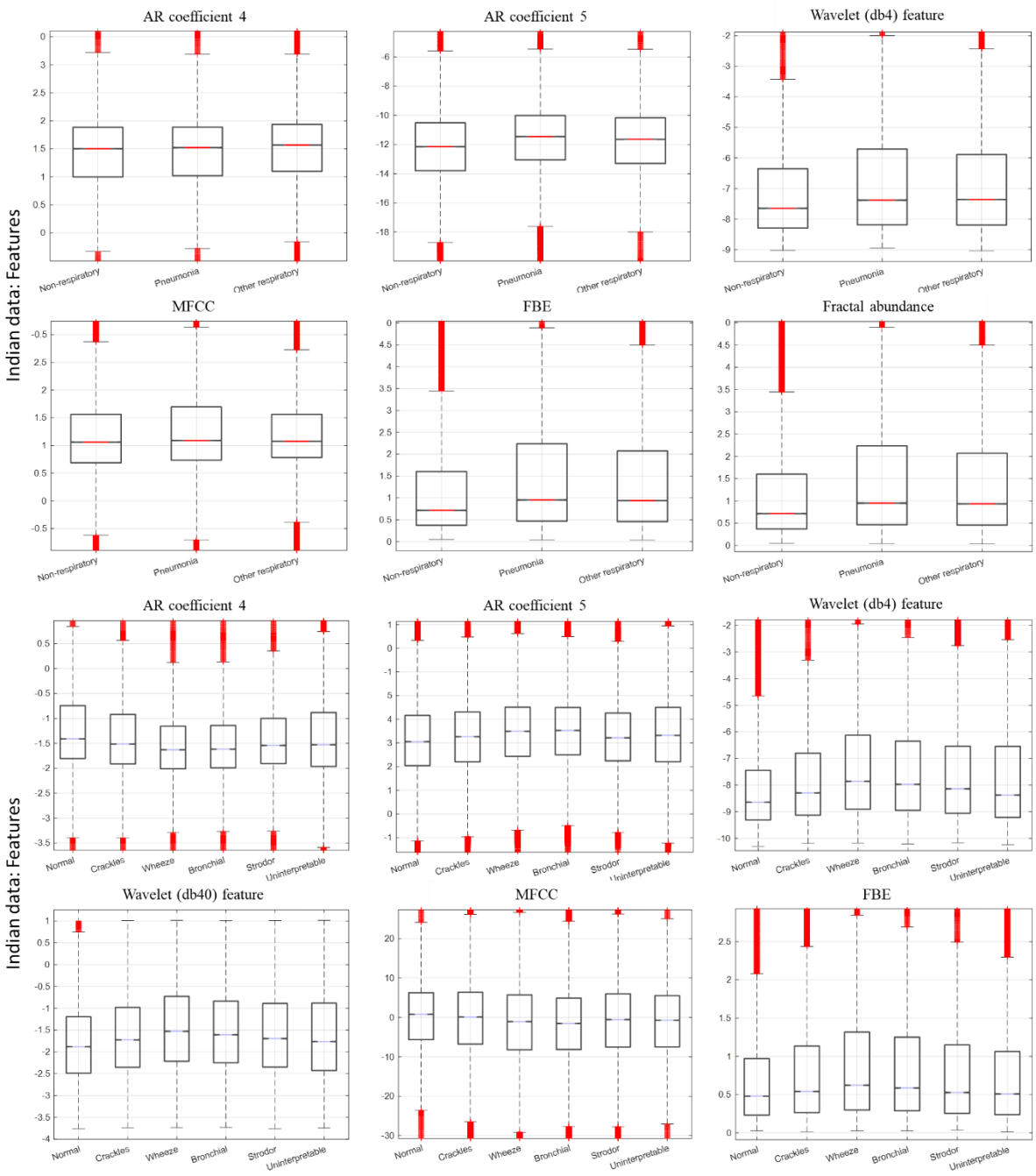


Figure 8.9: Boxplots of some of the features derived for the Indian dataset, divided across signal content labels. The range has been reduced to focus on the main distribution of the boxplots, excluding some of the outliers.

		Normal		Crackles	Wheeze	Bronchial breathing	Stridor
		Min	*Max*	*fraction containing outliers*			
AR	intercept vector	-0.027	0.001	2.81E-04	3.93E-04	4.98E-04	0
AR	ceofficient	-0.023	2.397	1.13E-03	1.96E-03	1.99E-03	0
AR	ceofficient	-3.339	0.958	8.44E-04	1.18E-03	0	0
AR	ceofficient	-0.960	15.705	2.81E-04	3.93E-04	0	0
AR	ceofficient	-17.161	1.337	0	0	0	0
AR	ceofficient	-1.266	1.399	0	3.93E-04	4.98E-04	0
AR	ceofficient	-1.274	4.961	0	3.93E-04	4.98E-04	0
AR	non-zero coeff	1.000	6.000	0	0	0	0
AR	noise covariance	0.000	0.044	0	0	0	0
AR	eigenmode	-0.957	1.000	0	7.85E-04	0	0
AR	eigenmode	-0.965	0.999	8.44E-04	3.93E-04	4.98E-04	0
AR	eigenmode	-0.974	0.998	0	3.93E-04	4.98E-04	0
AR	eigenmode	-0.971	1.000	1.41E-03	3.93E-04	9.96E-04	9.95E-04
AR	eigenmode	-0.976	0.988	2.81E-04	0	0	9.95E-04
AR	eigenmode	-0.959	0.996	2.81E-04	0	4.98E-04	0
EMD	IMF1	-2.146	-0.604	1.69E-03	1.96E-03	1.99E-03	0
EMD	IMF2	-0.390	1.877	8.44E-04	7.85E-04	4.98E-04	0
EMD	IMF3	-1.502	0.619	0	0	0	0
EMD	IMF4	-0.603	1.032	0	7.85E-04	4.98E-04	0
EMD	IMF5	-0.699	0.463	2.81E-04	0	0	9.95E-04
EMD	IMF6	-0.436	0.465	2.81E-04	0	0	9.95E-04
Wavelets	db4 ceofficient	-4.790	-0.009	0	0	0	0
Wavelets	db4 ceofficient	-4.444	0.334	0	0	0	0
Wavelets	db4 ceofficient	-9.590	-2.203	0	0	0	0
Wavelets	db4 ceofficient	-4.098	0.675	2.81E-04	0	0	0
Wavelets	db4 ceofficient	-9.484	-1.337	5.63E-04	3.93E-04	0	0
Wavelets	db4 ceofficient	-9.783	-2.379	0	0	0	0
Wavelets	db4 ceofficient	-9.511	-2.049	8.44E-04	7.85E-04	0	0
Wavelets	db4 ceofficient	-3.753	1.010	0	0	0	0
Wavelets	db4 ceofficient	-9.490	-0.549	0	0	0	0
Wavelets	db4 ceofficient	-9.532	-1.685	0	0	0	0
Wavelets	db4 ceofficient	-9.514	-1.140	8.44E-04	1.18E-03	0	9.95E-04
Wavelets	db4 ceofficient	-10.039	-2.375	0	0	0	0
Wavelets	db4 ceofficient	-9.712	-2.313	0	3.93E-04	0	0
Wavelets	db4 ceofficient	-9.544	-1.932	0	0	0	0
Wavelets	db4 ceofficient	-9.591	-2.184	0	0	0	0
Wavelets	db4 ceofficient	-3.410	1.349	0	0	0	0
Wavelets	db4 ceofficient	-9.542	0.215	0	0	0	0
Wavelets	db4 ceofficient	-9.624	-0.770	0	0	0	0
Wavelets	db4 ceofficient	-9.551	-0.317	0	7.85E-04	0	9.95E-04
Wavelets	db4 ceofficient	-9.623	-1.781	0	7.85E-04	0	0
Wavelets	db4 ceofficient	-9.667	-1.561	0	3.93E-04	0	0
Wavelets	db4 ceofficient	-9.566	-1.002	0	0	0	0
Wavelets	db4 ceofficient	-9.610	-1.334	2.81E-04	1.96E-03	0	9.95E-04
Wavelets	db4 ceofficient	-10.310	-2.288	0	3.93E-04	0	0
Wavelets	db4 ceofficient	-9.954	-2.358	0	0	0	0
Wavelets	db4 ceofficient	-9.691	-2.192	0	7.85E-04	0	0
Wavelets	db4 ceofficient	-9.855	-2.403	2.81E-04	0	0	9.95E-04
Wavelets	db4 ceofficient	-9.605	-1.841	2.81E-04	0	0	0
Wavelets	db4 ceofficient	-9.654	-1.989	0	0	0	0
Wavelets	db4 ceofficient	-9.676	-2.250	0	0	0	0
Wavelets	db4 ceofficient	-9.667	-2.081	0	0	0	0
Wavelets	db40 ceofficient	-4.790	-0.009	0	0	0	0
Wavelets	db40 ceofficient	-4.446	0.335	0	0	0	0
Wavelets	db40 ceofficient	-9.604	-2.272	0	0	0	0
Wavelets	db40 ceofficient	-4.104	0.675	2.81E-04	0	0	0
Wavelets	db40 ceofficient	-9.472	-1.396	5.63E-04	3.93E-04	0	0
Wavelets	db40 ceofficient	-9.771	-2.455	0	0	0	0
Wavelets	db40 ceofficient	-9.515	-2.100	1.13E-03	7.85E-04	0	9.95E-04
Wavelets	db40 ceofficient	-3.767	1.013	0	0	0	0
Wavelets	db40 ceofficient	-9.496	-0.653	0	0	0	0
Wavelets	db40 ceofficient	-9.532	-1.690	0	0	0	0
Wavelets	db40 ceofficient	-9.502	-1.203	5.63E-04	1.18E-03	4.98E-04	1.99E-03
Wavelets	db40 ceofficient	-10.032	-2.439	2.81E-04	7.85E-04	0	0
Wavelets	db40 ceofficient	-9.694	-2.433	0	0	0	0
Wavelets	db40 ceofficient	-9.544	-1.982	0	0	0	0

		Normal		Crackles	Wheeze	Bronchial breathing	Stridor
		Min	*Max*	*fraction containing outliers*			
Wavelets	db40 ceofficient	-3.437	1.356	0	0	0	0
Wavelets	db40 ceofficient	-9.539	0.248	2.81E-04	0	0	0
Wavelets	db40 ceofficient	-9.625	-0.810	0	0	0	0
Wavelets	db40 ceofficient	-9.556	-0.455	0	3.93E-04	0	9.95E-04
Wavelets	db40 ceofficient	-9.620	-1.807	0	0	0	0
Wavelets	db40 ceofficient	-9.632	-1.591	0	0	0	0
Wavelets	db40 ceofficient	-9.611	-1.071	0	3.93E-04	0	0
Wavelets	db40 ceofficient	-9.616	-1.320	0	7.85E-04	0	0
Wavelets	db40 ceofficient	-10.320	-2.475	0	7.85E-04	0	0
Wavelets	db40 ceofficient	-10.039	-2.409	0	0	0	0
Wavelets	db40 ceofficient	-9.699	-2.332	0	0	0	0
Wavelets	db40 ceofficient	-9.865	-2.408	0	0	4.98E-04	0
Wavelets	db40 ceofficient	-9.624	-1.904	0	0	0	0
Wavelets	db40 ceofficient	-9.653	-2.027	0	0	4.98E-04	0
Wavelets	db40 ceofficient	-9.680	-2.244	0	0	0	0
Wavelets	db40 ceofficient	-9.673	-2.159	2.81E-04	3.93E-04	0	0
MFCC	coefficient	44.865	96.781	0	0	0	0
MFCC	coefficient	-12.382	14.429	2.81E-04	0	0	0
MFCC	coefficient	-11.427	33.652	0	0	0	0
MFCC	coefficient	-21.881	23.048	1.13E-03	1.18E-03	9.96E-04	0
MFCC	coefficient	-39.171	13.965	2.25E-03	2.36E-03	1.99E-03	9.95E-04
MFCC	coefficient	-27.750	21.960	0	0	0	0
MFCC	coefficient	-24.368	28.231	8.44E-04	0	4.98E-04	9.95E-04
MFCC	coefficient	-20.595	27.645	7.32E-03	7.85E-04	4.98E-04	0
MFCC	coefficient	-29.337	25.661	8.44E-04	0	0	1.99E-03
MFCC	coefficient	-15.305	23.875	2.81E-04	3.93E-04	0	9.95E-04
MFCC	coefficient	-24.886	18.964	8.44E-04	1.18E-03	0	2.98E-03
MFCC	coefficient	-25.318	23.571	0	3.93E-04	4.98E-04	0
MFCC	coefficient	-24.582	18.473	1.13E-03	2.75E-03	0.00E+00	1.99E-03
MFCC	FBE	10322	4148369	2.81E-04	1.18E-03	0	0
MFCC	FBE	3502	4330516	0	3.93E-04	0	0
MFCC	FBE	1846	5485191	2.81E-04	7.85E-04	4.98E-04	0
MFCC	FBE	742	7904341	2.81E-04	0	0	0
MFCC	FBE	499	10008084	2.81E-04	3.93E-04	1.99E-03	5.97E-03
MFCC	FBE	422	12300805	2.81E-04	3.93E-04	0	9.95E-04
MFCC	FBE	292	13476480	0	0	0	0
MFCC	FBE	232	10227934	0	0	4.98E-04	0
MFCC	FBE	257	6889617	2.81E-04	1.18E-03	9.96E-04	0
MFCC	FBE	226	8813489	0	0	4.98E-04	0
MFCC	FBE	325	11041801	0	0	0	0
MFCC	FBE	328	12777901	0	0	0	0
MFCC	FBE	417	14559428	0	0	0	0
MFCC	FBE	515	14001323	0	0	0	0
MFCC	FBE	679	12667752	2.81E-04	1.57E-03	9.96E-04	0
MFCC	FBE	811	14339642	0	3.93E-04	0	0
MFCC	FBE	986	16407887	0	0	4.98E-04	0
MFCC	FBE	1369	14797666	0	0	0	0
MFCC	FBE	1359	15605514	0	7.85E-04	9.96E-04	0
MFCC	FBE	1531	17535250	0	7.85E-04	4.98E-04	0
Kurtosis	Kurtosis	1.015	643.039	5.63E-04	4.71E-03	0	9.95E-04
Skewness	Skewness	-11.433	13.807	0	7.85E-04	9.96E-04	0
Entropy	Entropy	1.656	742.620	0	7.85E-04	4.98E-04	0
Fractal ana	lacunarity	2631.0	2646.0	0	0	0	0
Fractal ana	lacunarity	1319.0	1323.0	0	0	0	0
Fractal ana	lacunarity	661.0	662.0	2.81E-04	0	0	0
Fractal ana	lacunarity	331.0	331.0	5.63E-04	0	0	0
Fractal ana	lacunarity	166.0	166.0	5.63E-04	0	0	0
Fractal ana	lacunarity	83.0	83.0	5.63E-04	0	0	0
Fractal ana	lacunarity	42.0	42.0	5.63E-04	0	0	0
Fractal ana	lacunarity	21.0	21.0	5.63E-04	0	0	0
Fractal ana	lacunarity	11.0	11.0	5.63E-04	0	0	0
Fractal ana	lacunarity	6.0	6.0	5.63E-04	0	0	0
Fractal ana	lacunarity	3.0	3.0	5.63E-04	0	0	0
Fractal ana	lacunarity	2.0	2.0	5.63E-04	0	0	0
Fractal ana	fractal dimension	1.0	1.0	5.63E-04	0	0	0
Fractal ana	fractal abundance	7.8	7.8	5.63E-04	0	0	0

Table 8.14: Feature range values for normal signals and incidence of outliers for each of the lung sound classes as per the majority vote of annotation. For each lung sound class, incidence value higher than the typical duration of that lung sound, as a fraction of the whole signal, are highlighted in blue.

8.6.2. Acoustic modelling

Using the methodology, described in 8.2, HMM models were trained for signals associated with the three main disease categories – pneumonia, other respiratory disease and non-respiratory disease. The number of possible states in each model was fine-tuned in the training stage of the model, where the likelihood probability was used to maximise correct sequence assignment to a disease class. The following numbers of states were found to be optimal for pneumonia, other respiratory disease and non-respiratory disease – 9, 9, 6 respectively. In order to convert the HMM predictions for all patient signals

into an overall symptom of pneumonia, a threshold was applied determining what percentage of all patient lung signals need to be predicted to have pneumonia in order for their lung sound data to be labelled as indicative of pneumonia. Figure 8.10 illustrates the impact of this varying threshold, used to convert the lung signal predictions into a binary, on the overall diagnosis. Bearing this in mind, the DNN-HMM delivered highest accuracy:

- Pneumonia DNN-HMM states delivered 71.6% accuracy in predicting a pneumonia diagnosis, provided 10% of a patient's signals containing pneumonia DNN-HMM states is taken as a threshold for that patient to be predicted to have pneumonia.
- Other respiratory disease DNN-HMM states delivered 70.6% accuracy in predicting a diagnosis of other respiratory disease, provided 70% of a patient's signals containing other respiratory disease DNN-HMM states is taken as a threshold for that patient to be predicted to have other respiratory disease.
- Non-respiratory disease DNN-HMM states delivered 64.5% accuracy in predicting a diagnosis of other respiratory disease, provided 12% of a patient's signals containing non-respiratory disease DNN-HMM states is taken as a threshold for that patient to be predicted to have other respiratory disease.

Next, the correlation coefficients between the derived HMM states and the majority vote of the annotators' labelled were computed and are presented in Table 8.15. Based on this, it was found that:

- "Non-respiratory HMM" states positively correlated with signals labelled as normal and negatively correlated with wheeze-, crackles- and stridor-containing signals;
- "Pneumonia HMM" states positively correlated with crackles- and wheeze-containing signals and negatively correlated with signals labelled as normal;
- "Other respiratory HMM" states negatively correlated with crackles- and stridor-containing signals.

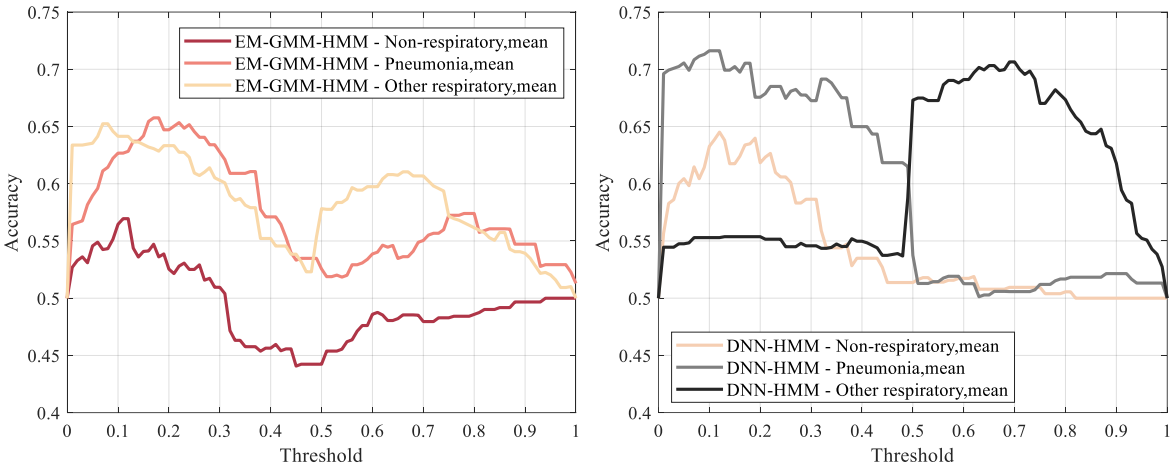


Figure 8.10: Accuracy of prediction of diagnostic outcome based on the HMM outcome derived from the two different methods. A threshold value between 0 and 1 is varied to convert the lung signal predictions into binary vectors of disease association.

	rho		Normal	Crackles	Wheeze	Bronchial breathing	Stridor	Can't interpret
Mean	Kmeans-GMM-HMM	Non-respiratory	0.1912	-0.0218	-0.1926	-0.0661	-0.0872	-0.0211
Mean	Kmeans-GMM-HMM	Pneumonia	-0.1746	0.1662	0.1197	-0.0084	0.0980	-0.0111
Mean	Kmeans-GMM-HMM	Other respiratory	0.0422	-0.1385	0.0089	0.0483	-0.0360	0.0231
Mean	EM-GMM-HMM	Non-respiratory	-0.0118	0.0261	0.0180	0.0759	0.0826	-0.1108
Mean	EM-GMM-HMM	Pneumonia	-0.0161	0.1366	-0.0094	-0.0710	0.0370	-0.0440
Mean	EM-GMM-HMM	Other respiratory	0.0240	-0.1429	-0.0063	0.0014	-0.1002	0.1294
Mean	DNN-HMM	Non-respiratory	0.2248	0.0024	-0.2060	-0.0746	-0.1035	-0.0651
Mean	DNN-HMM	Pneumonia	-0.2482	0.1780	0.1703	0.0287	0.1542	-0.0060
Mean	DNN-HMM	Other respiratory	0.1106	-0.1692	-0.0475	0.0139	-0.0886	0.0414
Std	Kmeans-GMM-HMM	Non-respiratory	0.0884	0.0119	-0.1233	-0.0178	-0.0802	0.0104
Std	Kmeans-GMM-HMM	Pneumonia	-0.1774	0.1324	0.0875	0.0317	0.0945	0.0185
Std	Kmeans-GMM-HMM	Other respiratory	0.0312	-0.1451	-0.0235	0.0375	-0.0645	0.0917
Std	EM-GMM-HMM	Non-respiratory	0.0363	0.0018	-0.0235	0.0387	0.0107	-0.0685
Std	EM-GMM-HMM	Pneumonia	0.0233	0.1180	-0.0390	-0.0912	-0.0110	-0.0261
Std	EM-GMM-HMM	Other respiratory	-0.0014	-0.1134	-0.0285	0.0077	-0.1189	0.1657
Std	DNN-HMM	Non-respiratory	0.2467	-0.1059	-0.1911	-0.0657	-0.1535	-0.0061
Std	DNN-HMM	Pneumonia	-0.1859	0.1258	0.0902	-0.0035	0.1085	0.0523
Std	DNN-HMM	Other respiratory	0.0872	-0.1473	-0.0561	0.0131	-0.1346	0.0880

	p		Normal	Crackles	Wheeze	Bronchial breathing	Stridor	Can't interpret
Mean	Kmeans-GMM-HMM	Non-respiratory	0.0001	0.6652	0.0001	0.1896	0.0834	0.6762
Mean	Kmeans-GMM-HMM	Pneumonia	0.0005	0.0009	0.0174	0.8683	0.0516	0.8264
Mean	Kmeans-GMM-HMM	Other respiratory	0.4031	0.0058	0.8594	0.3386	0.4753	0.6476
Mean	EM-GMM-HMM	Non-respiratory	0.8146	0.6055	0.7213	0.1323	0.1013	0.0276
Mean	EM-GMM-HMM	Pneumonia	0.7499	0.0065	0.8529	0.1592	0.4628	0.3828
Mean	EM-GMM-HMM	Other respiratory	0.6350	0.0044	0.9003	0.9772	0.0466	0.0100
Mean	DNN-HMM	Non-respiratory	0.0000	0.9615	0.0000	0.1391	0.0399	0.1964
Mean	DNN-HMM	Pneumonia	0.0000	0.0004	0.0007	0.5700	0.0021	0.9054
Mean	DNN-HMM	Other respiratory	0.0280	0.0007	0.3466	0.7829	0.0787	0.4118
Std	Kmeans-GMM-HMM	Non-respiratory	0.0794	0.8142	0.0142	0.7248	0.1115	0.8368
Std	Kmeans-GMM-HMM	Pneumonia	0.0004	0.0084	0.0825	0.5300	0.0606	0.7134
Std	Kmeans-GMM-HMM	Other respiratory	0.5366	0.0038	0.6415	0.4573	0.2010	0.0688
Std	EM-GMM-HMM	Non-respiratory	0.4716	0.9710	0.6408	0.4437	0.8328	0.1745
Std	EM-GMM-HMM	Pneumonia	0.6446	0.0190	0.4393	0.0701	0.8277	0.6051
Std	EM-GMM-HMM	Other respiratory	0.9775	0.0242	0.5717	0.8786	0.0181	0.0009
Std	DNN-HMM	Non-respiratory	0.0000	0.0353	0.0001	0.1929	0.0022	0.9031
Std	DNN-HMM	Pneumonia	0.0002	0.0124	0.0732	0.9446	0.0311	0.2999
Std	DNN-HMM	Other respiratory	0.0834	0.0033	0.2659	0.7945	0.0074	0.0806

Table 8.15: Correlation coefficients (top) and corresponding p-values (bottom) for the correlation between derived HMM states and the majority vote of annotations. Highlighted in blue: coefficients greater than |0.1| and p-values <0.05.

8.6.3. RF classification of pneumonia

The performance of an RF model for pneumonia classification, trained on the HMM disease features, with and without the addition of vital signs, is summarised in Table 8.16. When combined with vital signs, features derived from any one of the three HMM methods were seen to deliver good results. DNN-HMM features combined with vital signs delivered the highest accuracy, 82.9%, corresponding to **sensitivity of 81.3%** and **specificity of 84.5%**. Comparing this against RF classification using vital

signs alone (85.9% sensitivity and 69.7% specificity), the addition of the automated lung sound analysis improved specificity substantially whilst only impacting sensitivity minorly. In comparison, pneumonia classification performance with the annotators' labels was also inferior: 55.3% sensitivity and 88.9% specificity with the majority vote of the annotators' labels, and 66.8% sensitivity and 90.3% specificity with the majority vote of the annotators' labels and vital signs. Taking a closer look at the this, Figure 8.11 compares the ROC curve of the RF model with DNN-HMM features + vital signs against three other feature set configurations.

	Sensitivity		Specificity		Accuracy	
	Mean	(95% CI)	**Mean**	(95% CI)	**Mean**	(95% CI)
Benchmark: vital signs	**0.859**	(0.827-0.864)	**0.697**	(0.689-0.705)	**0.778**	(0.771-0.797)
Benchmark: annotator labels	**0.553**	(0.539-0.568)	**0.889**	(0.883-0.895)	**0.721**	(0.712-0.730)
Benchmark: annotator labels + vital signs	**0.668**	(0.659-0.676)	**0.903**	(0.899-0.907)	**0.786**	(0.781-0.790)
EM-GMM-HMM + Vital signs	**0.738**	(0.731-0.745)	**0.884**	(0.875-0.894)	**0.811**	(0.803-0.811)
DNN-HMM + Vital signs	**0.813**	(0.789-0.837)	**0.845**	(0.828-0.863)	**0.829**	(0.820-0.839)

Table 8.16: Sensitivity, specificity and accuracy of an RF classifier of pneumonia for different feature configurations involving the HMM disease states features and vital signs. Classification with annotator's labels and/or vital signs is included as benchmark comparison.

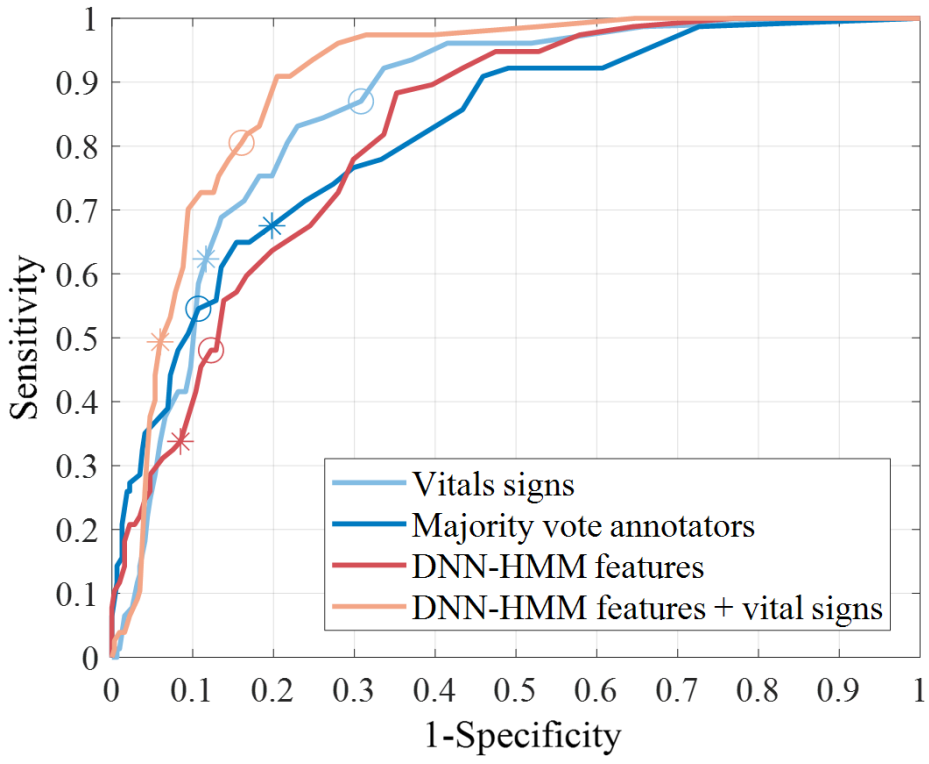


Figure 8.11: ROC curves for RF classification models of pneumonia based on four different sets of features. The star markers correspond to thresholding optimisation via accuracy and the circles to optimisation via MCC.

8.7. Discussion and conclusions

This chapter proposes two new models for lung sound analysis as means for classification of pneumonia – model 1 developed against delineated lung sounds and model 2 developed against overall diagnostic labelling of a signal.

Model 1 was developed on the Peru dataset using the annotations of one clinical expert. It delivered 84.4% (95% CI 80.6-88.2) sensitivity and 74.3% (95% CI 73.0-75.7) specificity in classifying pneumonia against the remainder of the Peru dataset (asthma, bronchitis and healthy cases). This substantially outperformed classification with the annotator's labels (sensitivity of 53.0% and specificity of 72.2%), improving sensitivity by more than 30% whilst also preserving specificity. A comparison between the acoustic modelling done via HMM and the annotator, presented in Section 8.5.2, highlighted specificity in agreement but poor sensitivity, i.e. the HMM predictions picked up on the majority of annotator's labels but also suggested additional signal areas that should be highlighted as pathological lung sounds. These results suggest that, whilst clinical annotators are good at providing a specific gold standard that can inform the training of acoustic models such as HMM, their annotation should not be treated as an exhaustive interpretation of the lung signal. These findings are consistent with auscultation analysis conducted by McCollum et al. on the basis of the PERCH dataset which suggested that digital stethoscopes capture greater amount of subclinical abnormalities compared to acoustic stethoscopes (McCollum et al. 2017). The proposed model 1 builds upon the clinical annotation to deliver a stronger diagnostic prediction, based only on lung signals. Additionally, variability in performance across the LOO cross validation and the 100 model repetitions is low, suggesting good reproducibility of results.

Model 2 was developed on the Indian study data which did not include lung sound delineations of specific sounds. Instead, the model explored the ability of more advanced HMM techniques to quantify the affiliation of whole lung signals with one of three disease classes – pneumonia, other respiratory and non-respiratory disease. Model 2 delivered 81.3% sensitivity (95% CI 78.9-83.7) and 84.5% specificity (95% CI 82.8-86.3), combining the derived acoustic features with vital signs. In comparison, classification with vital signs alone delivered 85.9% sensitivity and 69.7% specificity, and classification

with vital signs in combination with the majority vote of all annotators delivered 66.8% sensitivity and 90.3% specificity. The proposed approach preserves the high sensitivity of the former whilst almost matching the high specificity of the latter. Taking a closer look at the association between the clinical labels of the signals and the predicted disease affiliation, a clear and clinically consistent correlation was reported in Table 8.15 (e.g. signals predicted as pneumonia were correlated with the crackles labels in the majority annotation). The DNN-HMM derived acoustic states were seen to deliver highest agreement with overall diagnosis (Figure 8.10) and best classification performance when combined with vitals signs. This is consistent with findings of researchers in the speech analysis community who have repeatedly highlighted that DNN-HMM is able to better exploit information in neighbouring states and hence model context-dependent states. Variability across the LOO cross validation and 100 repetitions was recorded to be reasonably low. Unfortunately, the dataset available did not have sufficiently high representation of any other disease and did not allow for further stratification of the "other class". Nevertheless, given the high disease variability of the dataset and the realistic nature of the signals analysed (collected in both community and hospital setting), the results obtain in this chapter suggest that modelling of lung signals, without the need of clinical annotation as a baseline, can still improve diagnostic outcomes for childhood pneumonia. Moreover, as McCollum et al. have illustrated before, digital stethoscopes are capable of capturing additional information (e.g. subclinical abnormalities leading to 14% of control population exhibiting crackles and/or wheeze); hence, the diagnostic value of conventional lung sounds derived from digital stethoscopes signals might be weaker than disease-driven interpretation of the signal.

9. Summary of Contributions and Future Work

The overarching objective of this work was to develop tools that improve the diagnosis of childhood pneumonia in three main ways:

- **More accurate**: improve the specificity of diagnosis, whilst preserving sensitivity achieved via conventional tools such as the IMCI guidelines;
- **More reproducible and robust**: reduce inter-observer variability by designing tools for automated analysis that rely on reproducible and objective measurements, rather than subjective observations;
- **More accessible in community-settings**: design tools that enable minimally trained users to use affordable point-of-care devices to derive essential clinical insights, minimising barriers to access due to shortage of clinical expertise.

To deliver these objectives, work on this project followed two main verticals: (1) machine learning tools for improved diagnosis via fusion of symptoms; (2) signal processing and machine learning tools for automated auscultation and further diagnostic improvement. The contributions in each one of these verticals are discussed next in the context of previous and future work.

9.1. Machine learning tools for improved diagnosis of childhood pneumonia

In Chapter 6 & 7, this thesis presented a machine learning approach for the diagnosis of childhood pneumonia on the basis of conventional clinical markers. Models were developed on three separate datasets, where generalisability internally (within an independent test set from the development population) was found to be high. In comparison, the performance of the IMCI guidelines on each of the three populations was found to be inferior. Additionally, clinical decision rules and existing diagnostic models, as reviewed in Chapter 2, delivered at best sensitivity of 78% and specificity of 64%, and required the addition of both observational characteristics and blood test results: Osat, T, duration of fever, RR and CRP (Nijman et al. 2013).

Beyond validation within individual populations, the generalisability of the proposed models across populations was examined in Chapter 7, highlighting two main approaches to developing and validating a generalisable model: (1) developing the model on a 'rich' dataset including a wide range of diseases, such as the dataset collected in the Indian (Mumbai) field study described in Chapter 4; (2) developing a global model on an aggregated set of smaller and more homogeneous studies and validating performance internally.

Beyond the identification of pneumonia, Chapter 6 presented a machine learning framework for the tri-fold diagnostic evaluation of children who might have pneumonia, also including evaluation of severity and determination of aetiology. It was shown that severity can be determined through a combination of three vital signs and two lung sounds (72.4% sensitivity and 82.2% specificity); addition of a conventional biomarker (C-reactive protein) further improved severity predictions (89.1% sensitivity and 81.3% specificity). Finally, the analysis demonstrated that aetiology can be determined using three vital signs and a newly proposed biomarker (Lcn2) (81.8% sensitivity and 90.6% specificity).

Blood-based biomarkers were included in the analysis despite the fact that affordable point-of-care tools might not be commercially available for all of them yet. Nevertheless, research in this area has delivered promising results. For example, Martinez et al. reported a production price of US$0.01 for a paper-based analytical device and multiple applications for this type of technology have been explored (Martinez et al. 2010). Point-of-care assays for CRP have been developed by several commercial providers but are still not widely available. Lcn2 cannot be currently measured in a point-of-care setting. However, the results obtained in this study highlighted that CRP and Lcn2 could facilitate both severity and aetiology determination, supporting the need for the development of affordable point-of-care assays for both biomarkers. This approach could provide an alternative in settings where access to advanced medical technologies is limited. However, a substantially larger dataset is required to validate the ability of this approach to replace the use of X-rays and blood culture.

The diagnostic models developed in Chapters 6 & 7, demonstrated the ability of machine learning to combine highly sensitive features (such as vital signs) with highly specific features (such as lung sounds)

and deliver more balanced diagnostic performance. However, this approach assumes the availability of lung sound information (e.g. presence of crackles), which is currently only attainable via examination by a clinical expert. Hence, the second work vertical in this thesis was dedicated to the development of tools for the automated interpretation of digital stethoscope signals. An HMM-based model was developed on lung signals from the Indian (Mumbai) dataset, where no delineations of individual lung sounds were available; instead, the model was trained to associate the overall signal with one of three disease categories – pneumonia, other respiratory and non-respiratory conditions. The model predictions were again fed in an RF classifier for disease classification. This approach proposed an alternative to lung sound analysis, moving away from conventional clinical labels (such as crackles), to capture the overall acoustic sequence and its association with disease. Annotations performed by 8 clinical experts were used as a benchmark; agreement between annotators was found to be poor, suggesting clinical interpretation of lung signals is an unreliable gold standard. The proposed approach was found to outperform classification with both vital signs and clinical annotations (Table 9.1).

Dataset	**Model features***	**Accuracy**		**Sensitivity**		**Specificity**	
		Mean	(95% CI)	**Mean**	(95% CI)	**Mean**	(95% CI)
I(M)	**Benchmark: IMCI**	**0.667**	-	**0.5**	-	**0.833**	-
I(M)	Vital signs	0.859	(0.827-0.864)	0.697	(0.689-0.705)	0.778	(0.771-0.797)
I(M)	Vital signs + clinical annotation of lung signals	0.668	(0.659-0.676)	0.903	(0.899-0.907)	0.786	(0.781-0.790)
I(M)	**DNN-HMM + Vital signs**	**0.813**	(0.789-0.837)	**0.845**	(0.828-0.863)	**0.829**	(0.820-0.839)

Table 9.1: SUMMARY – Comparison between different techniques for identification of pneumonia applied to the Indian (Mumbai), I(M), dataset: (1) IMCI guidelines; (2) RF model developed on vital signs alone; (3) RF model developed on the combination of vital signs and the majority vote of the clinical annotation of the lung signals; (4) RF model developed on the combination of vital signs and DNN-HMM analysis of lung sounds.

9.2. Relevance of findings to low-resource settings

Beyond the algorithmic advancements presented in this thesis, work was conducted in a manner that maximises the potential impact of the proposed solutions to low-resource settings. In particular, feature selection, as described in Chapter 6, was governed by both data considerations as well as the objectives of minimising the number of devices required for measurements and the current availability of such point-of-care tools. Moreover, the field study in India, as described in Chapter 4, was designed to capture a realistic dataset, reflecting diagnostic scenarios in both a resource-constrained hospital and a poor community living in an urban slum. Design challenges had to be overcome to ensure suitability of the tools to both the user and the local health system. Consequently, the dataset collected captured a number of challenges inherent to the use of mobile health solutions and point-of-care devices in low-resource settings: (1) noisy signals; (2) disease heterogeneity, both within the pneumonia class and amongst the group of other diseases; (3) clinical uncertainty at community level. Unlike most data-driven solutions which are developed on idealistic datasets, the algorithms developed in the thesis on the Indian (Mumbai) data offer a more realistic assessment of the potential of machine learning to deliver accurate and reliable diagnosis in resource-constrained settings. Moreover, these algorithms were found to be generalisable across other populations, as discussed in Chapter 7.

9.3. Validation in clinical practice and relevance to health policy

There are a number of important considerations that should be discussed in the context of translating algorithm-driven diagnostic innovation, such as the models described in this thesis, into clinical practice and policy. First, an acceptable benchmark for diagnostic performance (i.e. minimal sensitivity and specificity) should be established in the context of diagnostic alternatives. With IMCI as the current gold standard for community-based triage/diagnosis, the proposed algorithms should outperform its sensitivity and specificity. Based on this, 85% sensitivity and 80% specificity were selected as thresholds in the sample size calculations of the India (Mumbai) study presented in Chapter 4. Classification performance attained when vitals signs were combined with the DNN-HMM-derived acoustic markers was seen to satisfy these benchmarks. Validation of the proposed approach should be performed via an appropriate randomised control trial, where the algorithms should be administered in a setup similar to that of the India (Mumbai) study and the control group should receive triage/monitoring only via IMCI.

Additionally, the intervention group receiving the algorithms should include a two-stage comparative process: (Stage 1) Including a combination of IMCI & proposed algorithms to prioritise sensitivity (i.e. if either of these suggests illness, a child should be referred to a doctor); (Stage 2) After the algorithms are proven to consistently outperform IMCI, then they should be used to dictate decisions in the intervention group.

9.4. Future work

There are a number of limitations associated with the work presented in this thesis; a short description of each of them is given here, alongside suggested routes for future work that can help expand findings.

- **Dataset limitations**: although most of the datasets used in this project were fairly large, only one of them, the Indian (Mumbai) data contained the data richness required to develop generalisable data-driven solutions. Since these data were collected with a limited budget and a timeframe of 6 months, it size was insufficient to deliver all clinical insights required for this analysis. For example, many of the 'non-pneumonia' diseases were under-sampled, making the 'non-pneumonia' class both heterogeneous and sparse. Additionally, most datasets contain very few, if any, neonatal cases, making generalisation of algorithms to this age group difficult. Larger datasets are rare but do exist (e.g. PERCH study dataset mentioned in Chapter 1) and open access to them would enable the development and validation of generalisable models.
- **Performance variability as a function of signal quality**: The quality of signals generated by different point-of-care devices is variable and likely to result in differences in the symptomatic data captured. Some of the differences in vital signs across the datasets analysed in this project, such as temperature or oxygen saturation, might be attributed to the use of different study equipment. However, insufficient information prohibited the more in depth analysis of this issue. Future work should explore opportunities to delineate signal quality differences from population-specific physiological differences. Data generated by device-comparative studies, such as the one conducted by the Malaria Consortium (Malaria 2016), can be used to inform such analysis.

- **Modality limitations:** symptomatic data in the study stemmed from a few modalities, mostly classical vital signs and lung signals, which were deemed most appropriate for community-based diagnosis. Yet, recent advancements have been reported regarding the use of machine learning techniques for analysis of X-rays and identification of pneumonia (Rajpurkar 2017). Whilst such tools will continue to be relevant only to hospital settings, cross-modality analysis, linking acoustic biomarkers from a digital stethoscope with image markers from a chest radiograph could help to improve the diagnostic value of point-of-care tools such as the digital stethoscope and provide an evidence-based alternative for the use of X-rays. Similarly, the use of ultrasound for diagnosis of childhood pneumonia has been explored more recently, delivering promising results (Claes et al. 2017). The same type of cross-modality analysis could deliver valuable medical insights. Additionally, better tools for evaluation of malnutrition are required to enable stratification of malnutrition levels and the development of more specific pneumonia algorithms.

10. Appendix

Fleiss κ

Fleiss kappa (Fleiss κ) is defined as: $\kappa = \frac{p_o - p_c}{1 - p_c}$

First, look at two annotators and two classes/categories:

$p_o = \frac{n_{11} + n_{22}}{n}$ is the observed agreement

$p_c = \left(\frac{n_{1+}}{n}\right)\left(\frac{n_{+1}}{n}\right) + \left(\frac{n_{2+}}{n}\right)\left(\frac{n_{+2}}{n}\right)$ is the estimated chance agreement

where n is the total number of items; n_{11} = number of items both annotators assigned to class 1; n_{22} = number of items both annotators assigned to class 2; n_{1+}= number of items annotator 1 assigned to class 1; n_{2+}= number of items annotator 1 assigned to class 2; n_{+1}= number of items annotator 2 assigned to class 1; n_{+2}= number of items annotator 2 assigned to class 2.

Then for multiple annotator and multiple categories, using nominal weights $w_{kl} = \begin{cases} 0 & if\ k \neq l \\ 1 & if\ k = l \end{cases}$

The observed agreement is now defined as: $p_o = \frac{1}{n'}\sum_{i=1}^{n'}\sum_{k=1}^{q}\frac{r_{ik}(r_{ik}^{*}-1)}{r_i(r_i-1)}$, where $r_{ik}^{*} = \sum_{l=1}^{q} w_{kl} r_{il}$

The estimated chance agreement is now defined as: $p_c = \sum_{k,l}^{q} w_{kl}(\bar{p}_{+k}\bar{p}_{k+} - \frac{s_{kl}^2}{r})$,

where $\bar{p}_{+k} = \frac{1}{r}\sum_{g=1}^{r} p_{gk}$ and $s_{kl}^2 = \frac{1}{r-1}(\sum_{g=1}^{r} p_{gk}p_{gl} - r\,\bar{p}_{+k}\bar{p}_{+l})$ and $p_{gl} = \frac{n_{gk}}{n_g}$

Here q is the total number of classes and r is the total number of annotators; w_{kl} = the weight associated with two annotators assigning an item to categories *k* and *l*; r_{il} = the number of annotators that assigned item *i* to category *l*; n' = number of items labelled by two or more annotators; r_{ik} = number of annotators that assigned item i to category *k*; r_i = number of annotators that assigned item i to any category; n_g = number of items that annotator g assigned to any category; n_{gk} = number of items that annotator g assigned to category *k*.

GMM

Gaussian mixture models (GMM) are a popular tool used for speaker identification. In the context presented in this thesis, the distribution of features (extracted) from a lung signal frame is modelled by a weighted sum of N Gaussian densities: $p(\mathbf{x}|\lambda) = \sum_{i=1}^{N} w_i p_i(\mathbf{x})$, where **x** is a D-dimensional vector, λ is the class model (e.g. a sound class) and $p_i(\mathbf{x})$ are the component densities with $i = 1, \dots, N$.

Each component density is a Gaussian with:

$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{0.5}} \exp\{-0.5(\mathbf{x}-\mu_i)' \sum_i^{-1}(\mathbf{x}-\mu_i)\}$, where μ_i is the density mean and $\sum_i$ is the density covariance matrix.

The complete Gaussian mixture density is hence parameterised by $\lambda = \{w_i, \mu_i, \sum_i\}, \ i = 1, \dots, N$, i.e. the mean vectors, the covariance matrix and mixture weights of all the Gaussians.

Each class in a given problem can be modelled by a specific GMM, $\lambda_k = \lambda_1, \lambda_2, \dots, \lambda_K$,trained on a given set of data via the Expectation Maximisation (EM) algorithm (Bahoura and Pelletier 2004). When presented with new data in the test set, the objective is to find a model λ_k which maximises the *a posteriori* probability:

$\hat{k} = \arg \max_{1 \le k \le K} p(\lambda_k|\mathbf{X})$, where $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_M}\}$ represents a sequence of observations.

Bayes rule can be used to express $\hat{k}$ as:

$$\hat{k} = \arg \max_{1 \le k \le K} \frac{p(\mathbf{X}|\lambda_k)p(\lambda_k)}{p(\mathbf{X})}$$

Which can be further simplified to $\hat{k} = \arg \max_{1 \le k \le K} p(\mathbf{X}|\lambda_k)$, assuming equally likely classes, i.e. p(**X**) is the same for all classes.

DNN

Deep neural network (DNN) is a feed-forward artificial neural network with multiple hidden layers. A non-linear function is used to map the feature input from one layer unit to the next. In the application presented in this thesis, the traditional logistic function is used, where the output, **y**, is calculated as: $\mathbf{y} = \frac{1}{1+e^{-(b+\mathbf{xw})}}$, with **x** is the input feature, **w** the weights between connections, b denotes the bias. There are two methods to train a DNN, as described below (Li 2013).

- Unsupervised pre-training

Stacked Restricted Boltzmann Machine (RBM) is used to initialise the DNN. The RBM is an undirected graphical model with two constituent layers: (1) a layer of binary stochastic hidden units; (2) a layer of stochastic visible units (Hinton 2006). Every configuration of visible and hidden state vectors, **v** and **h** respectively, is used by the RBM to derive an associated energy via:

$E(\mathbf{v}, \mathbf{h}) = \sum_i \frac{(v_i - \mu_i)^2}{2\sigma_i^2} - \sum_j b_j h_j - \sum_{i,j} w_{ij} \frac{v_i}{\sigma_i} h_j$, where **w** is as defined above, b_j denotes the bias of the hidden unit h_j, and μ_i and σ_i are the mean and variance of the hidden variable v_i. Coupled with this, the free energy for that same configuration of **v** and **h** is defined as:

$F(\mathbf{v}) = -\log(\sum_h e^{-E(\mathbf{v},\mathbf{h})})$

Hence, a pre-training log likelihood can be defined as: $\xi(\theta) = -F(\mathbf{v}) - \log(\sum_v e^{-F(\mathbf{v})})$, where θ denotes the model parameters.

Next, the w_{ij} model parameters are updated iteratively according to:

$\Delta w_{ij}(t+1) = m\Delta w_{ij}(t) - \epsilon \frac{\partial \xi(\theta)}{\partial w_{ij}}$, where t denotes the iteration number, m is a momentum factor for smoothing of weight updates, ϵ is the learning rate of RBM, and

$\epsilon \frac{\partial \xi(\theta)}{\partial w_{ij}} = < v_i h_j >_{data} - < v_i h_j >_{recon}$, where the former term denotes the expectation of the training data and the latter the expectation of the reconstructed data.

At the end of the unsupervised pre-training, the resulting parameters are used as initial values of the DNN and back propagation (BP) is employed for fine-tuning of the network.

- Discriminative pre-training (supervised pre-training)

Supervised pre-training of DNN can be performed according to the next steps (Dahl 2012): (1) use BP to train a one-hidden layer DNN using every frame's state label to reach full convergence; (2) replace the *softmax* layer with another randomly initialised hidden layer and a new random *softmax* layer on top; (3) train the network to full convergence again; (4) repeat the process until the target number of hidden layers is reached.

11. References

Abbasi, S.; Derakhshanfar, R.; Abbasi, A.; Sarbaz, Y. 2013. "Classification of normal and abnormal lung sounds using neural network and support vector machines. ." *Proceedings of the 21st Iranian Conference on Electrical Engineering*:1-4.

Acharya, D., K. S. Prasanna, S. Nair, and R. S. Rao. 2003. "Acute respiratory infections in children: a community based longitudinal study in south India." *Indian J Public Health* 47 (1):7-13.

Addo-Yobo, E., N. Chisaka, M. Hassan, P. Hibberd, J. M. Lozano, P. Jeena, W. B. MacLeod, I. Maulen, A. Patel, S. Qazi, D. M. Thea, and N. T. Nguyen. 2004. "Oral amoxicillin versus injectable penicillin for severe pneumonia in children aged 3 to 59 months: a randomised multicentre equivalency study." *Lancet* 364 (9440):1141-8. doi: 10.1016/S0140-6736(04)17100-6.

Agrawal, P. K., S. Agrawal, S. Ahmed, G. L. Darmstadt, E. K. Williams, H. E. Rosen, V. Kumar, U. Kiran, R. C. Ahuja, V. K. Srivastava, M. Santosham, R. E. Black, and A. H. Baqui. 2012. "Effect of knowledge of community health workers on essential newborn health care: a study from rural India." *Health Policy Plan* 27 (2):115-26. doi: 10.1093/heapol/czr018.

Ahlstrom, C., P. Hult, P. Rask, J. E. Karlsson, E. Nylander, U. Dahlstrom, and P. Ask. 2006. "Feature extraction for systolic heart murmur classification." *Ann Biomed Eng* 34 (11):1666-77. doi: 10.1007/s10439-006-9187-4.

Ahmed, T., M. Ali, M. M. Ullah, I. A. Choudhury, M. E. Haque, M. A. Salam, G. H. Rabbani, R. M. Suskind, and G. J. Fuchs. 1999. "Mortality in severely malnourished children with diarrhoea and use of a standardised management protocol." *Lancet* 353 (9168):1919-22. doi: 10.1016/S0140-6736(98)07499-6.

Ali, A., A. R. Khowaja, M. Z. Bashir, F. Aziz, S. Mustafa, and A. Zaidi. 2013. "Role of human metapneumovirus, influenza A virus and respiratory syncytial virus in causing WHO-defined severe pneumonia in children in a developing country." *PLoS One* 8 (9):e74756. doi: 10.1371/journal.pone.0074756.

Allain, C., and M. Cloitre. 1991. "Characterizing the lacunarity of random and deterministic fractal sets." *Phys Rev A* 44 (6):3552-3558.

Alsmadi, S.; Kahya, Y. 2002. "Online classification of lung sounds using DSP." *Proceedings of the 2nd Joint Engineering in Medicine and Biology*:1771-1772.

Analysis, Swedish Agency for Growth Policy. 2013. "India's healthcare system - overview and quality improvements. ." *Stockholm*.

Ashraf, H., T. Ahmed, M. I. Hossain, N. H. Alam, R. Mahmud, S. M. Kamal, M. A. Salam, and G. J. Fuchs. 2007. "Day-care management of children with severe malnutrition in an urban health clinic in Dhaka, Bangladesh." *J Trop Pediatr* 53 (3):171-8. doi: 10.1093/tropej/fmm005.

Ashraf, H., S. A. Jahan, N. H. Alam, R. Mahmud, S. M. Kamal, M. A. Salam, and N. Gyr. 2008. "Day-care management of severe and very severe pneumonia, without associated co-morbidities such as severe malnutrition, in an urban health clinic in Dhaka, Bangladesh." *Arch Dis Child* 93 (6):490-4. doi: 10.1136/adc.2007.118877.

Ashraf, H., R. Mahmud, N. H. Alam, S. A. Jahan, S. M. Kamal, F. Haque, M. A. Salam, and N. Gyr. 2010. "Randomized controlled trial of day care versus hospital care of severe pneumonia in Bangladesh." *Pediatrics* 126 (4):e807-15. doi: 10.1542/peds.2009-3631.

Ayari, F.; Ksouri, M.; Alouani, A. 2012. "A new scheme for automatic classification of pathologic lung sounds." *Int. J. Comput. Sci.* 9:448-458.

Aydore, S., I. Sen, Y. P. Kahya, and M. Mihcak. 2009. "Classification of respiratory signals by linear analysis." *Conf Proc IEEE Eng Med Biol Soc* 2009:2617-20. doi: 10.1109/IEMBS.2009.5335395.

Bachur, R., H. Perry, and M. B. Harper. 1999. "Occult pneumonias: empiric chest radiographs in febrile children with leukocytosis." *Ann Emerg Med* 33 (2):166-73.

Bahoura, M. 2009. "Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes." *Comput Biol Med* 39 (9):824-43. doi: 10.1016/j.compbiomed.2009.06.011.

Bahoura, M., and C. Pelletier. 2004. "Respiratory sounds classification using cepstral analysis and Gaussian mixture models." *Conf Proc IEEE Eng Med Biol Soc* 1:9-12. doi: 10.1109/IEMBS.2004.1403077.

Bahoura, M.; Lu, X. 2006. "Separation of crackles from vesicular sounds using wavelet packet transform." *ICASSP* II (1076-1079).

Balarajan, Y., S. Selvaraj, and S. V. Subramanian. 2011. "Health care and equity in India." *Lancet* 377 (9764):505-15. doi: 10.1016/S0140-6736(10)61894-6.

Bang, A. T., R. A. Bang, S. B. Baitule, M. H. Reddy, and M. D. Deshmukh. 1999. "Effect of home-based neonatal care and management of sepsis on neonatal mortality: field trial in rural India." *Lancet* 354 (9194):1955-61. doi: 10.1016/S0140-6736(99)03046-9.

Bang, A. T., R. A. Bang, O. Tale, P. Sontakke, J. Solanki, R. Wargantiwar, and P. Kelzarkar. 1990. "Reduction in pneumonia mortality and total childhood mortality by means of community-based intervention trial in Gadchiroli, India." *Lancet* 336 (8709):201-6.

Bari, A., S. Sadruddin, A. Khan, I. Khan, I. A. Lehri, W. B. Macleod, M. P. Fox, D. M. Thea, and S. A. Qazi. 2011. "Community case management of severe pneumonia with oral amoxicillin in children aged 2-59 months in Haripur district, Pakistan: a cluster randomised trial." *Lancet* 378 (9805):1796-803. doi: 10.1016/S0140-6736(11)61140-9.

Berkley, J. A., K. Maitland, I. Mwangi, C. Ngetsa, S. Mwarumba, B. S. Lowe, C. R. Newton, K. Marsh, J. A. Scott, and M. English. 2005. "Use of clinical syndromes to target antibiotic prescribing in seriously ill children in malaria endemic area: observational study." *BMJ* 330 (7498):995. doi: 10.1136/bmj.38408.471991.8F.

Bhutta, Z. A., J. K. Das, N. Walker, A. Rizvi, H. Campbell, I. Rudan, R. E. Black, and Lancet Diarrhoea and Pneumonia Interventions Study Group. 2013. "Interventions to address deaths from childhood pneumonia and diarrhoea equitably: what works and at what cost?" *Lancet* 381 (9875):1417-29. doi: 10.1016/S0140-6736(13)60648-0.

Black, R. E., S. Cousens, H. L. Johnson, J. E. Lawn, I. Rudan, D. G. Bassani, P. Jha, H. Campbell, C. F. Walker, R. Cibulskis, T. Eisele, L. Liu, C. Mathers, W. H. O. Child Health Epidemiology Reference Group of, and Unicef. 2010. "Global, regional, and national causes of child mortality in 2008: a systematic analysis." *Lancet* 375 (9730):1969-87. doi: 10.1016/S0140-6736(10)60549-1.

Bleeker, S. E., K. G. Moons, G. Derksen-Lubsen, D. E. Grobbee, and H. A. Moll. 2001. "Predicting serious bacterial infection in young children with fever without apparent source." *Acta Paediatr* 90 (11):1226-32.

Breiman, L. 2001. "Statistical modelling: the two cultures." *Statistical Science* 16 (3):199–231.

Broor, S., S. Parveen, P. Bharaj, V. S. Prasad, K. N. Srinivasulu, K. M. Sumanth, S. K. Kapoor, K. Fowler, and W. M. Sullender. 2007. "A prospective three-year cohort study of the epidemiology and virology of acute respiratory infections of children in rural India." *PLoS One* 2 (6):e491. doi: 10.1371/journal.pone.0000491.

Bryce, J., E. Gouws, T. Adam, R. E. Black, J. A. Schellenberg, F. Manzi, C. G. Victora, and J. P. Habicht. 2005. "Improving quality and efficiency of facility-based child health care through Integrated Management of Childhood Illness in Tanzania." *Health Policy Plan* 20 Suppl 1:i69-i76. doi: 10.1093/heapol/czi053.

Bryce, J., C. G. Victora, and Mce-Imci Technical Advisors. 2005. "Ten methodological lessons from the multi-country evaluation of integrated Management of Childhood Illness." *Health Policy Plan* 20 Suppl 1:i94-i105. doi: 10.1093/heapol/czi056.

Burges, C. 1998. "A tutorial on support vector machines for pattern recognition. ." *Data Mining Knowledge Dis.* 2:121–167.

Campbell, H., Z. Biloglav, and I. Rudan. 2008. "Reducing bias from test misclassification in burden of disease studies: use of test to actual positive ratio--new test parameter." *Croat Med J* 49 (3):402-14.

Caulfield, L. E., M. de Onis, M. Blössner, and R. E. Black. 2004. "Undernutrition as an underlying cause of child deaths associated with diarrhea, pneumonia, malaria, and measles." *Am J Clin Nutr* 80 (1):193-8.

César, J. A., C. G. Victora, F. C. Barros, I. S. Santos, and J. A. Flores. 1999. "Impact of breast feeding on admission for pneumonia during postneonatal period in Brazil: nested case-control study." *BMJ* 318 (7194):1316-20.

Chang, C.; Lin, C. . 2011. "LIBSVM: A library for support vector machines." *ACM Transactions on Intelligent Systems and Technology* 2 (27):1–27.

Charbonneau, G.; Ademovic, E.; Cheetham, B.; Malmberg, L.; Vanderschoot, J.; Sovijarvi, A. 2000. "Basic techniques for respiratory sound analysis." *European Respiratory Review*.

Charleston-Villalobos, S., R. Gonzalez-Camarena, G. Chi-Lem, and T. Aljama-Corrales. 2007. "Crackle sounds analysis by empirical mode decomposition. Nonlinear and nonstationary signal analysis for distinction of crackles in lung sounds." *IEEE Eng Med Biol Mag* 26 (1):40-7.

Cherian, T., E. K. Mulholland, J. B. Carlin, H. Ostensen, R. Amin, M. de Campo, D. Greenberg, R. Lagos, M. Lucero, S. A. Madhi, K. L. O'Brien, S. Obaro, M. C. Steinhoff, and WHO Radiology Working Grp. 2005. "Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies." *Bulletin of the World Health Organization* 83 (5):353-359.

Chisti, M. J., M. Tebruegge, S. La Vincente, S. M. Graham, and T. Duke. 2009. "Pneumonia in severely malnourished children in developing countries - mortality risk, aetiology and validity of WHO clinical signs: a systematic review." *Trop Med Int Health* 14 (10):1173-89. doi: 10.1111/j.1365-3156.2009.02364.x.

Claes, A. S., P. Clapuyt, R. Menten, N. Michoux, and D. Dumitriu. 2017. "Performance of chest ultrasound in pediatric pneumonia." *European Journal of Radiology* 88:82-87. doi: 10.1016/j.ejrad.2016.12.032.

Clemmensen, L.; Hastie, T.; Witten, D.; Ersboll, B. 2011. "Sparse discriminant analysis." *Technometrics* 53 (4):406-413.

Commissioner, Registrar General and Census. 2014. "Annual health survey (AHS)." *Second bulletin*.

Consulting, Vital Wave. 2009. "Health Information Systems in developing countries: a landscape analysis. Research paper and strategic briefing.".

Craig, J. C., G. J. Williams, M. Jones, M. Codarini, P. Macaskill, A. Hayen, L. Irwig, D. A. Fitzgerald, D. Isaacs, and M. McCaskill. 2010. "The accuracy of clinical symptoms and signs for the diagnosis of serious bacterial infection in young febrile children: prospective cohort study of 15 781 febrile illnesses." *BMJ* 340:c1594.

Dahl, G.; Yu, D.; Deng, L.; Acero, A. 2012. "Context-dependent pretrained deep neural networks for large-vocabulary speech recognition." *Audio, Speech, and Language Processing, IEEE Transactions* 20 (1):30-42.

Dahl, G.; Yu, D.; L. Deng, L.; Acero, A. 2011. "Context-Dependent Pre-Trained Deep Neural Networks for Large Vocabulary Speech Recognition." *IEEE Trans. Speech and Audio Proc., Special Issue on Deep Learning for Speech and Language Processing*.

Donders, A. R., G. J. van der Heijden, T. Stijnen, and K. G. Moons. 2006. "Review: a gentle introduction to imputation of missing values." *J Clin Epidemiol* 59 (10):1087-91. doi: 10.1016/j.jclinepi.2006.01.014.

Dongre, A. R., P. R. Deshmukh, and B. S. Garg. 2008. "Perceptions and health care seeking about newborn danger signs among mothers in rural Wardha." *Indian J Pediatr* 75 (4):325-9. doi: 10.1007/s12098-008-0032-7.

Dorantes-Mendez, G., S. Charleston-Villalobos, R. Gonzalez-Camarena, G. Chi-Lem, J. G. Carrillo, and T. Aljama-Corrales. 2008. "Crackles detection using a time-variant autoregressive model." *Conf Proc IEEE Eng Med Biol Soc* 2008:1894-7. doi: 10.1109/IEMBS.2008.4649556.

Du, G.; Yeo, T. 2002. "A novel lacunarity estimation method applied to SAR image segmentation." *IEEE Transactions on Geoscience and Remote Sensing* 40 (12):2687 - 2691.

Du, M.; Chan, F.; Lam, F.; Sun, J. 1997. "Crackle detection and classification based on matched waveletanalysis." *The 19th IEEE Engineering in Medicine and Biology Society Conference Proceedings, Chicago, Illinois, USA,* 4:1638-1641.

Duke, T., and C. S. Mgone. 2003. "Measles: not just another viral exanthem." *Lancet* 361 (9359):763-73. doi: 10.1016/S0140-6736(03)12661-X.

Duke, T., J. Mgone, and D. Frank. 2001. "Hypoxaemia in children with severe pneumonia in Papua New Guinea." *Int J Tuberc Lung Dis* 5 (6):511-9.

Duncan, H., J. Hutchison, and C. S. Parshuram. 2006. "The Pediatric Early Warning System score: a severity of illness score to predict urgent medical need in hospitalized children." *J Crit Care* 21 (3):271-8. doi: 10.1016/j.jcrc.2006.06.007.

Ebell, M. H. 2010. "Clinical diagnosis of pneumonia in children." *Am Fam Physician* 82 (2):192-3.

Economist, The. June 2012. "Squeezing the doctor – The future of medicine."26.

Edejer, T. T., M. Aikins, R. Black, L. Wolfson, R. Hutubessy, and D. B. Evans. 2005. "Cost effectiveness analysis of strategies for child health in developing countries." *BMJ* 331 (7526):1177. doi: 10.1136/bmj.38652.550278.7C.

Egdell, P., L. Finlay, and D. K. Pedley. 2008. "The PAWS score: validation of an early warning scoring system for the initial assessment of children in the emergency department." *Emerg Med J* 25 (11):745-9. doi: 10.1136/emj.2007.054965.

Ellington, L. E., D. Emmanouilidou, M. Elhilali, R. H. Gilman, J. M. Tielsch, M. A. Chavez, J. Marin-Concha, D. Figueroa, J. West, and W. Checkley. 2014. "Developing a reference of normal lung sounds in healthy Peruvian children." *Lung* 192 (5):765-73. doi: 10.1007/s00408-014-9608-3.

Ellington, L. E., R. H. Gilman, J. M. Tielsch, M. Steinhoff, D. Figueroa, S. Rodriguez, B. Caffo, B. Tracey, M. Elhilali, J. West, and W. Checkley. 2012. "Computerised lung sound analysis to improve the specificity of paediatric pneumonia diagnosis in resource-poor settings: protocol and methods for an observational study." *BMJ Open* 2 (1):e000506. doi: 10.1136/bmjopen-2011-000506.

Emmanouilidou, D., K. Patil, J. West, and M. Elhilali. 2012. "A multiresolution analysis for detection of abnormal lung sounds." *Conf Proc IEEE Eng Med Biol Soc* 2012:3139-42. doi: 10.1109/EMBC.2012.6346630.

Emrani, S.; Krim, H. 2013. "Wheeze detection and location using spectro-temporal analysis of lung sounds." *Proceedings of the 29th Southern Biomedical Engineering Conference*:37-38.

Enarson, P., S. La Vincente, R. Gie, E. Maganga, and C. Chokani. 2008. "Implementation of an oxygen concentrator system in district hospital paediatric wards throughout Malawi." *Bull World Health Organ* 86 (5):344-8.

EXIST. 2014. "Registrar General of India. SRS Bulletin, Sample Registration System."

EXIST. 2015. "Ministry of Health and Family Welfare. Rural health statistics bulletin 2015. New Delhi: Government of India."

Falade, A. G., H. Tschäppeler, B. M. Greenwood, and E. K. Mulholland. 1995. "Use of simple clinical signs to predict pneumonia in young Gambian children: the influence of malnutrition." *Bull World Health Organ* 73 (3):299-304.

Flandrin, P.; Rilling, G.; Goncalves, P. 2004. "Empirical mode decomposition as a filter bank." *IEEE Signal Processing Letters*:112-114.

Fu, L. Y., R. Ruthazer, I. Wilson, A. Patel, L. M. Fox, T. A. Tuan, P. Jeena, N. Chisaka, M. Hassan, J. Lozano, I. Maulen-Radovan, D. M. Thea, S. Qazi, and P. Hibberd. 2006. "Brief hospitalization and pulse oximetry for predicting amoxicillin treatment failure in children with severe pneumonia." *Pediatrics* 118 (6):e1822-30. doi: 10.1542/peds.2005-2673.

Gakidou, E., S. Oza, C. Vidal Fuertes, A. Y. Li, D. K. Lee, A. Sousa, M. C. Hogan, S. Vander Hoorn, and M. Ezzati. 2007. "Improving child survival through environmental and nutritional interventions: the importance of targeting interventions toward the poor." *JAMA* 298 (16):1876-87. doi: 10.1001/jama.298.16.1876.

GBD. 2015. "Mortality and Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013." *Lancet* 385:117-71.

Gefen, Y.; Meir, Y.; Mandelbrot, B.; Aharony, A. . 1983. "Geometric implementation of hypercubic lattices with noninteger dimensionality by use of low lacunarity fractal lattices." *Phys. Rev. Lett.* 50 (3):145–148.

General, Office of Registrar. 2003. "Registrar General of India. Maternal mortality in India: 1997–2003 trends, causes and risk factors. ." *New Delhi*.

General, Office of Registrar. 2005. "Registrar General of India. SRS Bulletin. Sample Registration System." *New Delhi* 39 (1).

Ghimire, M., S. K. Bhattacharya, and J. P. Narain. 2012. "Pneumonia in South-East Asia Region: public health perspective." *Indian J Med Res* 135 (4):459-68.

Gladstone, B. P., J. P. Muliyil, S. Jaffar, J. G. Wheeler, A. Le Fevre, M. Iturriza-Gomara, J. J. Gray, A. Bose, M. K. Estes, D. W. Brown, and G. Kang. 2008. "Infant morbidity in an Indian slum birth cohort." *Arch Dis Child* 93 (6):479-84. doi: 10.1136/adc.2006.114546.

Gnitecki, J.; Moussavi, Z. 2005. "The fractality of lung sounds: A comparison of three waveform fractal dimension algorithms. ." *Chaos, Solitons Fractals* 26:1065-1072.

Gogia, S., and H. S. Sachdev. 2010. "Home visits by community health workers to prevent neonatal deaths in developing countries: a systematic review." *Bull World Health Organ* 88 (9):658-666B. doi: 10.2471/BLT.09.069369.

Gowraiah, V., S. Awasthi, R. Kapoor, D. Sahana, P. Venkatesh, B. Gangadhar, A. Awasthi, A. Verma, N. Pai, and M. Seear. 2014. "Can we distinguish pneumonia from wheezy diseases in tachypnoeic children under low-resource conditions? A prospective observational study in four Indian hospitals." *Arch Dis Child* 99 (10):899-906. doi: 10.1136/archdischild-2013-305740.

Graham, S. M., M. English, T. Hazir, P. Enarson, and T. Duke. 2008. "Challenges to improving case management of childhood pneumonia at health facilities in resource-limited settings." *Bull World Health Organ* 86 (5):349-55.

Gross, V.; Penzel, T.; Hadjileontiadis, L.; Koehler, U.; Vogelmeier, C. 2002. "Electronic auscultation based on wavelet transformation in clinical use." *IEEE, Proceedings of the Second Joint EMBS/BMES Conference*:1531–2.

Gudwani, A; Mitra, P; Puri, A; Vaidya, M. 2012. "India healthcare: inspiring possibilities, challenging journey." *McKinsey & Co*.

Guler, E. C., B. Sankur, Y. P. Kahya, and S. Raudys. 2005. "Two-stage classification of respiratory sound patterns." *Comput Biol Med* 35 (1):67-83. doi: 10.1016/j.compbiomed.2003.11.001.

Guler, I., H. Polat, and U. Ergun. 2005. "Combining neural network and genetic algorithm for prediction of lung sounds." *J Med Syst* 29 (3):217-31.

Guntupalli, K. K., P. M. Alapat, V. D. Bandi, and I. Kushnir. 2008. "Validation of automatic wheeze detection in patients with obstructed airways and in healthy subjects." *J Asthma* 45 (10):903-7. doi: 10.1080/02770900802386008.

Gurung, A., C. G. Scrafford, J. M. Tielsch, O. S. Levine, and W. Checkley. 2011. "Computerized lung sound analysis as diagnostic aid for the detection of abnormal lung sounds: a systematic review and meta-analysis." *Respir Med* 105 (9):1396-403. doi: 10.1016/j.rmed.2011.05.007.

Guyon, I.; Elisseeff, A. 2003. "An introduction to variable and feature selection." *J. Mach. Learn. Res.* 3:1157–1182.

Hadi, Abdullahel. 2003. "Management of acute respiratory infections by community health volunteers : experience of Bangladesh Rural Advancement Committee (BRAC)." *Bulletin of the World Health Organisation* 81 (3):183-9.

Hadjileontiadis, L. 2003. "Discrimination analysis of discontinuous breath sounds using higher‐order crossings." *Medical and Biological Engineering and Computing* 41:445-55.

Hadjileontiadis, L. J. 2007. "Empirical mode decomposition and fractal dimension filter. A novel technique for denoising explosive lung sounds." *IEEE Eng Med Biol Mag* 26 (1):30-9.

Hadjileontiadis, L. J. 2009. "A texture-based classification of crackles and squawks using lacunarity." *IEEE Trans Biomed Eng* 56 (3):718-32. doi: 10.1109/TBME.2008.2011747.

Haines, A., D. Sanders, U. Lehmann, A. K. Rowe, J. E. Lawn, S. Jan, D. G. Walker, and Z. Bhutta. 2007. "Achieving child survival goals: potential contribution of community health workers." *Lancet* 369 (9579):2121-31. doi: 10.1016/S0140-6736(07)60325-0.

Haines, C., M. Perrott, and P. Weir. 2006. "Promoting care for acutely ill children-development and evaluation of a paediatric early warning tool." *Intensive Crit Care Nurs* 22 (2):73-81. doi: 10.1016/j.iccn.2005.09.003.

Hashemi, A.; Arabalibiek, H.; Agin, K. . 2011. "Classification of wheeze sounds using wavelets and neural networks." *International Conference on Biomedical Engineering and Technology IPCBEE* 11.

Hastie, T.; Tibshirani, R.; Friedman, J. 2009. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Edited by Second: Springer.

Heiskanen-Kosma, T., M. Korppi, C. Jokinen, S. Kurki, L. Heiskanen, H. Juvonen, S. Kallinen, M. Sten, A. Tarkiainen, P. R. Ronnberg, M. Kleemola, P. H. Makela, and M. Leinonen. 1998. "Etiology of childhood pneumonia: serologic results of a prospective, population-based study." *Pediatr Infect Dis J* 17 (11):986-91.

Hinton, G.; Deng, L.; Yu, D.; Dahl, G.; Mohamed, A.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.; Kingsbury, B. 2012. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups." *IEEE Signal Processing Magazine* 29 (6).

Hinton, G.; Osindero, S.; Teh, W. 2006. "A fast learning algorithm for deep belief nets." *Neural computation* 18 (7):1527-1554.

HLEG. 2011. "High level expert group on Universal Health Coverage Report."120.

Hotelling, H. 1933. "Analysis of a complex of statistical variables into principal components." *J. Educ. Psychol* 24:417–441.

Huang, H., 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. 2014. "Discovery and validation of biomarkers to guide clinical management of pneumonia in African children." *Clin Infect Dis* 58 (12):1707-15. doi: 10.1093/cid/ciu202.

Huang, N.; Shen, Z.; Long, S.; Wu, M.; Shih, H.; Zheng, Q.; Yen, N.; Tung, C.; Liu, H. 1998 "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis." *Proc. R. Soc. Lond.* 454 903-995.

Icer, S.; Gengec, S. 2014. "Classification and analysis of non-stationary characteristics of crackle and rhonchus lung adventitious sounds." *Digital Signal Process* 28:18-27.

IHME. 2014. "Pushing the pace: progress and challenges in fighting childhood pneumonia." *Institute for Health Metrics and Evaluation* (Seattle).

India, Government of. 2014. "A Strategic Approach to Reproductive, Maternal, Newborn, Child and Adolescent Health (RMNCH+A) in India. Ministry of Health and Family Welfare." *New Delhi*.

India, Government of. 2015a. "Central Bureau of Health Intelligence. National health profile 2015. New Delhi: Ministry of Health & Family Welfare." *New Delhi*.

India, Government of. 2015b. "Ministry of Women and Child Development. Rapid survey on children 2013–14. India fact sheet. ." *New Delhi*.

India, Government of. 2015c. "National Sample Survey Office. Key indicators of social consumption in India health (January–June 2014). ." *New Delhi*.

India, New Delhi: Government of. 2015d. "Ministry of Health and Family Welfare. Rural health statistics bulletin 2015. ."

India, Planning Commission of. 2011. "High Level Expert Group Report on universal health coverage for India. ." *New Delhi: Government of India*.

India, Registrar General of. 2015e. "Maternal mortality ratio bulletin 2011–13. New Delhi: Office of Registrar General."

Iyer, V. K., P. A. Ramamoorthy, and Y. Ploysongsang. 1989. "Autoregressive modeling of lung sounds: characterization of source and transmission." *IEEE Trans Biomed Eng* 36 (11):1133-7. doi: 10.1109/10.40821.

Jain, A., and J. Vepa. 2008. "Lung sound analysis for wheeze episode detection." *Conf Proc IEEE Eng Med Biol Soc* 2008:2582-5. doi: 10.1109/IEMBS.2008.4649728.

Jain, D. L., V. Sarathi, and S. Jawalekar. 2013. "Predictors of treatment failure in hospitalized children [3-59 months] with severe and very severe pneumonia." *Indian Pediatr* 50 (8):787-9.

Jane, R., S. Cortes, J. A. Fiz, and J. Morera. 2004. "Analysis of wheezes in asthmatic patients during spontaneous respiration." *Conf Proc IEEE Eng Med Biol Soc* 5:3836-9. doi: 10.1109/IEMBS.2004.1404074.

Jin, F.; Sattar, F.; Goh, D. 2008. "Automatic wheeze detection using histograms of sample entropy. ." *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society,*:1890-1893.

Jin, F.; Sattar, F.; Goh, D. 2014. "New approaches for spectro-temporal feature extraction with applications to respiratory sound classification." *Neurocomputing* 123.

Joachims, T. . 1998. "Text categorization with Support Vector Machines: Learning with many relevant features." *Machine Learning: ECML-98. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), Springer, Berlin, Heidelberg* 1398.

Juven, T., J. Mertsola, M. Waris, M. Leinonen, O. Meurman, M. Roivainen, J. Eskola, P. Saikku, and O. Ruuskanen. 2000. "Etiology of community-acquired pneumonia in 254 hospitalized children." *Pediatr Infect Dis J* 19 (4):293-8.

Kahya, Y. P., M. Yeginer, and B. Bilgic. 2006. "Classifying respiratory sounds with different feature sets." *Conf Proc IEEE Eng Med Biol Soc* 1:2856-9. doi: 10.1109/IEMBS.2006.259946.

Kahya, Y.; Guler, E.; Sankur, B. 1999. "Statistical analysis of lung sound data." *Proceedings of the 21st Annual Conference and the 1999 Annual Fall Metering of the Biomedical Engineering Society*:1015-1015.

Kandaswamy, A.; Kumar, C.; Ramanathan, R.; Jayaraman, S.; Malmurugan, N. 2004. "Neural classification of lung sounds using wavelet coefficients." *Comput. Biol. Med.* 34:523-537.

Kayha, Y.; Yerer, S.; Cerid, O. . 2001. "A Wavelet-based instrument for detection of crackles in pulmonary sounds." *Engineering in Medicine and Biology Society IEEE-EMBS* (4):3175–8.

Kim, T.; White, H. 2004. "On more robust estimation of skewness and kurtosis." *Finance Research Letters* 1 (1):56-73.

Kumar, V., S. Mohanty, A. Kumar, R. P. Misra, M. Santosham, S. Awasthi, A. H. Baqui, P. Singh, V. Singh, R. C. Ahuja, J. V. Singh, G. K. Malik, S. Ahmed, R. E. Black, M. Bhandari, G. L. Darmstadt, and Group Saksham Study. 2008. "Effect of community-based behaviour change management on neonatal mortality in Shivgarh, Uttar Pradesh, India: a cluster-randomised controlled trial." *Lancet* 372 (9644):1151-62. doi: 10.1016/S0140-6736(08)61483-X.

Lanata, C. F., I. Rudan, C. Boschi-Pinto, L. Tomaskovic, T. Cherian, M. Weber, and H. Campbell. 2004. "Methodological and quality issues in epidemiological studies of acute lower respiratory infections in children in developing countries." *Int J Epidemiol* 33 (6):1362-72. doi: 10.1093/ije/dyh229.

Landis, J. R., and G. G. Koch. 1977. "The measurement of observer agreement for categorical data." *Biometrics* 33 (1):159-74.

Leventhal, J. M. 1982. "Clinical predictors of pneumonia as a guide to ordering chest roentgenograms." *Clin Pediatr (Phila)* 21 (12):730-4. doi: 10.1177/000992288202101205.

Levine, O. S., K. L. O'Brien, M. Deloria-Knoll, D. R. Murdoch, D. R. Feikin, A. N. DeLuca, A. J. Driscoll, H. C. Baggett, W. A. Brooks, S. R. Howie, K. L. Kotloff, S. A. Madhi, S. A. Maloney, S. Sow, D. M. Thea, and J. A. Scott. 2012. "The Pneumonia Etiology Research for Child Health Project: a 21st century childhood pneumonia etiology study." *Clin Infect Dis* 54 Suppl 2:S93-101. doi: 10.1093/cid/cir1052.

Li, L.; Xu, W.; Hong, Q.; Tong, F.; Wu, J. 2017. "Classification between normal and adventitious lung sounds using deep neural network." *2016 10th International Symposium on Chinese Spoken Language Processing*.

Li, L.; Zhao, Y.; Jiang, D.; Zhang, Y. 2013. "Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition." *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* 312-17.

Lin, B.; Wu, H.; Chong, F.; Chen, S. 2006. "Wheeze recognition based on 2D bilateral filtering of spectrogram." *Biomed. Eng. Applic., Basis Commun.* 18:29-38.

Liu, H.; Yu, L. 2005. "Toward integrating feature selection algorithms for classification and clustering." *IEEE Trans. Knowl. Data Eng.* 17:491–502.

Liu, L., S. Oza, D. Hogan, Y. Chu, J. Perin, J. Zhu, J. E. Lawn, S. Cousens, C. Mathers, and R. E. Black. 2016. "Global, regional, and national causes of under-5 mortality in 2000-15: an updated systematic analysis with implications for the Sustainable Development Goals." *Lancet* 388 (10063):3027-3035. doi: 10.1016/S0140-6736(16)31593-8.

Liu, L., S. Oza, D. Hogan, J. Perin, I. Rudan, J. E. Lawn, S. Cousens, C. Mathers, and R. E. Black. 2015. "Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: an updated systematic analysis." *Lancet* 385 (9966):430-40. doi: 10.1016/S0140-6736(14)61698-6.

Loeb, M., and K. High. 2005. "The effect of malnutrition on risk and outcome of community-acquired pneumonia." *Respir Care Clin N Am* 11 (1):99-108. doi: 10.1016/j.rcc.2004.10.002.

Lozano, J. M. 2001. "Epidemiology of hypoxaemia in children with acute lower respiratory infection." *Int J Tuberc Lung Dis* 5 (6):496-504.

Lu, X.; Bahoura, M. 2007. "An automatic system for crackles detection and classification." *Canadian Conference on Electrical and Computer Engineering, 2006. CCECE '06.* .

Lynch, T., R. Platt, S. Gouin, C. Larson, and Y. Patenaude. 2004. "Can we predict which children with clinically suspected pneumonia will have the presence of focal infiltrates on chest radiographs?" *Pediatrics* 113 (3 Pt 1):e186-9.

Madhi, S. A., O. S. Levine, R. Hajjeh, O. D. Mansoor, and T. Cherian. 2008. "Vaccines to prevent pneumonia and improve child survival." *Bull World Health Organ* 86 (5):365-72.

Maguire, J. L., D. M. Kulik, A. Laupacis, N. Kuppermann, E. M. Uleryk, and P. C. Parkin. 2011. "Clinical prediction rules for children: a systematic review." *Pediatrics* 128 (3):e666-77. doi: 10.1542/peds.2011-0043.

Mahabee-Gittens, E. M., J. Grupp-Phelan, A. S. Brody, L. F. Donnelly, S. E. Bracey, E. M. Duma, M. L. Mallory, and G. B. Slap. 2005. "Identifying children with pneumonia in the emergency department." *Clin Pediatr (Phila)* 44 (5):427-35. doi: 10.1177/000992280504400508.

Malaria, Consortium;. 2016. "Malaria Consortium: Pneumonia Diagnostics Project."

Marsh, D. R., K. E. Gilroy, R. Van de Weerdt, E. Wansi, and S. Qazi. 2008. "Community case management of pneumonia: at a tipping point?" *Bull World Health Organ* 86 (5):381-9.

Marten, R., D. McIntyre, C. Travassos, S. Shishkin, W. Longde, S. Reddy, and J. Vega. 2014. "An assessment of progress towards universal health coverage in Brazil, Russia, India, China, and South Africa (BRICS)." *Lancet* 384 (9960):2164-71. doi: 10.1016/S0140-6736(14)60075-1.

Martinez-Hernandez, H. G., C. T. Aljama-Corrales, R. Gonzalez-Camarena, V. S. Charleston-Villalobos, and G. Chi-Lem. 2005. "Computerized classification of normal and abnormal lung sounds by multivariate linear autoregressive model." *Conf Proc IEEE Eng Med Biol Soc* 6:5999-6002. doi: 10.1109/IEMBS.2005.1615858.

Martinez, A. W., S. T. Phillips, G. M. Whitesides, and E. Carrilho. 2010. "Diagnostics for the Developing World: Microfluidic Paper-Based Analytical Devices." *Analytical Chemistry* 82 (1):3-10. doi: 10.1021/ac9013989.

Mathew, J. L., A. K. Patwari, P. Gupta, D. Shah, T. Gera, S. Gogia, P. Mohan, R. Panda, and S. Menon. 2011. "Acute respiratory infection and pneumonia in India: a systematic review of literature for advocacy and action: UNICEF-PHFI series on newborn and child health, India." *Indian Pediatr* 48 (3):191-218.

Mathews, B., S. Shah, R. H. Cleveland, E. Y. Lee, R. G. Bachur, and M. I. Neuman. 2009. "Clinical predictors of pneumonia among children with wheezing." *Pediatrics* 124 (1):e29-36. doi: 10.1542/peds.2008-2062.

May, C., K. Roth, and P. Panda. 2014. "Non-degree allopathic practitioners as first contact points for acute illness episodes: insights from a qualitative study in rural northern India." *BMC Health Serv Res* 14:182. doi: 10.1186/1472-6963-14-182.

Mayorga, P., C. Druzgalski, O. H. Gonzalez, and H. S. Lopez. 2012. "Modified classification of normal lung sounds applying Quantile vectors." *Conf Proc IEEE Eng Med Biol Soc* 2012:4262-5. doi: 10.1109/EMBC.2012.6346908.

Mayorga, P.; Druzgalski, C.; González, O.; Zazueta, A.; Criollo, M. 2011. "Expanded quantitative models for assessment of respiratory diseases and monitoring." *Pan American Health Care Exchanges (PAHCE)*.

Mayorga, P.; Druzgalski, C.; Morelos, R.; Gonzalez, O.; Vidales, J. 2010. "Acoustics Based Assessment of Respiratory Diseases using GMM Classification." *Conf Proc IEEE Eng Med Biol Soc.* :6312-6.

Mazic, I.; Sovilj, S.; Magjarevic, R. 2003. "Analysis of respiratory sounds in asthmatic infants." *Measure. Sci. Rev.* (3):9-12.

McCracken, G. H. 2000a. "Diagnosis and management of pneumonia in children." *Pediatric Infectious Disease Journal* 19 (9):924-928.

McCracken, G. H. 2000b. "Etiology and treatment of pneumonia." *Pediatric Infectious Disease Journal* 19 (4):373-377. doi: Doi 10.1097/00006454-200004000-00032.

McNally, L. M., P. M. Jeena, K. Gajee, S. A. Thula, A. W. Sturm, S. Cassol, A. M. Tomkins, H. M. Coovadia, and D. Goldblatt. 2007. "Effect of age, polymicrobial disease, and maternal HIV status on treatment response and cause of severe pneumonia in South African children: a prospective descriptive study." *Lancet* 369 (9571):1440-51. doi: 10.1016/S0140-6736(07)60670-9.

Meslier, N., G. Charbonneau, and J. L. Racineux. 1995. "Wheezes." *Eur Respir J* 8 (11):1942-8.

Mission, New Delhi: National Rural Health. 2005. "National Rural Health Mission. Accredited Social Health Activist guidelines. ."

Monaghan, A. 2005. "Detecting and managing deterioration in children." *Paediatr Nurs* 17 (1):32-5. doi: 10.7748/paed2005.02.17.1.32.c964.

Mondal, A.; Bhattacharya, P.; Saha, G. 2014. "Detection of Lungs Status Using Morphological Complexities of Respiratory Sounds." *The Scientific World Journal* 2014:9.

Murphy, C. G., A. C. van de Pol, M. B. Harper, and R. G. Bachur. 2007. "Clinical predictors of occult pneumonia in the febrile child." *Acad Emerg Med* 14 (3):243-9. doi: 10.1197/j.aem.2006.08.022.

Murphy, R. L., Jr., E. A. Del Bono, and F. Davidson. 1989. "Validation of an automatic crackle (rale) counter." *Am Rev Respir Dis* 140 (4):1017-20. doi: 10.1164/ajrccm/140.4.1017.

Murphy, R. L., A. Vyshedskiy, V. A. Power-Charnitsky, D. S. Bana, P. M. Marinelli, A. Wong-Tse, and R. Paciej. 2004. "Automated lung sound analysis in patients with pneumonia." *Respir Care* 49 (12):1490-7.

Nair, H., E. A. Simões, I. Rudan, B. D. Gessner, E. Azziz-Baumgartner, J. S. Zhang, D. R. Feikin, G. A. Mackenzie, J. C. Moïsi, A. Roca, H. C. Baggett, S. M. Zaman, R. J. Singleton, M. G. Lucero, A. Chandran, A. Gentile, C. Cohen, A. Krishnan, Z. A. Bhutta, A. Arguedas, A. W. Clara, A. L. Andrade, M. Ope, R. O. Ruvinsky, M. Hortal, J. P. McCracken, S. A. Madhi, N. Bruce, S. A. Qazi, S. S. Morris, S. El Arifeen, M. W. Weber, J. A. Scott, W. A. Brooks, R. F. Breiman, H. Campbell, and Severe Acute Lower Respiratory Infections Working Group. 2013. "Global and regional burden of hospital admissions for severe acute lower respiratory infections in young children in 2010: a systematic analysis." *Lancet* 381 (9875):1380-90. doi: 10.1016/S0140-6736(12)61901-1.

Nannan, N., R. Norman, M. Hendricks, M. A. Dhansay, D. Bradshaw, and South African Comparative Risk Assessment Collaborating Group. 2007. "Estimating the burden of disease attributable

to childhood and maternal undernutrition in South Africa in 2000." *S Afr Med J* 97 (8 Pt 2):733-9.

Neuman, M. I., M. C. Monuteaux, K. J. Scully, and R. G. Bachur. 2011. "Prediction of pneumonia in a pediatric emergency department." *Pediatrics* 128 (2):246-53. doi: 10.1542/peds.2010-3367.

NICE. 2013. "Fever in under 5s: assessment and initial management." *National Institute for Health and Care Excellence*.

Nijman, R. G., Y. Vergouwe, M. Thompson, M. van Veen, A. H. J. van Meurs, J. van der Lei, E. W. Steyerberg, H. A. Moll, and R. Oostenbrink. 2013. "Clinical prediction model to aid emergency doctors managing febrile children at risk of serious bacterial infections: diagnostic study." *Bmj-British Medical Journal* 346. doi: ARTN f1706

10.1136/bmj.f1706.

Nsona, H., A. Mtimuni, B. Daelmans, J. A. Callaghan-Koru, K. Gilroy, L. Mgalula, T. Kachule, and T. Zamasiya. 2012. "Scaling up integrated community case management of childhood illness: update from Malawi." *Am J Trop Med Hyg* 87 (5 Suppl):54-60. doi: 10.4269/ajtmh.2012.11-0759.

Ono, H., Y. Taniguchi, K. Shinoda, T. Sakamoto, S. Kudoh, and A. Gemma. 2009. "Evaluation of the usefulness of spectral analysis of inspiratory lung sounds recorded with phonopneumography in patients with interstitial pneumonia." *J Nippon Med Sch* 76 (2):67-75.

Oud, M., E. H. Dooijes, and J. S. van der Zee. 2000. "Asthmatic airways obstruction assessment based on detailed analysis of respiratory sound spectra." *IEEE Trans Biomed Eng* 47 (11):1450-5. doi: 10.1109/10.880096.

Palaniappan, R.; Sundaraj, K. 2013a. "Respiratory sound classification using cepstral features and support vector machine." *IEEE Recent Advances in Intelligent Computational Systems (RAICS)*.

Palaniappan, R.; Sundaraj, K.; Ahamed, N. 2013b. "Machine learning in lung sound analysis: A systematic review." *Biocybernetics and Biomedical Engineering* 33 (3):129–135.

Palaniappan, R.; Sundaraj, K.; Ahamed, N.; Arjunan, A.; Sundaraj, S. 2013c. "Computer-based respiratory sound analysis: a systematic review." *IETE Technical Review* 30 (3):248-256.

Pandey, M. R., J. S. Boleij, K. R. Smith, and E. M. Wafula. 1989. "Indoor air pollution in developing countries and acute respiratory infection in children." *Lancet* 1 (8635):427-9.

Pasterkamp, H., S. S. Kraman, P. D. DeFrain, and G. R. Wodicka. 1993. "Measurement of respiratory acoustical signals. Comparison of sensors." *Chest* 104 (5):1518-25.

Pasterkamp, H., S. S. Kraman, and G. R. Wodicka. 1997. "Respiratory sounds. Advances beyond the stethoscope." *Am J Respir Crit Care Med* 156 (3 Pt 1):974-87. doi: 10.1164/ajrccm.156.3.9701115.

Patel, V.; Parikh, R.; Nandraj,S.; Balasubramaniam,P.; Narayan,K.; Paul,V.K.; Kumar,A.K.; Chatterjee,M.; Reddy,K.S. 2015. "Assuring health coverage for all in India." *Lancet* 386 (10011):2422-2435.

Paul, V. K., H. S. Sachdev, D. Mavalankar, P. Ramachandran, M. J. Sankar, N. Bhandari, V. Sreenivas, T. Sundararaman, D. Govil, D. Osrin, and B. Kirkwood. 2011. "Reproductive health, and child health and nutrition in India: meeting the challenge." *Lancet* 377 (9762):332-49. doi: 10.1016/S0140-6736(10)61492-4.

Pépin, J., A. M. Demers, F. Mberyo-Yaah, S. Jaffar, C. Blais, P. Somsé, G. Bobossi, and P. Morency. 2001. "Acute lower respiratory infections among children hospitalized in Bangui, Central African Republic: toward a new case-management algorithm." *Trans R Soc Trop Med Hyg* 95 (4):410-7.

Phung, D.; Tran, D.; Ma, W.; Nguyen, P.; Pham, T. 2014. "Using Shannon entropy as EEG signal feature for fast person identification." *ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.

Piirila, P., and A. R. Sovijarvi. 1995. "Crackles: recording, analysis and clinical significance." *Eur Respir J* 8 (12):2139-48.

Plotnick, R. E., R. H. Gardner, W. W. Hargrove, K. Prestegaard, and M. Perlmutter. 1996. "Lacunarity analysis: A general technique for the analysis of spatial patterns." *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 53 (5):5461-5468.

Powell-Jackson T, Acharya A, Mills A. 2013. "An assessment of the quality of primary health care in India." *Econ Polit Wkly* 48:53-61.

Qazi, S., S. Aboubaker, R. MacLean, O. Fontaine, C. Mantel, T. Goodman, M. Young, P. Henderson, and T. Cherian. 2015. "Ending preventable child deaths from pneumonia and diarrhoea by 2025. Development of the integrated Global Action Plan for the Prevention and Control of Pneumonia and Diarrhoea." *Arch Dis Child* 100 Suppl 1:S23-8. doi: 10.1136/archdischild-2013-305429.

Qiu, Y., A. R. Whittaker, M. Lucas, and K. Anderson. 2005. "Automatic wheeze detection based on auditory modelling." *Proc Inst Mech Eng H* 219 (3):219-27. doi: 10.1243/095441105X28551.

Rabiner, L. 1989. "A tutorial on hidden markov models and selected applications in speech recognition." *Proceedings of the IEEE* 77 (2):257–286.

Rabiner, L.; Biing-Hwang, J. 1993. *Fundamentals of Speech Recognition*: Prentice Hall.

Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; Lungren, M.; Ng, A. 2017. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." *Cornel University Library*.

Rasmussen, Z., A. Pio, and P. Enarson. 2000. "Case management of childhood pneumonia in developing countries: recent relevant research and current initiatives." *Int J Tuberc Lung Dis* 4 (9):807-26.

Reddaiah, V. P., and S. K. Kapoor. 1991. "Effectiveness of ARI control strategy on underfive mortality." *Indian J Pediatr* 58 (1):123-30.

Reddy, K. S., V. Patel, P. Jha, V. K. Paul, A. K. Kumar, L. Dandona, and Healthcare Lancet India Group for Universal. 2011. "Towards achievement of universal health care in India by 2020: a call to action." *Lancet* 377 (9767):760-8. doi: 10.1016/S0140-6736(10)61960-5.

Reichert, S., R. Gass, C. Brandt, and E. Andres. 2008. "Analysis of respiratory sounds: state of the art." *Clin Med Circ Respirat Pulm Med* 2:45-58.

Rice, A. L., L. Sacco, A. Hyder, and R. E. Black. 2000. "Malnutrition as an underlying cause of childhood deaths associated with infectious diseases in developing countries." *Bull World Health Organ* 78 (10):1207-21.

Riella, R. J., P. Nohama, and J. M. Maia. 2009. "Method for automatic detection of wheezing in lung sounds." *Braz J Med Biol Res* 42 (7):674-84.

Rizal, A.; Hidayat, R.; Nugroho, H. 2015. "Signal Domain in Respiratory Sound Analysis: Methods, Application and Future Development." *Journal of Computer Science* 11 (10):1005-1016.

Roth, D. E., L. E. Caulfield, M. Ezzati, and R. E. Black. 2008. "Acute lower respiratory infections in childhood: opportunities for reducing the global burden through nutritional interventions." *Bull World Health Organ* 86 (5):356-64.

Rudan, I., C. Boschi-Pinto, Z. Biloglav, K. Mulholland, and H. Campbell. 2008. "Epidemiology and etiology of childhood pneumonia." *Bull World Health Organ* 86 (5):408-16.

Rudan, I., K. L. O'Brien, H. Nair, L. Liu, E. Theodoratou, S. Qazi, I. Lukšić, C. L. Fischer Walker, R. E. Black, H. Campbell, and Child Health Epidemiology Reference Group (CHERG). 2013. "Epidemiology and etiology of childhood pneumonia in 2010: estimates of incidence, severe morbidity, mortality, underlying risk factors and causative pathogens for 192 countries." *J Glob Health* 3 (1):010401. doi: 10.7189/jogh.03.010401.

Rudan, I., L. Tomaskovic, C. Boschi-Pinto, H. Campbell, and WHO Child Health Epidemiology Reference Group. 2004. "Global estimate of the incidence of clinical pneumonia among children under five years of age." *Bull World Health Organ* 82 (12):895-903. doi: /S0042-96862004001200005.

Ruuskanen, O., and J. Mertsola. 1999. "Childhood community-acquired pneumonia." *Semin Respir Infect* 14 (2):163-72.
Sanchez Morillo, D., S. Astorga Moreno, M. A. Fernandez Granero, and A. Leon Jimenez. 2013. "Computerized analysis of respiratory sounds during COPD exacerbations." *Comput Biol Med* 43 (7):914-21. doi: 10.1016/j.compbiomed.2013.03.011.
Sankur, B., Y. P. Kahya, E. C. Guler, and T. Engin. 1994. "Comparison of AR-based algorithms for respiratory sounds classification." *Comput Biol Med* 24 (1):67-76.
Sazawal, S., R. E. Black, and Pneumonia Case Management Trials Group. 2003. "Effect of pneumonia case management on mortality in neonates, infants, and preschool children: a meta-analysis of community-based trials." *Lancet Infect Dis* 3 (9):547-56.
School, Harvard Business. 2009. "Hospital pioneers new model."
Sciences, Mumbai: International Institute for Population. 2007. "International Institute for Population Sciences (IIPS) and Macro International. National Family Health Survey (NFHS-3), 2005–06, volume 1. ."
Scott, J. A., C. Wonodi, J. C. Moisi, M. Deloria-Knoll, A. N. DeLuca, R. A. Karron, N. Bhat, D. R. Murdoch, J. Crawley, O. S. Levine, K. L. O'Brien, D. R. Feikin, and Group Pneumonia Methods Working. 2012. "The definition of pneumonia, the assessment of severity, and clinical standardization in the Pneumonia Etiology Research for Child Health study." *Clin Infect Dis* 54 Suppl 2:S109-16. doi: 10.1093/cid/cir1065.
Seide, F.; Li, G.; Chen, X.; Yu, D. 2011. "Feature Engineering in Context-Dependent Deep Neural Networks for Conversational Speech Transcription." *Proc. ASRU*:24-29.
Serbes, G.; Sakar, C.; Kahya, Y.; Aydin, N. 2013. "Pulmonary crackle detection using time-frequency and time-scale analysis." *Digital Signal Process A Rev. J.* 23:1012-1021.
Shah, S., B. Mathews, M. I. Neuman, and R. Bachur. 2010. "Detection of occult pneumonia in a pediatric emergency department." *Pediatr Emerg Care* 26 (9):615-21. doi: 10.1097/PEC.0b013e3181ef04a3.
Shah, S. N., R. G. Bachur, D. L. Simel, and M. I. Neuman. 2017. "Does this child have pneumonia? the Rational Clinical Examination systematic review (vol 318, pg 462, 2017)." *Jama-Journal of the American Medical Association* 318 (13):1284-1284. doi: 10.1001/jama.2017.13624.
Shann, F. 1995. "The management of pneumonia in children in developing countries." *Clin Infect Dis* 21 Suppl 3:S218-25.
Shepherd-Smith, A. 2012. "Free drugs for India's poor." *Lancet* 380 (9845):874.
Shimeles, D., and S. Lulseged. 1994. "Clinical profile and pattern of infection in Ethiopian children with severe protein-energy malnutrition." *East Afr Med J* 71 (4):264-7.
Sjostrand, K.; Clemmensen, L.; Larsen, R.; Ersboll, B. "SpaSM: a Matlab toolbox for sparse statistical modeling."
Smith, K. R., J. P. McCracken, M. W. Weber, A. Hubbard, A. Jenny, L. M. Thompson, J. Balmes, A. Diaz, B. Arana, and N. Bruce. 2011. "Effect of reduction in household air pollution on childhood pneumonia in Guatemala (RESPIRE): a randomised controlled trial." *Lancet* 378 (9804):1717-26. doi: 10.1016/S0140-6736(11)60921-5.
Sovijarvi, A.; Dalmasso, F.; Vanderschoot, J.; Malmberg, L.; Righini, G.; Stoneman, S. 2000a. "Definition of terms for applications of respiratory sounds." *Eur Respir Rev* 77:597-610.
Sovijärvi, A.; Vanderschoot, J.; Earis, J. 2000. "Standardization of computerized respiratory sound analysis." *European Respiratory Review* 10:585.
Sovijarvi, A.R.; Malmberg, L.P.; Charbonneau, G.; Vandershoot, J. 2000b. "Characteristics of breath sounds and adventitious respiratory sounds." *European Respiratory Rev* 10:591-6.
Stoppiglia, H.; Dreyfus, G.; Dubois, R.; Y. Oussar, Y. 2003. "Ranking a random feature for variable and feature selection." *Journal of Machine Learning Research* 3:1399–1414.
Sudhinaraset, M., M. Ingram, H. K. Lofthouse, and D. Montagu. 2013. "What is the role of informal healthcare providers in developing countries? A systematic review." *PLoS One* 8 (2):e54978. doi: 10.1371/journal.pone.0054978.

Sundararaman, T. 2007. "Community health-workers: scaling up programmes." *Lancet* 369 (9579):2058-9. doi: 10.1016/S0140-6736(07)60326-2.

Suskind, R. M. 1988. "Malnutrition and the immune response." *Beitr Infusionther Klin Ernahr* 19:1-25.

Taplidou, S. A., and L. J. Hadjileontiadis. 2007. "Wheeze detection based on time-frequency analysis of breath sounds." *Comput Biol Med* 37 (8):1073-83. doi: 10.1016/j.compbiomed.2006.09.007.

Taplidou, S. A., L. J. Hadjileontiadis, I. K. Kitsas, K. I. Panoulas, T. Penzel, V. Gross, and S. M. Panas. 2004. "On applying continuous wavelet transform in wheeze analysis." *Conf Proc IEEE Eng Med Biol Soc* 5:3832-5. doi: 10.1109/IEMBS.2004.1404073.

Taplidou, S.; Hadjileontiadis, L.; Penzel, P.; Gross, V. 2003. "WED: An efficient wheezing-episode detector based on breath sounds spectrogram analysis." *Proceedings of the 25th Annual International Conference of the IEEE Date of Conference Engineering in Medicine and Biology Society*:2531-2534.

Thompson, M., N. Coad, A. Harnden, R. Mayon-White, R. Perera, and D. Mant. 2009. "How well do vital signs identify children with serious infections in paediatric emergency care?" *Arch Dis Child* 94 (11):888-93. doi: 10.1136/adc.2009.159095.

Tibballs, J., S. Kinney, T. Duke, E. Oakley, and M. Hennessy. 2005. "Reduction of paediatric in-patient cardiac arrest and death with a medical emergency team: preliminary results." *Arch Dis Child* 90 (11):1148-52. doi: 10.1136/adc.2004.069401.

Tibshirani, R. 2011. "Regression shrinkage and selection via the lasso: A retrospective." *Journal of the Royal Statistical Society.Series B: Statistical Methodology* 73 (3):273–282.

Tiewsoh, K., R. Lodha, R. M. Pandey, S. Broor, M. Kalaivani, and S. K. Kabra. 2009. "Factors determining the outcome of children hospitalized with severe pneumonia." *BMC Pediatr* 9:15. doi: 10.1186/1471-2431-9-15.

Torgerson, W. . 1952. "Multidimensional scaling: I. Theory and method. ." *Psychometrika* 17:401–419.

Trautner, B. W., A. C. Caviness, G. R. Gerlacher, G. Demmler, and C. G. Macias. 2006. "Prospective evaluation of the risk of serious bacterial infection in children who present to the emergency department with hyperpyrexia (temperature of 106 degrees F or higher)." *Pediatrics* 118 (1):34-40. doi: 10.1542/peds.2005-2823.

Tsanas, A. 2012. "Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning." DPhil Biomedical Engineering, Department of Mathematics, University of Oxford.

Tsanas, A.; Little, M.; McSharry, P. 2013. "Chapter 7: A methodology for the analysis of medical data." In *Handbook of Systems and Complexity in Health*. Springer.

UNICEF. 2013. Pneumonia diagnostics: current outlook and perspectives.

UNICEF, WHO. 2006. Pneumonia. The forgotten killer of children. New York: United Nation's Children's Fund.

Van den Bruel, A., B. Aertgeerts, R. Bruyninckx, M. Aerts, and F. Buntinx. 2007. "Signs and symptoms for diagnosis of serious infections in children: a prospective study in primary care." *British Journal of General Practice* 57 (540):538-546.

van der Maaten, L., and G. Hinton. 2008. "Visualizing Data using t-SNE." *Journal of Machine Learning Research* 9:2579-2605.

Vannuccini, L., M. Rossi, and G. Pasquali. 1998. "A new method to detect crackles in respiratory sounds." *Technol Health Care* 6 (1):75-9.

Vapnik, V. 1999. "An overview of statistical learning theory." *IEEE Trans. Neural Netw.* 10:988–999.

Victora, C. G., B. R. Kirkwood, A. Ashworth, R. E. Black, S. Rogers, S. Sazawal, H. Campbell, and S. Gove. 1999. "Potential interventions for the prevention of childhood pneumonia in developing countries: improving nutrition." *Am J Clin Nutr* 70 (3):309-20.

Waitman, L.; Clarkson, K.; Barwise, J.; King, P. 2000. "Representation and classification of breath sounds recorded in an intensive care setting using neural networks." *J Clin Monit Comput.* 16 (2):95-105.

Walia, K. 2013. "Point of care investigations in pediatric care to improve health care in rural areas." *Indian J Pediatr* 80 (7):576-84. doi: 10.1007/s12098-013-1016-9.

Walker, C. L., I. Rudan, L. Liu, H. Nair, E. Theodoratou, Z. A. Bhutta, K. L. O'Brien, H. Campbell, and R. E. Black. 2013. "Global burden of childhood pneumonia and diarrhoea." *Lancet* 381 (9875):1405-16. doi: 10.1016/S0140-6736(13)60222-6.

Wang, B.; Miao, L.; Dong, H.; Zheng, Z. 2012. "The research of lung sound signals based on cepstrum analysis." *Proceedings of the International Conference on Biomedical Engineering and Biotechnology*:934-938.]. .

WHO. 2010. *Recommendations on the Management of Diarrhoea and Pneumonia in HIV-Infected Infants and Children: Integrated Management of Childhood Illness (IMCI)*. Geneva.

WHO. 2014a. "Global health expenditure database NHA indicators."

WHO. 2014b. "Integrated Management of Childhood Illness, Chart Booklet."

WHO. 2014c. Revised WHO classification and treatment of childhood pneumonia at health facilities. Evidence Summaries.

WHO. 2017. "Global Health Observatory. Life expectancy: life expectancy data by country. ."

WHO/UNICEF. 2005. Handbook Integrated Management of Childhood Illness.

Willis, J. R., V. Kumar, S. Mohanty, A. Kumar, J. V. Singh, R. C. Ahuja, R. P. Misra, P. Singh, V. Singh, A. H. Baqui, S. Sidhu, M. Santosham, G. L. Darmstadt, and Group Saksham Study. 2011. "Utilization and perceptions of neonatal healthcare providers in rural Uttar Pradesh, India." *Int J Qual Health Care* 23 (4):487-94. doi: 10.1093/intqhc/mzr030.

Wubbel, L., L. Muniz, A. Ahmed, M. Trujillo, C. Carubelli, C. McCoig, T. Abramo, M. Leinonen, and G. H. McCracken, Jr. 1999. "Etiology and treatment of community-acquired pneumonia in ambulatory children." *Pediatr Infect Dis J* 18 (2):98-104.

Xu, J., J. Cheng, and Y. Wu. 1998. "A cepstral method for analysis of acoustic transmission characteristics of respiratory system." *IEEE Trans Biomed Eng* 45 (5):660-4. doi: 10.1109/10.668757.

Yamashita, M.; Himeshima, M.; Matsunaga, S. 2014. "Robust classification between normal and abnormal lung sounds using adventitious-sound and heartsound models." *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*:4451-4455.

Yeginerand, M.; Kahya, Y. 2005. "Modeling of pulmonary crackles using wavelet networks." *Engineering in Medicine and Biology Society IEEE-EMBS* 8:7560–3.

Yilmaz, C. A., and Y. P. Kahya. 2006. "Multi-channel classification of respiratory sounds." *Conf Proc IEEE Eng Med Biol Soc* 1:2864-7. doi: 10.1109/IEMBS.2006.259385.

Young, S. 1996. "Large vocabulary continuous speech recognition: A review." *IEEE Signal Processing Mag.* 13 (5):45-57.

Yu, D.; L. Deng, L.; Dahl, G. 2010. "Roles of Pre-training and Fine-Tuning in Context-Dependent DNN-HMMs for Real-World Speech Recognition." *Proc. of NIPS Workshop on Deep Learning and Unsupervised Feature Learning,* .

Zou, H.; Hastie, T. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society.Series B: Statistical Methodology* 67 (2):301–320.