Towards a Rapid Breath-Based Diagnostic for Pulmonary Tuberculosis in Pediatric Patients

A Thesis
Submitted to the Faculty
in partial fulfillment of the requirements for the
degree of

Bachelor of Engineering

by

Lloyd May

Thayer School of Engineering
Dartmouth College
Hanover, New Hampshire

June 2018

Thesis Committee:

Advisor _____
Prof. Jane E. Hill

Chair _____
Prof. Erland M. Schulson

ABSTRACT

Pulmonary tuberculosis (TB) is a curable disease that claims the lives of hundreds of thousands of children annually. The disease burden is highest in resource limited settings which are often unable to provide timely disease diagnosis. Therefore, there is a need for a sensitive, rapid diagnostic for pulmonary TB in pediatric patients, a large and vulnerable subset of those infected with TB. Human breath contains many volatile metabolites that are produced by the human host as well as by the microbes present inside the host. This volatile organic compound (VOC) profile can be characterized using highly sensitive mass spectrometry methods. The resultant profile can then be used as a tool to diagnose or screen for disease. There are sputum and breath models that have been developed in both non-human primates and adult humans that have shown great promise in using VOC profiles to diagnose TB. Yet no model has been developed for pediatric patients. 34 pediatric breath samples were collected from the Red Cross Memorial Children's hospital in Cape Town, South Africa. These samples were stored on thermal desorption tubes and misanalysed via two dimensional time of flight mass spectrometry. The VOC profile of each sample was used to develop a random forest model to predict TB status from the VOC profile of a sample. A suite of 20 novel, discriminatory compounds were identified.

Acknowledgements

This project was completed under the guidance and supervision of Prof. Jane E. Hill. I would like to thank all members of the Hill Lab for their guidance and patience over the past four years. I would also like to thank our project collaborators in Capetown: Dr. Heather Zar, Sr. Margareta Prins, and Cynthia Whitman.

Contents

1 Introduction and Background **1**

2 Methods **4**
2.1 Breath Collection Kits 4
2.2 Breath Collection Procedure 6
2.3 Mass Spectrometry 7
2.4 Data Formatting and Storage 8

3 Analysis **9**
3.1 Pre-Processing 9
3.2 Model Generation 12
3.3 Feature Selection 13

4 Results & Conclusions **15**
4.1 Results 15
4.2 Conclusion 20

List of Tables

3.1 Number of statistically significant compounds present at varying p-value cut-off. 11

List of Figures

2.1 Components used in every breath kit. A)1.5L Tedlar breath collection bag, B) Mouth piece, C) 0.22 μm filter, D) Thermal disorption tube (TDT) . . . 5
2.2 Example two-dimensional gas chromatogram of 1L breath from a TB positive pediatric patient. 8

3.1 Data reduction and feature selection flow chart 10
3.2 First 3 principal components of all 819 features (right) and the reduced list of 267 features (left) 12
3.3 Flow chart of the implementation and iteration of random forest models. . . 14

4.1 Visualization of the first 3 principal components of the 20 most important features in a balanced two class breath model. 16
4.2 A dendrogram showing heat-map expression of the 20 most important compounds as well as clustering via class. 17
4.3 The averaged predictive probability generated by an RF trained iteratively on the 20 most important features. 18
4.4 List of generic compound identification numbers generated by ChromaTOF as well as their chemical formulas. 19
4.5 Frequency list of compounds present in the dataset that have appeared in animal or adult breath models. 20

Chapter 1

Introduction and Background

This project involves identifying volatile organic molecules (VOCs) present in the breath of pediatric patients that can be used to identify children with pulmonary *M.tuberculosis* (TB) infections. The TB status of each child will be determined using the gold standard technique of culture, as well as nucleic acid amplification (i.e. GeneXpert^{T}M. Breath samples were collected from patients prior to treatment at the Red Cross Memorial Children's Hospital in Cape Town, South Africa. Breath molecules were concentrated and stored on carbon tubes which were analyzed via two-dimensional gas chromatography coupled to time of flight mass spectrometry (GCxGC TOFMS). In this way, the breath molecules present in each sample are separated and relatively quantified. Statistical and machine learning methods will be implemented to determine a suite of discriminatory breath biomarkers that discriminate TB-positive from TB-negative pediatric patients.

In 2016, the World Health Organization reported that 1 million children under the age of 15 contracted TB. This disease now ranks as one of the top 10 causes of childhood mortality for children under the age of 5 [1]. In 2014, 136,000 children lost their lives to the disease [2]. However, this number is likely underreported, as in over a third of reporting countries, children who are co-infected with HIV/AIDS and TB are often reported to have passed away from only HIV/AIDS [3]. Recent studies have shown that tuberculosis and

pneumonia co-infections occur in approximately 1%-23% of pediatric pneumonia cases but are underreported due to difficulties diagnosing TB in pediatric patients [4].

The requirements of conventional diagnostic methods are obstacles in providing adequate care in resource limited settings. Current methods are either rapid but not sensitive to children, such as GeneXpert [5], or take on the order of 4-6 weeks to produce a result, such as culture [3]. All current diagnostic methods require a sputum sample to be coughed up from the lung, a sample that is rarely produced by children [4]. Therefore, induced sputum samples are normally collected in lieu of this. Induced sputum sample collection requires nebulization and can usually only be performed in resourced clinical or hospital settings. Due to these diagnostic limitations, the WHO called for the development of a rapid, non-sputum-based biomarker test that did not require bacterial isolation for the detection of active TB in 2014 [6, 7]. The non-invasive and rapid analysis of exhaled breath could fulfill these requirements and improve outcomes among children, especially in low-resource settings.

Pulmonary TB manifests differently in adults and pediatrics. Adults generally form granulomas, allowing them to produce sputum samples. The majority of the corpus of TB literature focuses exclusively on TB in adults. Pediatric patients, in contrast, do not form granulomas but rather miliary TB which spreads throughout the lung [8]. Pediatrics are far more susceptible to disseminated TB in other parts of the body, such as the eye or bone. These factors contribute to pediatric patients not being able to produce sputum samples and sputum induction methods, such as saline nebulization, are used to gather dilute sputum samples [8].

Breath as a diagnostic tool for disease has been explored in a variety of contexts [9, 10, 11]. Several animal and adult human models have been developed which use volatile organic compound profiles to diagnose or screen for TB. Mgode has used trained Giant Pouched rats to be used as a first line screening test in Tanzania. The rats are trained to identify TB positive sputum samples by sniffing the head-space of the prepared sputum

sample. When this screening strategy was applied as a first-line screening diagnostic for pediatric patients, detection rates of TB increased by 67.6% in 982 cases [12].

A model was developed using longitudinal breath samples from 5 macaques. The model included pre-infection and room air control samples as well as multiple breath samples at different time points post-infection. A suite of 19 discriminatory compounds indicative of active TB infection were found [13].

Breath samples from 34 adults in Haiti were used in conjunction with room air controls to develop a model which differentiated between TB positive and negative patients based on their respective VOC breath profiles. A suite of 22 compounds indicative of active TB was developed [14].

No model has been developed that has explored the use of breath samples to diagnose or screen for TB in pediatric patients, a large and vulnerable portion of the TB population. Breath-based technologies with a high sensitivity could also prove useful in low-resource settings as a screening mechanism. A sensitive, rapid pulmonary TB screening device could help alleviate extended wait-times and I argue that breath may be a viable clinical sample to meet these diagnostic criteria. Volatile molecules on the breath of TB suspect pediatric patients can be used to discriminate between TB positive and TB negative patients.

Chapter 2

Methods

Breath samples were collected from a pediatric population in Cape Town, South Africa. The TB status of recruited patients was determined after breath collection. Confirmed TB negative recruited patients were used as a negative control throughout the study. One room air control sample was collected at the time of each breath collection in order to account for molecules that may be present in the air that the patients are breathing and do not contribute to the biomarkers present in breath sample. The noise present in the data can be reduced by controlling for molecules present in room air in this fashion.

2.1 Breath Collection Kits

The 1.5L Tedlar bags (A in figure 2.1) were chosen for their inert chemical properties as well as their compatibility with the collection system used in this study [14]. Containments in the 1.5L Tedlar breath collection bags were removed by flushing each bag with nitrogen gas three times. Flushing consists of filling each bag with nitrogen gas via an inert vinyl tube and leaving the bag closed in a 32° C room for 30 minutes. The bag is then emptied and the process is repeated for a total of three times. Once each bag has been emptied for a third time, the bag is sealed and tightly rolled for storage in a hermetically sealed storage container. The mouth piece (B in figure 2.1) consists of a standard drinking straw fitted with

Figure 2.1: Components used in every breath kit. A)1.5L Tedlar breath collection bag, B) Mouth piece, C) 0.22 μm filter, D) Thermal disorption tube (TDT)

inert vinyl tubing to allow for the straw to connect directly with the Tedlar bag. The mouth piece is used to make the breath collection process more intuitive and less frightening for children who have often been in the hospital for extended periods of time before breath collection. The mouth piece also allows researchers to collect breath without having to touch any part of the system that has been in direct contact with the patient's mouth as the mouth piece is removed by pulling on the vinyl tubing at the bottom of the straw.

The 0.22 μm filters are prepared by attaching 4 short segments of varying diameter vinyl tubing to a sterile filter. These segments of vinyl tubing allow for the filter to be placed seamlessly in-line during breath sample evacuation. The filters ensure that no bacteria or other organic matter, such as traces of saliva or sputum, are drawn onto the TDTs. This makes the TDTs safer to handle and allows an increased robust storage time of the TDTs before they are processed [15].

The TDTs consist of a three-bed layering containing Carbopack Y, Carbopack X, and Carboxen 1000 (Supelco, Bellefonte, PA), a sorbent combination that has been previously

optimized for the collection of a wide range of breath molecules [15]. The TDTs were conditioned by purging them with nitrogen gas 350° for 2 hours. The tubes are then verified through a GCxGC TOFMS *fast* method that detects the presence of compounds on the tube. The resultant chromatograms are visually inspected to ensure the tube has a low background signal. An additional shipment of hermetically sealed TDTs were included with the shipment of breath kits. These additional TDTs were used for the collection of room air samples.

2.2 Breath Collection Procedure

1.5L whole breath samples were collected from the Red Cross War Memorial Children's Hospital in Cape Town, South Africa. These samples will be classified as *M. Tuberculosis* positive or *M. Tuberculosis* negative based on the gold standard diagnostic of culture as well as nucleic acid amplification (GeneXpert). If a patient's culture or GeneXpert result was positive, that patient's sample was classified as TB positive. Patients were recruited according to the following inclusion criteria:

- Between 2 - 15 years of age
- Have a definitive TB status and drug resistance status as determined by two induced sputum (IS) Gene Xpert/RIF assays as well as culture
- Be healthy enough to provide a voluntary whole breath sample.
- Must not have received any TB medication for more than 72 hours before the date of breath collection.
- Known HIV status.
- The patient's legal guardian(s) are available and willing to provide consent

Patients who did not meet the inclusion criteria were excluded from the study. The patient was given water to rinse the inside of their mouth to reduce the presence of compounds associated with the mouth and oral cavity in the collected sample. The patient was then asked to breath normally into a 1L Tedlar bag until the bag was full. These bags were chosen for their inert properties and low background VOC emissions. The bag was then sealed and the contents of the bag drawn through a 13mm, 0.22μm PTFE filter to remove any potential pathogens, and then drawn onto a three-bed TDT to optimally collect a wide range of breath molecules. The tubes were then sealed in individual airtight bags and stored at 2° C until shipment to Hill Lab for analysis through GCxGC-TOFMS.

2.3 Mass Spectrometry

The samples will be analyzed via GCxGC-TOFMS, a system that allows for increased amplitude of chemical signal and the separation of normally co-eluting molecules, enabling the robust detection of trace elements in breath [13]. The breath molecules that have been collected on the thermal desorption tubes will be desorbed at 330°C into a cryogenically cooled (-120°C) inlet liner of a GCxGC-TOFMS instrument. After desorption, the inlet will be rapidly heated from -120°C to 270°C and the trapped breath molecules will be injected onto an Rxi- 624Sil-MS/Stabilwax chromatography column combination which is well suited for complex mixtures consisting of non-polar and semi-polar volatile molecules, such as breath [13]. After separation, breath molecules will be identified and relatively quantified through mass spectrometry. Mass spectra will be collected over the range of m/z 30-500 at a rate of 200Hz. Putatively identified biomarkers will be confirmed by the injection of standards.

Each black dot in figure 2.2 represents a compound identified by the the analysis after integration. The X-axis represents first dimension retention time in minutes while the y-axis second dimension retention time in seconds. The color represents the intensity of the

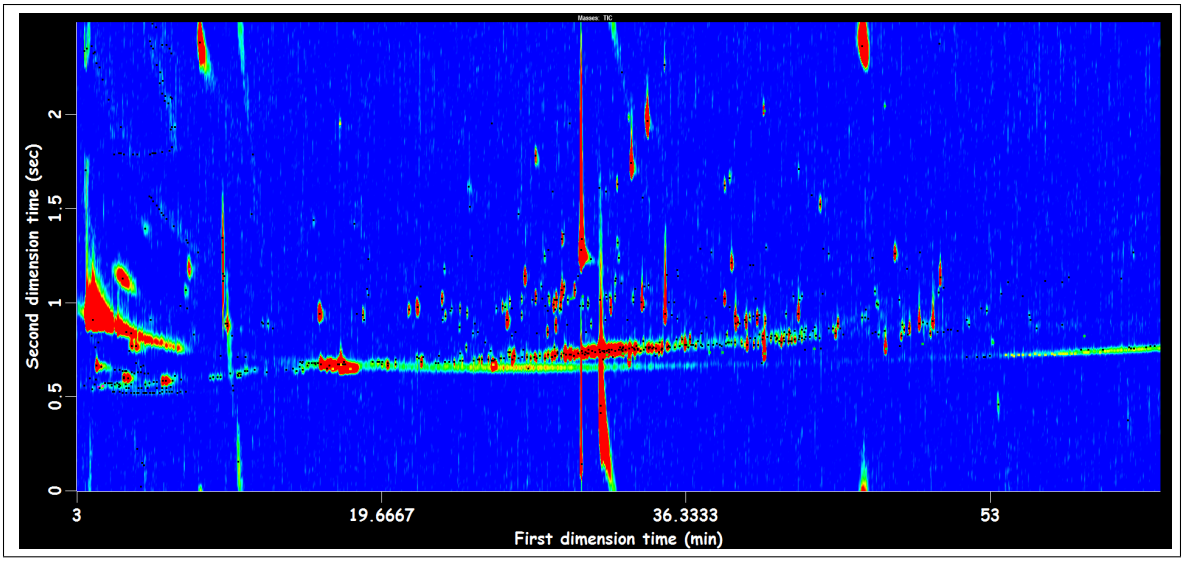

Figure 2.2: Example two-dimensional gas chromatogram of 1L breath from a TB positive pediatric patient.

compound while the area represents the signal relative to the compound. Therefore a higher area is indicative of a higher relative concentration.

2.4 Data Formatting and Storage

De-identified patient data was collected from the Red Cross Memorial Children's Hospital, including patient sex, age, HIV status as well as tuberculosis diagnostic test results (GeneXpert and culture results). Patient information was linked to patient breath samples through use of a de-identified numbering system. The results of the GCxGC-TOFMS analysis of the patient's breath samples, as well as room air, was rendered as .csv containing the relative abundance, signal to noise ratio as well as the first and second retention times. This raw file was formated into a Python Pandas dataframe containing all de-identified patient information as well as the relative abundance of all compounds present in samples. A class column was included which summarized the samples status as either TB positive, TB negative or a room air sample. All data was stored locally as well as back-ups on Thayer, Kite as well as Jupyter servers.

Chapter 3

Analysis

Raw data from the ChromaTOF software was used as the dataset for analysis. This dataset was formatted and normalized before pre-processing. The pre-processing steps focus on noise reduction and signal preservation. A more robust model can be developed with a dataset that contains the highest signal to noise ratio. This was achieved by the removal of features contributed by room air as well as by the removal of sparse features from the dataset.

3.1 Pre-Processing

A list of 1080 compounds were identified by the GCxGC-TOFMS analysis. This list contained the compound name, identification number as well as relative abundance, signal to noise ratio and retention times for each compound in each sample. A 75% similarity cutoff was used when identifying a putative compound name. If a sample did not contain a compound, a zero was inserted in all entries for that compound-sample pair. All pre-processing and noise reduction techniques are shown in figure 3.1. This list was manually screened for known contaminants and plastacids that are introduced by the collection system's bag and pump as well as from the GCxGC-TOFMS column bleed [16]. After cleaning, a suite of 819 compounds remained for analysis. The $\log_{10}$ of the abundance of each compound was

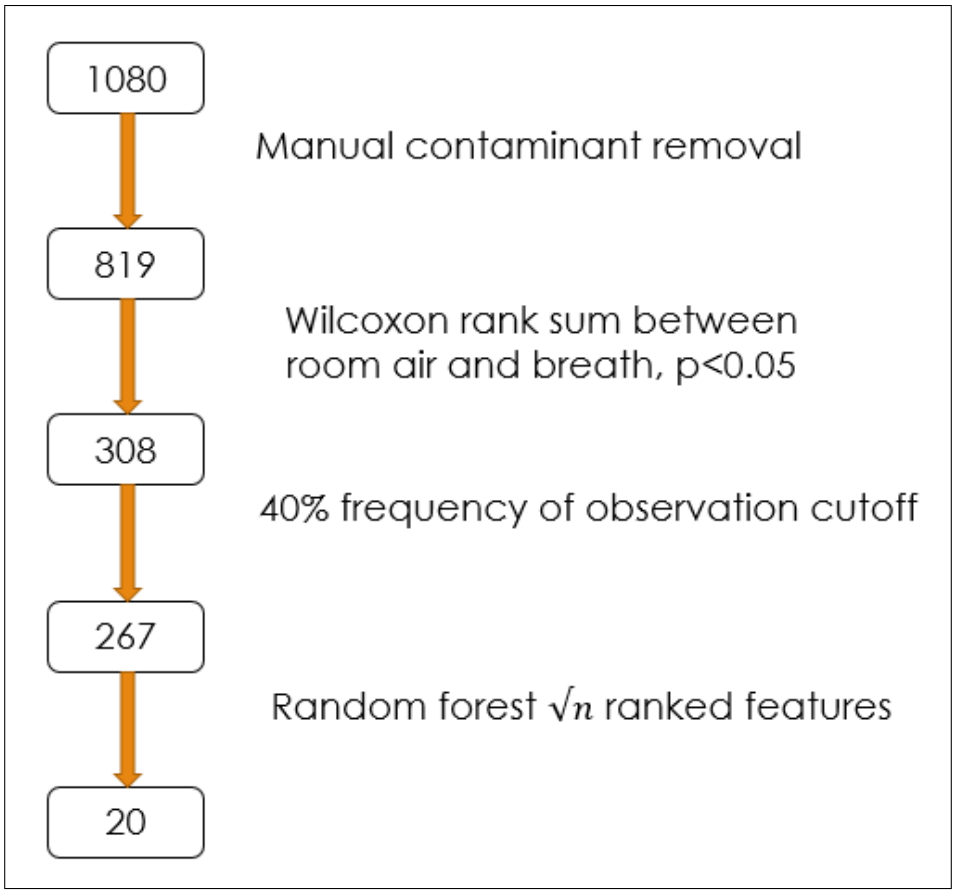

Figure 3.1: Data reduction and feature selection flow chart

taken as well as a probabilistic quotient normalization (PQN) implemented to normalize the data [17]. This normalization aids in reducing model bias towards compounds with a higher relative abundance.

A list of the relative abundances of all 819 compounds present in breath samples was generated as well as the same list for room air samples. A Wilcoxen sum-rank test was then performed across these two lists to generate a new list of compounds present in breath samples that are statistically different from room air samples [18]. This is done to reduce the noise present in the data, removing compounds from the dataset that were contributed by room air. Patients inhale room air and many compounds do not interact with the respiratory system and are simply exhaled. These compounds that do not add significant information into the system are removed from the analysis. A Wilcox-rank sum test was chosen as

normality does not have to be assumed and the test does not have to be parameterized or optimized in any way [19, 20].

Table 3.1: Number of statistically significant compounds present at varying p-value cut-off.

P-value Cut-off	Number of Compounds
0.2	459
0.1	397
0.05	308
0.01	135
0.005	93

Various p-value cut-offs were explored and resulted in a different number of qualified compounds. A cut-off of $p < 0.05$ was used which resulted in a suite of 267 compounds that were statistically different from room air. This cut-off was chosen to reduce noise as far as possible without a significant loss of information within the system.

A 40% frequency of observation cutoff was implemented to remove sparse features from the dataset. This cutoff removed features that were not present in at least 40% of either of the two classes, TB positive and TB negative breath samples. After this cutoff, 267 compounds were present in the dataset.

As seen in figure 3.2, there is significant separation between room air and breath samples based on the first 3 principal components alone. The angle of perspective which showed greatest separation between classes was chosen for each plot separately in figure 3.2. There is no apparent separation between TB positive and negative breath samples after pre-processing. If separation between the two groups was present after this step, then a further investigation into confounding collection practices between groups would need to be conducted. As the samples were collected before the TB status of the patient was known, there is no reason to assume confounding factors were present in collection procedures.

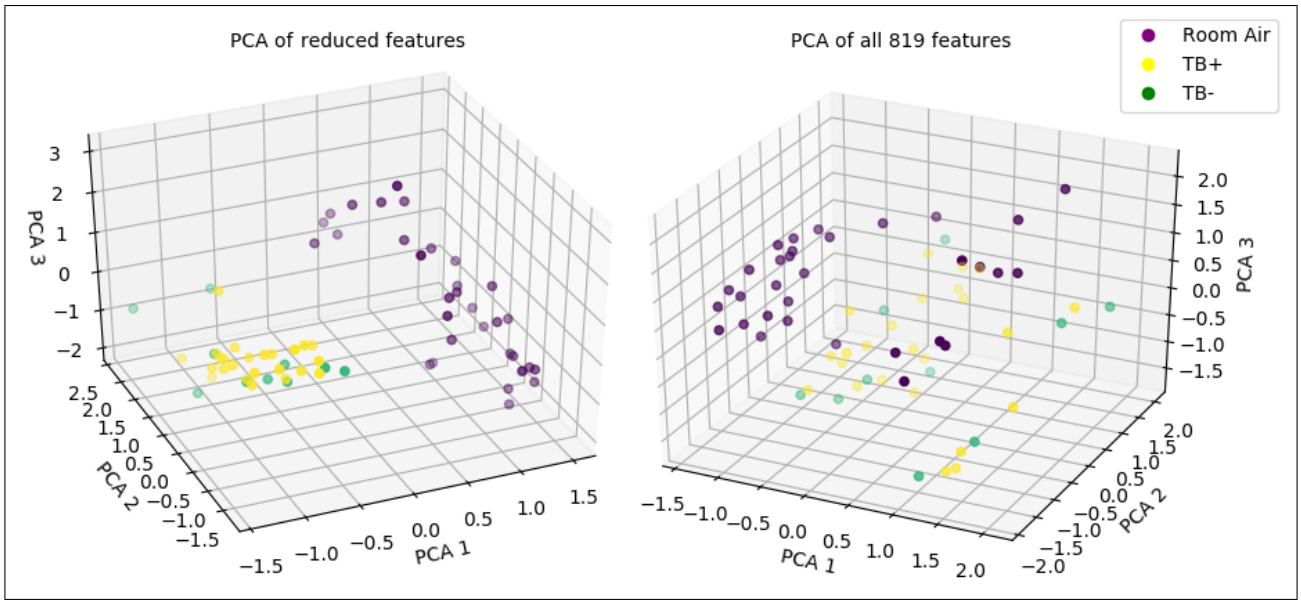

Figure 3.2: First 3 principal components of all 819 features (right) and the reduced list of 267 features (left)

3.2 Model Generation

A random forest (RF) model was selected to generate a list of discriminatory compounds. The model is generated by implementing a large forest of decision trees which uses a gini operation to generate a probabilistic classification of each sample. Each decision tree is given a vote and these votes and scores are used to generate the model. Random forest was chosen as it generates a list of feature importance at the conclusion of the model as well as allowing for weighting and balancing hyper parameterization at each node [19]. This is desirable as the intended outcome of the analysis is to determine a reduced suite of discriminatory molecules that could be used to design a screening test or diagnostic for pediatric TB. More complex models such as recursive neural net and non-linear support vector machines (SVM) would achieve better classification results, but would not generate a list of feature importances. It may also not be possible to translate the results of these complex methods into a physical screening or diagnostic device as the models would be built on all 267 features. The relative abundance of these 267 compounds would then have to be collected by any diagnostic system which leverages this information. As the proposed

focus of these devices is a rapid and cost-effective unit, requiring the device to detect 267 compounds accurately is an unnecessary and inefficient design constraint.

A linear SVM was also explored as the square of the coefficients of the model can be interpreted as feature importance if it can be assumed that the data has been sufficiently normalized. The linear SVM model is a coarser approach when compared to RF as it creates a linear hyperplane decision boundary between classes whereas RF uses a generative ensemble model.

3.3 Feature Selection

The data set comprising of 267 compounds present in 34 breath samples (10 TB positive and 24 TB negative) was used in an implementation of an RF analysis. All steps in the RF implementation and iteration are shown in figure 3.3. The classes were randomly balanced at the start of the model by randomly selecting 10 negative breath samples and all of the positive breath samples. This was done to avoid the bias towards sole negative classification, favoring sensitivity over specificity. A training-test split of $\frac{1}{3}$ was then used, partitioning the data into a training and testing set. The split was balanced, ensuring that the same number of TB positive and negative samples appeared in the training and testing datasets at each iteration.

A 500 tree RF was then trained on the 14x267 training data. A list of compound importances was generated and the top 20 most important features were taken and used to train a new RF model. 20 was chosen as it is the closest integer approximation to the $\sqrt[2]{n}$, which is commonly used as an importance cutoff in RF analysis [14]. This new RF model trained on the 14x20 training set was validated using the 6x20 testing set that was withheld at the start of the iteration. A frequency list of compounds that appeared in the 20 selected compounds was updated at the conclusion of each iteration as well as the predictive probability for each of testing samples. This process was repeated 500 times and a grand list of all

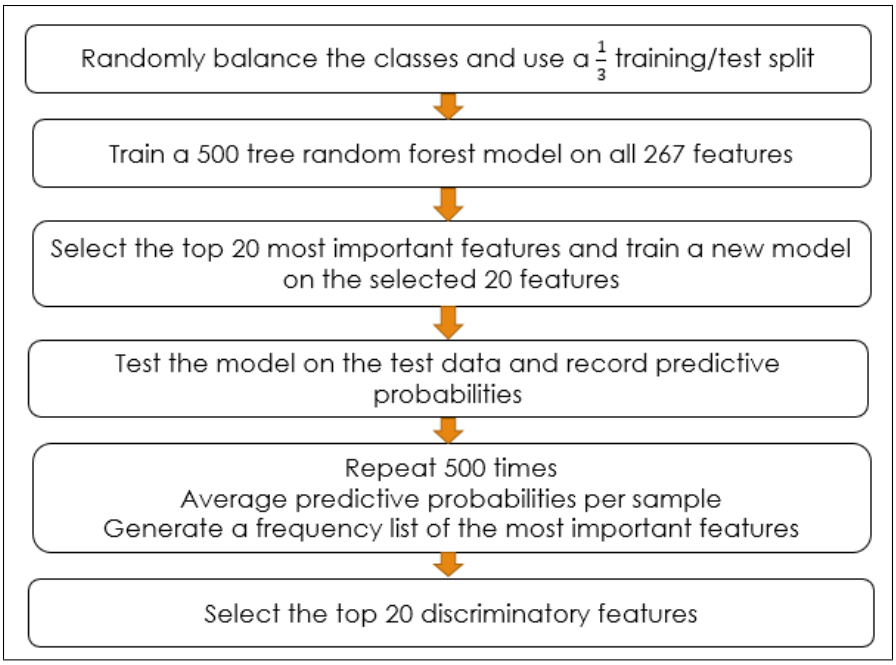

Figure 3.3: Flow chart of the implementation and iteration of random forest models.

sample predictive probabilities and compound importance frequency list was generated.

Chapter 4

Results & Conclusions

The feature reduction strategy implemented in Chapter 3 resulted in a suite of 20 discriminatory compounds. The expression, variation and presence of these compounds in the literature will be explored in this chapter. The list of compounds of interest present in literature will also be cross-referenced to the list of all compounds found both room and air samples. The putative compound names were withheld for intellectual property protection and chemical formulas were given where possible. Steps for future work and longitudinal study design will be explored at the conclusion of the chapter as well as concluding statements and findings.

4.1 Results

The variability in the suite of 20 discriminatory compounds, as listed in figure 4.4, is visualized using a principal components analysis (PCA) in figure 4.1. A PCA transforms multidimensional data into a lower-dimensional space while maximizing the variance. There is a degree of clustering of TB positive samples in this transformed space. The lack of clustering of the TB negative samples in the visualization could imply that there is greater variability between these samples when the 20 compounds of interest are reduced to their respective principal components.

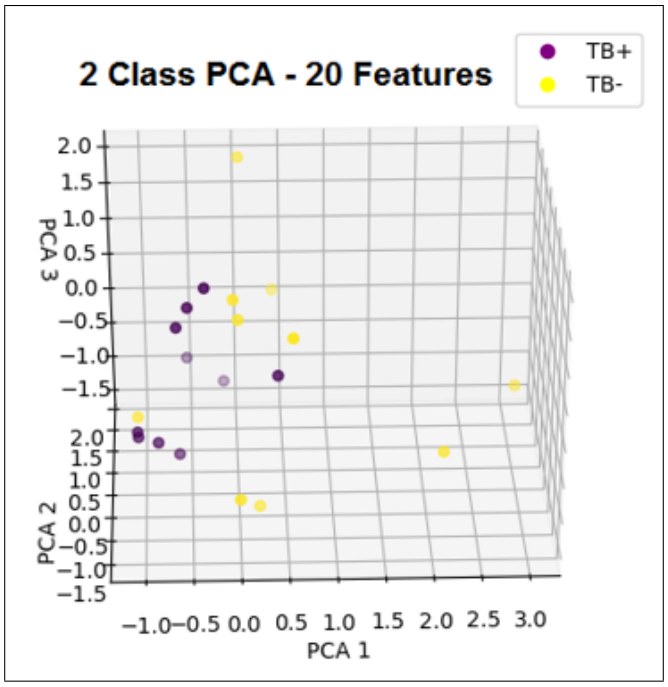

Figure 4.1: Visualization of the first 3 principal components of the 20 most important features in a balanced two class breath model.

A dendrogram of the 20 compounds of interest showed a large degree of sample clustering in figure 4.2. The dendrogram calculates the distance between two samples using the weighted city-block protocol and clusters according to this distance. The expression level of each compound of interest can be seen in the denrogram associated heat-map. The color shows the normalized intensity of expression of each of the 20 compounds of interest in each breath sample. The compound identification number can be linked to a chemical formula or analyte tag as shown in figure 4.4.

The predictive probability of the RF model trained on the 20 most important compounds found by the RF trained on all compounds was stored after every iteration. Each individual sample was part of the testing data split a minimum of 150 out of the 500 iterations. Therefore, an average predictive probability can be generated for every sample in the dataset by averaging the stored predictive probability over the number of times that sample appeared in the testing data split. This method of aggregated probabilities was implemented as the dataset was not large enough for a portion of the dataset to be completely withheld as an

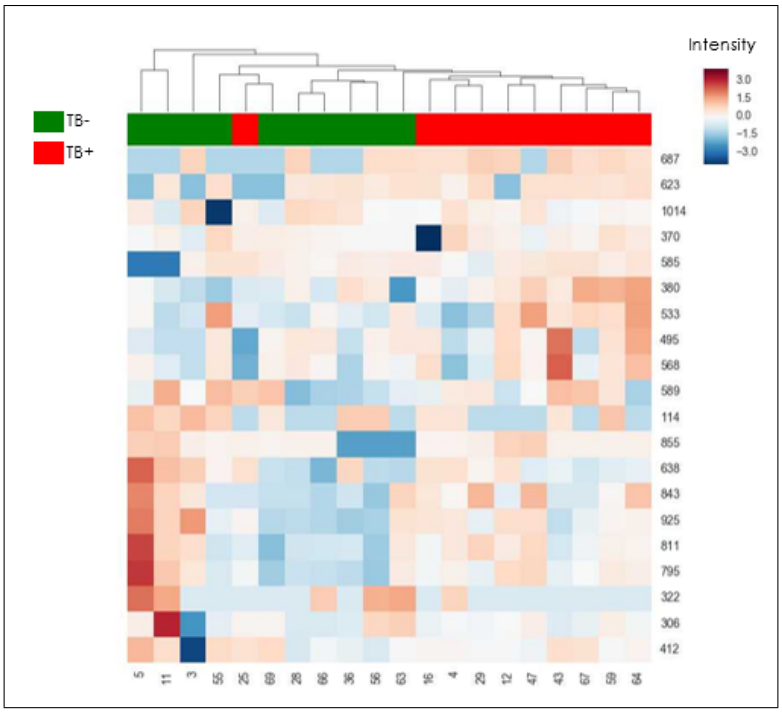

Figure 4.2: A dendrogram showing heat-map expression of the 20 most important compounds as well as clustering via class.

external validation set. Therefore the model suffered from the same slight over fitting that is present in traditional k-fold or leave-one-out cross-validation strategies. However, the averaged predictive probabilities were used to estimate the potential of the suite of compounds to be used for diagnostic purposes rather than to determine the power or accuracy of the compounds as a diagnostic tool itself. The averaged predictive probabilities visualized in figure 4.3 show that the suite of 20 compounds has the ability to predictive TB status in this model. This finding illustrates the need to continue the study, expanding the patient sample pool as well as extending the study to included longitudinal sample collection.

The list of compounds present in the suite of 20 with highest RF feature importance was cross-re fenced to compounds previously cited in animal and adult models [13, 14]. When the putative compound names were cross-referenced to these lists, none of the suite of 20 compounds had been previously identified in the literature. The chemical classification and name of the compounds would need to be verified through analysis of standards and other known compounds before definitive compound classification could occur.

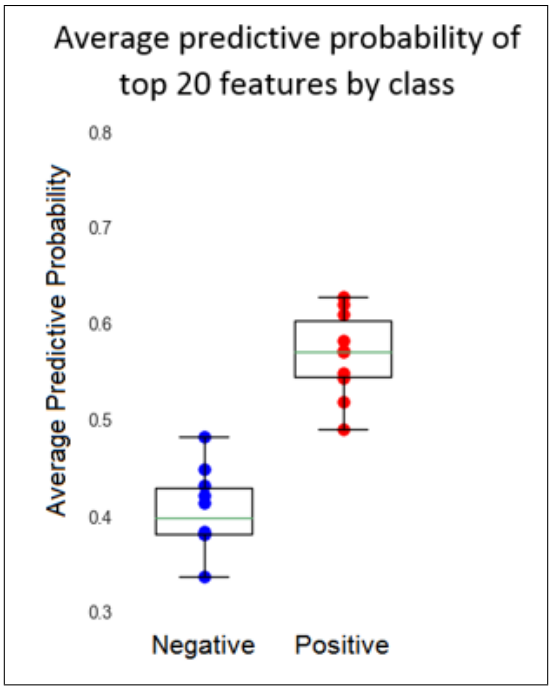

Figure 4.3: The averaged predictive probability generated by an RF trained iteratively on the 20 most important features.

The compound identification number along the y-axis of figure 4.2 refers to the list indicated in figure 4.4.

The total list of 819 compounds present in the dataset of manual contamination removal was compared with those compounds identified in previously cited in animal and adult models [13, 14]. Two compounds were present on both lists. The mean and standard deviation of the PQN $\log_{10}$ of expression of each compound as well as a count of the number of samples in each class the compound appeared in is provided in figure 4.5. A cyclic hydrocarbon was cited in an adult model developed from breath samples of 34 adults in Haiti and had increased expression in TB positive adults [14]. This cyclic hydrocarbon was present in 33% of TB positive pediatric breath samples, 12.5% of TB negative pediatric breath samples and none of the room air samples.

The linear hydrocarbon was cited in a model developed on longitudinal breath samples of 5 macaques [13]. The compound had higher expression in the breath of macaques post-infection. The compound appeared in all breath samples as well as 52% of room air

Compound #	Formula	Compound #	Formula
687	$C_7H_{15}Cl$	114	C_2H_6S
623	$C_{11}H_{22}$	855	C_9H_{18}
1014	$C_8H_{18}O$	638	Analyte 638
370	$C_6H_{12}O_2$	843	C_8H_{16}
585	$C_{13}H_{28}$	925	$C_{12}H_{26}$
380	Analyte 380	811	$C_{13}H_{28}$
533	$C_{12}H_{26}$	795	$C_8H_{16}O$
495	C_9H_{12}	322	$C_4H_8O_2$
568	$C_9H_{10}O$	306	$C_5H_{10}O$
		412	$C_7H_{14}O$

Figure 4.4: List of generic compound identification numbers generated by ChromaTOF as well as their chemical formulas.

samples. The mean expression of the compound indicated the compound does not appear to be differentially expressed between TB positive and TB negative breath sample classes.

The presence of the these previously cited compounds in the data as well as the predictive capabilities of a model developed from the VOC profile of pediatric breath samples motivate the continuation and expansion of the study. Through increased patient recruitment, an external validation set can be withheld from the model, allowing for a more robust evaluation of the model. An increased number of TB positive breath samples will also allow other balancing and weighting methods to be implemented other than random manual balancing which excludes information from the model. Longitudinal follow-up samples would allow for a more robust model to be developed that includes the effect of TB treatment on pediatric VOC breath profiles over time as well as the inclusion of hierarchical time series data to the model. The effect of potential confounding factors, such as HIV status, age and sex, as well as the interaction of these factors with compound expression, can be included in the development of a more robust model with a larger number of samples and time series information.

	RA		TB+		TB-	
	Count	Mean	Count	Mean	Count	Mean
Cyclic hydrocarbon	0	0	3	3.53 (0.48)	3	3.88 (0.33)
Hydrocarbon	17	3.50 (0.42)	10	3.79 (0.36)	22	3.75 (0.38)

Figure 4.5: Frequency list of compounds present in the dataset that have appeared in animal or adult breath models.

4.2 Conclusion

This is the first study, to our knowledge, which has examined the VOC profile of pediatric patients and developed a model to classify TB disease status based on the profile. The study used a novel and potentially clinically relevant sample, i.e. breath, in a clinical research setting in Cape Town, South Africa. A suite of 20 compounds of interest were identified via an iterative RF model and a list of feature importance and sample predictive probabilities was generated. The average predictive probabilities generated from the model indicated that the model could differentiate between TB positive and TB negative breath samples. None of these 20 compounds had been previously cited in adult or animal breath models.

The model is likely slightly over-fit as it suffers from the same information containment issues associated with other traditional cross-validation approaches. The results are still indicative that a breath-based device may be a viable option for diagnosing or screening for TB disease in pediatric patients, specifically those in resource limited settings. These findings motivate the continuation of sample collection and analysis, both in the form of increased patient recruitment and the collection of longitudinal follow-up sample collection. A balanced sample set of 20 TB positive and 20 TB negative breath samples would serve as the minimum amount required to develop a model with a withheld validation set if the effect size identified is indicative of the population.

Bibliography

[1] World Health Organization. Global tuberculosis report 2016. page 214, 2016.

[2] Helen E. Jenkins. Global burden of childhood tuberculosis. *Pneumonia*, 8(1):24, 2016.

[3] Stephen M. Graham, Charalambos Sismanidis, Heather J. Menzies, Ben J. Marais, Anne K. Detjen, and Robert E. Black. Importance of tuberculosis control to address child survival, 2014.

[4] Jacquie N. Oliwa, Jamlick M. Karumbi, Ben J. Marais, Shabir A. Madhi, and Stephen M. Graham. Tuberculosis as a cause or comorbidity of childhood pneumonia in tuberculosis-endemic areas: A systematic review. *The Lancet Respiratory Medicine*, 3(3):235–243, 2015.

[5] Mark P. Nicol and Heather J. Zar. New specimens and laboratory diagnostics for childhood pulmonary TB: Progress and prospects, 2011.

[6] WHO. High-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting. *WHO Meeting Report*, (April):1–98, 2014.

[7] Li Liu, Shefali Oza, Daniel Hogan, Jamie Perin, Igor Rudan, Joy E. Lawn, Simon Cousens, Colin Mathers, and Robert E. Black. Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: An updated systematic analysis. *The Lancet*, 385(9966):430–440, 2015.

[8] Ann M. Loeffler. *Pediatric Tuberculosis*. 2003.

[9] Anton Amann, Wolfram Miekisch, Jochen Schubert, Bogusław Buszewski, Tomasz Ligor, Tadeusz Jezierski, Joachim Pleil, and Terence Risby. Analysis of Exhaled Breath for Disease Detection. *Annual Review of Analytical Chemistry*, 2014.

[10] Bogusław Buszewski, Martyna Kesy, Tomasz Ligor, and Anton Amann. Human exhaled air analytics: Biomarkers of diseases, 2007.

[11] Wolfram Miekisch, Jochen K Schubert, and Gabriele F.E Noeldge-Schomburg. Diagnostic potential of breath analysis-focus on volatile organic compounds. *Clinica Chimica Acta*, 2004.

[12] Georgies F Mgode, Christophe L Cox, Stephen Mwimanzi, and Christiaan Mulder. Pediatric tuberculosis detection using trained African giant pouched rats. *Pediatric Research*, 00(July 2017):1–5, 2018.

[13] Theodore R. Mellors, Lionel Blanchet, JoAnne L. Flynn, Jaime Tomko, Melanie O'Malley, Charles A. Scanga, Philana L. Lin, and Jane E. Hill. A new method to evaluate macaque health using exhaled breath: A case study of M. tuberculosis in a BSL-3 setting. *Journal of Applied Physiology*, 122(3):695–701, 2017.

[14] Marco Beccaria, Theodore R. Mellors, Jacky S. Petion, Christiaan A. Rees, Mavra Nasir, Hannah K. Systrom, Jean W. Sairistil, Marc Antoine Jean-Juste, Vanessa Rivera, Kerline Lavoile, Patrice Severe, Jean W. Pape, Peter F. Wright, and Jane E. Hill. Preliminary investigation of human exhaled breath for tuberculosis diagnosis by multidimensional gas chromatography - Time of flight mass spectrometry and machine learning. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, 2018.

[15] Mark Libardoni, P. T. Stevens, J. Hunter Waite, and Richard Sacks. Analysis of human breath samples with a multi-bed sorption trap and comprehensive two-dimensional gas chromatography (GC X GC). *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, 842(1):13–21, 2006.

[16] Heather D Bean, Jean-Marie D Dimandja, and Jane E Hill. Bacterial volatile discovery using solid phase microextraction and comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry. *J. Chromatogr.*, 2012.

[17] Frank Dieterle, Alfred Ross, Götz Schlotterbeck, and Hans Senn. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in1H NMR metabonomics. *Analytical Chemistry*, 2006.

[18] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 1947.

[19] A. Smolinska, A. Ch Hauschild, R. R.R. Fijten, J. W. Dallinga, J. Baumbach, and F. J. Van Schooten. Current breathomics - A review on data pre-processing techniques and machine learning in metabolomics breath analysis. *Journal of Breath Research*, 8(2), 2014.

[20] BarunK Nayak and Avijit Hazra. How to choose the right statistical test? *Indian Journal of Ophthalmology*, 59(2):85, 2011.