A data mining approach to analysing airborne wood particulate concentration and atmospheric data

Subana Shanmuganathan
Auckland University of Technology (AUT), New Zealand
subana.shanmuganathan@aut.ac.nz

Rosziati Ibrahim
Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja
rosziati@uthm.edu.my

Siti Halimah Bt Bakori
Kolej PolyTech MARA
Batu Pahat, Johor
ct_halimah@gapps.kptm.edu.my

***Abstract*— Exposure to airborne wood (hard and soft) dust can lead to a number of diseases, such as asthma, emphysema, bronchitis and upper respiratory tract cancers, lately even proven to be linked to elevated risks for chromosomal instability in cells of the aerodigestive tract. In this context, the paper investigated the particulate wood dust concentrations in a university environment near a timber mill using a data mining approach consisting of JRip, J48 algorithms and a multilayer perceptron (MLP). The data collected consists of particulate wood concentrations and related atmospheric conditions recorded over a few days at four different locations within the university situated next to the timber mill. The results reveal that ORICC is the location most exposed to high concentrations of wood dust (up to 1.57 MG/M^3 at times). This exceeds the recommended exposure limit of 1 MG/M^3 for humans if the dust particles were of hardwood hence, more tests are recommended to establish the airborne particulate wood dust composition from the factory.**

Keywords—WEKA, J48, JRIP,

I. INTRODUCTION

Exposure to airborne particulate, wood dust (hard as well as soft wood) can lead to a number of diseases, such as asthma, emphysema, bronchitis and upper respiratory tract cancers [1, 2]. Recently it has been shown to be even linked to elevated risks for chromosomal instability in cells of the aerodigestive tract [3]. In view of these facts, the paper investigated a data mining approach to analysing the elevated levels of airborne wood dust in four different locations within UTHM (Universiti Tun Hussein Onn Malaysia) located next to a timber factory. The factory's proximity is of concern as the university staff and student are exposed to considerably high levels (1.57 MG/M^3) of particulate wood dust (could be from hard or soft wood, it is unknown at this moment) at certain areas (ORICC). The recommended limit for **hardwood** dust exposure for humans is 1MG/M^3 [4]. Meanwhile, effects to relocate the timber factory in the past have failed as it employs around a thousand people from the local community.

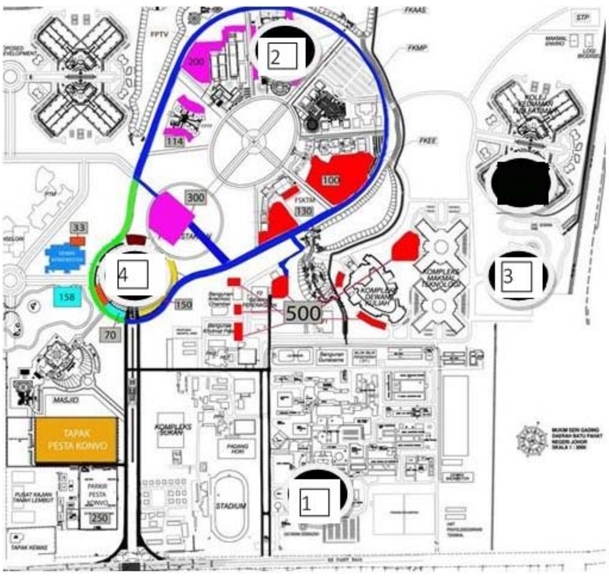

Fig. 1. Diagrammatic view of UTHM locations airborne wood dust were collected and the timber factory.

The wood dust as well as other atmospheric data collected using a piece of E-Sampler Particulate Monitor [4] over a few different days at four different points at irregular intervals are analysed using data mining methods, namely J48, JRip algorithms and a MLP (an artificial neural network (ANN) architecture) in WEKA software [5]. Full details of the data collected and the location are presented in section II following which the methodology is elaborated. In section IV, the results obtained are discussed. Finally, the conclusions of this investigation are summarised.

II. BACKGROUND

A. UTHM and the timber factory

Since the transformation of Pusat Latihan Staf Politeknik (PLSP) into a full university called Universiti Tun Hussein Onn Malaysia (UTHM), in Batu Pahat, Johor in western Malaysia in 2007, the local community has thrived enormously. It has indeed brought a lot of vibe, jobs and students into this part of the country. From such humble beginning as a staff training institute in the engineering field in 1993, to a full university producing several thousand graduates a year mostly engineering, the institution's growth within the last two decades has been remarkable. The township has accommodated the institute's growth gracefully. However, the recent deadlock between a nearby timber factory and the university has prompted the authors to investigate into the concentration of airborne wood dust from the factory especially, into the university premises.

B. E-Sampler data

One piece of E-Sample particulate monitor was set up to collect data on the airborne wood particulate concentrations at four points (fig. 1) near to different faculties/ offices within the university compound.

The E-Sampler is a type of nephelometer and it can automatically measure and record real-time airborne PM10, PM2.5, or TSP particulate concentration levels. It uses the principle of forward laser light scatter system. While in operation the E-Sampler draws sample air into the equipment and passes the air through the laser optical module. The particulate in the sample air stream will scatter the laser light through reflective and refractive properties. A photodiode detector at a near-forward angle collects the scattered light and processes the resulting electronic signal to determine the concentrations producing a continuous, real-time measurement of airborne particulate mass [4].

The same E-Sample monitor was set up at four different locations within UTHM for a consecutive few days to measure airborne wood particulate concentrations and the atmospheric conditions. The following are the location related details (fig 1):

L code	Location	Time recorded	interval
1	Dtii	26 Aug-26 Sep 2013	hourly
2	FKAAS	12-15 April 2013	10 mins
3	ORICC	17-27 June 2013	hourly
4	Library	19-21 June 2013	15 mins

The data elements collected by E-Sample are as follows (note: underlined data are the vectors used in the analysis):

1) area code: code to define the areas i.e., 1-4.
2) Conce: Real-time particulate concentration, in milligrams per cubic meter.
3) Flow: Real-time sample flow rate, in actual litre/minute
4) AT: Ambient temperature in degrees C.
5) BP: Ambient barometric pressure in Pascals.
6) RHx: External ambient relative humidity
7) Rhi: Internal filter sample relative humidity.
8) WS: Wind speed in meters per second
9) WD: Wind direction in degrees (if equipped).
10) BV: Battery voltage (or DC input voltage).

III. The Methodology

The data gathered is analysed using simple graphs and then with data mining techniques, namely JRip, J48 and multilayer perceptron (MLP) in WEKA. JRip is the open source Java implementation of a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER) that was originally proposed by William W. Cohen. Meanwhile, J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool. The C4.5 is a modified version of ID3, the modifications include: handling of both continuous and discrete attributes by creating a threshold and then splitting the list into those whose attribute value is above the threshold and those that are less than or equal to it, training data with missing attribute values, attributes with differing costs and pruning of trees by replacing the not useful branches with leaf nodes [6].

MLP in WEKA is an artificial neural network (ANN) architecture with a single hidden layer [7]. It uses an optimization class by minimizing the squared error plus a quadratic penalty with the BFGS or Broyden–Fletcher–Goldfarb–Shanno algorithm method. ANNs are biologically inspired networks of processing elements called neurons that again mimic animal (human) brain cells. ANNs provide an approach to incorporate heuristics into conventional algorithmic computing. The latter needs step-by-step instructions to solve a problem. Hence, using conventional computing, complex problems, such as

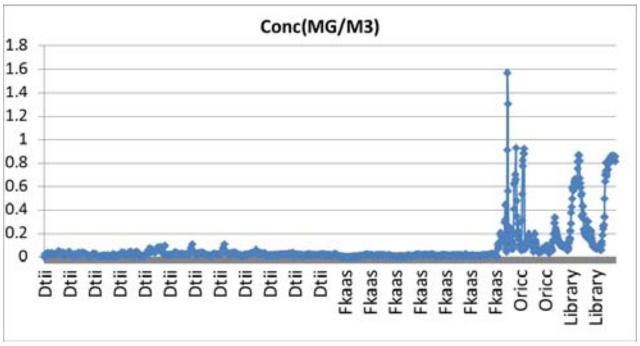

Fig. 2. Graph showing the difference in concentrations of airborne wood dust at the four different locations within UTHM (shown in fig 1).

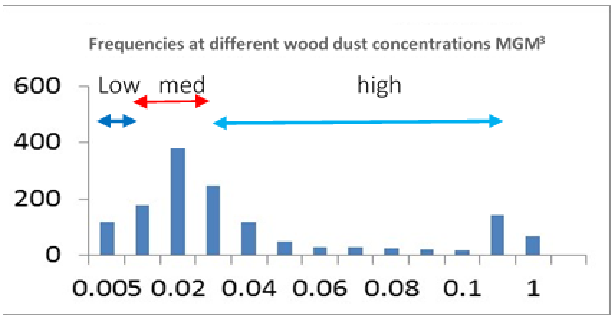

Fig. 3. Distribution of airborne wood dust data from all 4 UTHM locations.

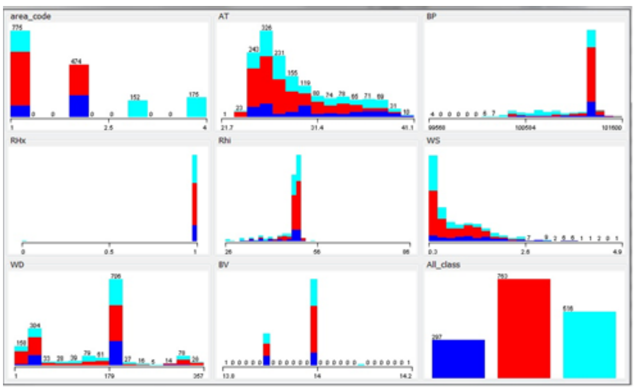

Fig. 4. Distribution of different attribute data (atmospheric conditions relating to different wood particulate concentrations).

pattern recognition, data mining cannot be performed. Data mining techniques are useful when analysing ad-hoc data sets, such as the one studied in this work that cannot be analysed using conventional rigorous statistical methods.

Initially, the data collected from all four UTHM locations were analysed together to get an understanding of the data set. The variability in the wood particulate concentrations and atmospheric conditions experienced in the different locations can be seen in figs 2-4. As the J48 and JRip classifiers require a categorical target variable, wood dust concentration is classified into three classes based on the data distribution and the classes are: "low" <=0.01, "med"<=0.03 and ,"high"> 0.03 (fig. 3 and 4). The same target classes are used for the MLP method. In the second stage, the target variable was reclassified to suit the wood particulate concentrations of the individual locations.

IV. RESULTS AND DISCUSSIONS

The results of the data mining approach investigated (JRIP, J48 and MLP) are discussed in this section. A graph (fig. 2) plotted to see the airborne wood dust concentrations at the four different locations within UTHM show that ORICC and library as the areas exposed to dust at a higher range (0.2-1.6 MG/M^3). ORICC has the highest 1.571 MG M^3 recorded on 19 June 2013 at11:00:00_am. In the other two locations the range has been 0-0.1 MG/M3. Attribute data distribution (fig 4.) as well shows that the wood particle concentrations as high at all time in locations 3 and 4. Meanwhile, location 1 has low, med and high whereas, location 2 has only low and medium concentrations.

A. JRip results of all four UTHM locations

Initially, data collected from all four locations is analysed together using the JRip classification algorithm to see the patterns in terms of atmospheric conditions relating to airborne wood particle dispersal and are discussed in this section.

Atmospheric conditions for high airborne wood particulate concentrations (>0.03 MG/M^3) at the four locatins are found in JRip rules no. 8-14 (fig. 5). Rule no 8, could be interpreted as all 325 instances falling under the high class for this rule to be from locations 3 and 4 (ORICC and library) with no exceptions. This can be confirmed by the graph in fig 2 as well. Rules 9-14 relate to location 1 (Dtii), class high wood particulate concentration. The following are the interpretation for these rules (fig 5):

Rule no. 9; wind from north-northeast direction (WD) <= 40^o) and at speeds less than 0.3 meters/second (WS <= 0.3), with relative humidity equal or greater that 50% (Rhi >= 50) and at temperatures less or equal to 25^{o}C (AT <= 25) lead to **high** wood particulate concentrations at this location with all 13 instances and no exceptions (13.0/0.0) at Dtii. Rule no. 10; wind from north-34^o -40^o northeast direction (WD <= 40) and (WD >= 34) and at speeds less of equal to 0.5 meters/second (WS <= 0.5) have led to **high** wood particulate concentrations, 24 instances with 3 exceptions (24.0/3.0). Rule no. 11; wind form north-northwest (WD<=40) and at speeds >= 40 meters/sec when atmospheric pressure is <= 101306 P, have led to **high**, 29 instances with 7 exceptions. Rule no. 12; temperatures greater or equal to 35.3^{o}C (AT >= 35.3) and less than 37.5oC (AT <= 37.5) and relative humidity less than or equal to 41% (Rhi <= 41), have led to **high** particulate concentrations, 20 instances with one exception. Rule no.13; wind speeds less than 0.3 meters/second (WS <= 0.3) and from the north (WD <= 1) and at relative humidity greater or equal to 50% (Rhi >= 50) and at temperature less or equal to 27.3^{o}C (AT <= 27.3) have led to **high** with 34 instances with 9 exceptions. Rule no. 14; temperatures between 31.1-33.9 ^{o}C (AT >= 31.1) and (AT <= 33.9) meaning on very hot days, wind speeds less or equal to 0.6 meters/second (WS <= 0.6) have led to **high** with 13 instance no exceptions (13.0/0.0).

The stratified cross validation summary (fig 6) gives the correctly classified instances as 80.84% which is good. In the accuracy ratings, class high has the highest precision (92%) meanwhile, class medium has the highest recall (90%). The confusion matrix at 10-fold cross validation shows the classification accuracy of the three classes.

1) (WS >= 0.9) and (WD <= 181) and (AT >= 34.8) and (WD >= 181) and (Rhi <= 44) => All_class=low (77.0/0.0)
2) (WD <= 181) and (WS >= 0.8) and (WD >= 181) and (WS >= 1.7) => All_class=low (55.0/14.0)
3) (WD <= 42) and (WD >= 41) and (BP <= 101267) => All_class=low (47.0/12.0)
4) (BP >= 101170) and (WS >= 0.9) and (BV >= 14) and (AT >= 29.3) and (AT <= 29.9) => All_class=low (32.0/12.0)
5) (WD <= 42) and (WD >= 41) and (AT >= 25.8) and (AT <= 25.9) => All_class=low (12.0/1.0)
6) (WD <= 181) and (WD >= 41) and (WD <= 41) and (BP >= 101326) and (AT >= 26.6) => All_class=low (16.0/2.0)
7) (WD <= 181) and (Rhi >= 53) => All_class=low (5.0/0.0)
8) (area_code >= 3) => All_class=high (325.0/0.0)
9) (area_code <= 1) and (WD <= 40) and (WS <= 0.3) and (Rhi >= 50) and (AT <= 25) => All_class=high (13.0/0.0)
10) (area_code <= 1) and (WD <= 40) and (WD >= 34) and (WS <= 0.5) => All_class=high (24.0/3.0)
11) (area_code <= 1) and (WD <= 40) and (WS <= 0.8) and (WD >= 40) and (BP <= 101306) => All_class=high (29.0/7.0)
12) (area_code <= 1) and (AT >= 35.3) and (AT <= 37.5) and (Rhi <= 41) => All_class=high (20.0/1.0)
13) (area_code <= 1) and (WS <= 0.3) and (WD <= 1) and (Rhi >= 50) and (AT <= 27.3) => All_class=high (34.0/9.0)
14) (area_code <= 1) and (AT >= 27.5) and (WS <= 0.6) and (AT >= 31.1) and (AT <= 33.9) => All_class=high (13.0/0.0)
⇨ All_class=med (874.0/171.0)
⇨ Number of Rules : 15

Fig. 5. JRip rules for all four UTHM locations (bold lettering is used to emphasise the **high** wood particulate concentration).

=== Stratified cross-validation ===		
=== Summary ===		
Correctly Classified Instances	1274	80.84%
Incorrectly Classified Instances	302	19.16%
Kappa statistic	0.6812	
Mean absolute error	0.188	
Root mean squared error	0.3204	
Relative absolute error	45.26%	
Root relative squared error	70.32%	
Total Number of Instances	1576	

=== Detailed Accuracy By Class ===

	TPRate	FPRate	Preci	Recall	FMea	ROC	Class
	0.566	0.036	0.785	0.566	0.658	0.847	low
	0.9	0.272	0.757	0.9	0.822	0.834	med
	0.812	0.033	0.923	0.812	0.864	0.908	high
WAvg.	0.808	0.149	0.816	0.808	0.805	0.861	

=== Confusion Matrix ===

a	b	c		<-- classified
168	127	2	|	a = low
43	687	33	|	b = med
3	94	419	|	c = high

Fig. 6. JRip stratified cross validation summary (WA: weighted average) all four location data.

B. J48 results of all four UTHM locations

J48 Rules as well confirm the JRip conditions for the different wood concentration classes for all four locations. The J48 cross validation shows 84% accuracy (fig. 7). The precision for class high is 90 % with a recall rate 86%. Class med has a precision of 80% with 89% recall, meanwhile, low has 83% and 67% of precision and recall respectively.

=== Stratified cross-validation === Summary ===		
Correctly Classified Instances	1324	84.01%
Incorrectly Classified Instances	252	15.99%
Kappa statistic	0.7375	
Mean absolute error	0.1401	
Root mean squared error	0.2955	
Relative absolute error	33.72%	
Root relative squared error	64.85%	
Coverage of cases (0.95 level)	96.26%	
Mean rel. region size (0.95 level)	51.73%	
Total Number of Instances	1576	

=== Detailed Accuracy === By Class ===

	TP Rate	FP Rate	Prec	Recall	F-Mea	MCC	ROC	PRC	Class
	0.67	0.03	0.836	0.67	0.74	0.7	0.92	0.78	low
	0.893	0.21	0.803	0.89	0.85	0.69	0.88	0.82	med
	0.86	0.04	0.906	0.86	0.88	0.83	0.96	0.91	high
WA	0.84	0.12	0.843	0.84	0.84	0.74	0.91	0.84	

=== Confusion Matrix ===

a	b	c	<-- classified as
199	95	3	| a = low
39	681	43	| b = med
0	72	444	| c = high

Fig. 7. J48 stratified cross validation summary – all four location data.

C. MLP results of all four UTHM locations

The MLP results (fig 8) show better precision and recall percentages than JRip for all four location data for the classes high and medium.

Stratified cross-validation Summary ===		
Correctly Classified Instances	1303	82.68%
Incorrectly Classified Instances	273	17.32%
Kappa statistic	0.708	
Mean absolute error	0.148	
Root mean squared error	0.271	
Relative absolute error	35.54%	
Root relative squared error	59.50%	
Coverage of cases (0.95 level)	99.37%	
Mean rel. region size (0.95 level)	54.15%	
Total Number of Instances	1576	

=== Detailed Accuracy By Class ===

P Rate	FP Rate	Preci	Recall	F-Mea	MCC	ROC	area	Class
0.38	0.021	0.807	0.38	0.517	0.49	0.918	0.756	low
0.923	0.261	0.769	0.923	0.839	0.67	0.919	0.898	med
0.942	0.032	0.935	0.942	0.938	0.91	0.995	0.989	high
WA0.827	0.141	0.83	0.827	0.811	0.72	0.944	0.901	

=== Confusion Matrix ===

a	b	c	<-- classified as
113	183	1	| a = low
26	704	33	| b = med
1	29	486	| c = high

Fig. 8. MLP stratified cross validation summary all four location data.

D. Data mining results of ORICC and Library

To further analyse the conditions that had led to high wood particulate concentrations at locations 3 and 4 (ORICC and library), the two location data is studied using 2D graphs, JRip, J48 and MLP classifiers. For the data mining algorithm analysis, the target variable is reclassified with OLlow <0.1, OLmed <0.5 and OLhigh>= 0.5 MGM3 based on the data distribution for these two locations (fig 9b).

A graph plotted for location 3 (ORICC) data shows the peaks in wood dust particulate concentration recorded for three days consecutively (19-21 June 2013) but at different times (fig 10). Thereafter, the concentration goes down to almost nil by 27 June 2013, just before the equipment was removed to a different location. On 19 June 2013, the highest concentration for this site was recorded (1.571 MG/M^3). This is over the recommended international limit for **hardwood** exposure to humans which is 1G/M^3 [4] even though more tests are required to establish the airborne wood dust composition of this timber factory.

The highest recorded concentration for location 4 library is 0.867 MG/M^3 on 20 June 2013 at 9.45 am (fig. 11). In the library area, the dust concentration seem to be peaking in the morning around 9.00 am. This has been observed for two consecutive days when the equipment was at this site. The seven sets of J48 tree conditions that had led to high (>0.5 MGM3) at ORICC and library are presented in fig 9a.

1; AT<=37.4°C,BP>100584<=101092P,WD<=178° and Rhi <=47%

2; AT<=37.4°C, BP >100584<=101092P,WD <=178>158° and Rhi >47%

3; AT<=37.4 <=27.0°C, BP >100584 <100759P and WD>178°

4; AT<=37.4 >28.5 °C, BP >100584 <=100720P, WD>178° and WS<0.9 for location 3 (ORICC)

5; AT<=37.4 <=27.1 °C, BP >100584 >101092P and Rhi>36%

6; AT<=37.4 >27.1 °C, BP >100584 >101092P and Rhi<=49%

Fig 9a: J48 tree rules generated for location 3 and 4 data (fig 13 for tree).

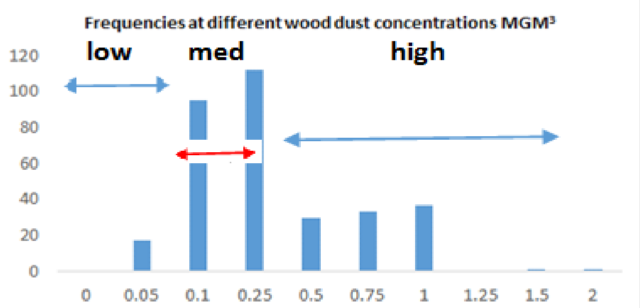

Fig. 9b. Airborne wood particulate concentration data distribution in Locations 3 and 4.

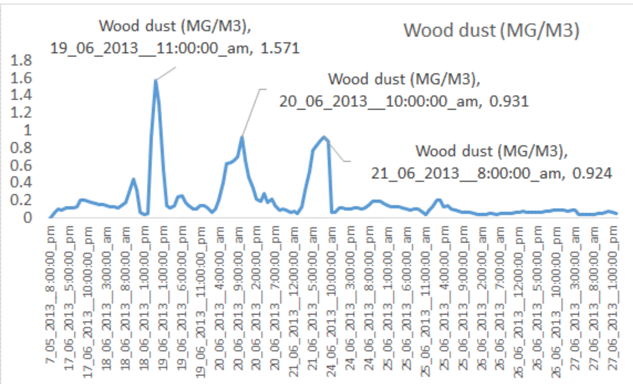

Fig. 10. Graph showing the variability in wood dust concentration collected at Location 3 (ORICC) from 17-27 June 2014 at hourly intervals.

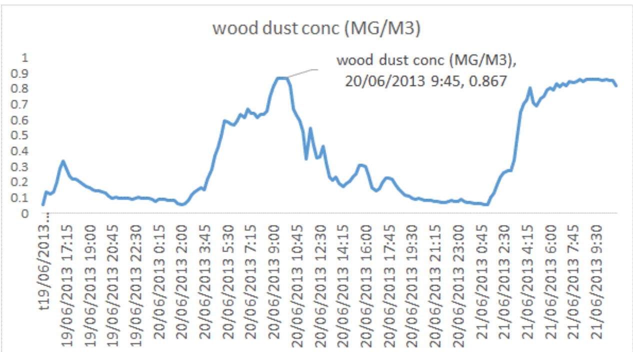

Fig. 11. Graph showing the variability in wood dust concentration collected at location 4 Library from 7 May -27 June 2014 at 15 min intervals.

Number of Leaves: 29
Size of the tree: 57
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
===Summary ===

Correctly Classified Instances	251	76.76%
Incorrectly Classified Instances	76	23.24%
Kappa statistic	0.6361	
Mean absolute error	0.1788	
Root mean squared error	0.3641	
Relative absolute error	41.69%	
Root relative squared error	78.64%	
Coverage of cases (0.95 level)	87.77%	
Mean rel. region size (0.95 level)	54.94%	
Total Number of Instances	327	

== Detailed Accuracy By Class==

	TPRate	FPRate	Preci	Recall	F-Mea	MCC	ROC	Area	Class
	0.768	0.126	0.761	0.768	0.764	0.641	0.86	0.73	low
	0.797	0.19	0.765	0.797	0.781	0.605	0.84	0.775	med
	0.708	0.055	0.785	0.708	0.745	0.678	0.86	0.703	high
WA	0.768	0.138	0.768	0.768	0.767	0.633	0.85	0.74	

a	b	c	<-- classified as
86	21	5	| a = low
20	114	9	| b = med
7	14	51	| c = high

Fig. 12: J48 stratified cross validation summary of location 3 and 4

From the J48 tree rules, it is can be stated that the ambient temperature (AT) has been <= 37.4 °C when wood particulate concentrations at ORICC were high (OLhigh >0.5 MGM3). This also reveals that the wind coming from SSE (178°) at speeds 0.9 meters/sec generates OLhigh wood dust scenarios at ORICC. The accuracy achieved for the J48 classification is 76%. Precision and recall for all three classes are over 75% (fig 12).

V. Conclusions

In light of the recommended limit for hardwood exposure for humans, the paper looked at the wood particulate concentrations at four different locations within UTHM situated next to a wood processing factory. In this first ever study on the wood dust at this location, the observational data of E-Sampler collected over consecutive days at the locations show the levels of likely exposure to particulate dust endured by UTHM staff and students. In summary, of the four locations studied, location 3 (ORICC) located literally next to the timber processing factory has been the most exposed location. The highest concentration recorded (1.57 MG/M3) at this location is over the recommended exposure limit (1.0 MG/M3) for humans, if the dust particles were from **hardwood** hence further studies are recommended to establish the type of airborne wood particles released from the factory. The concentrations measured near the Library area as well have been almost nearing the recommended limit for **hardwood**.

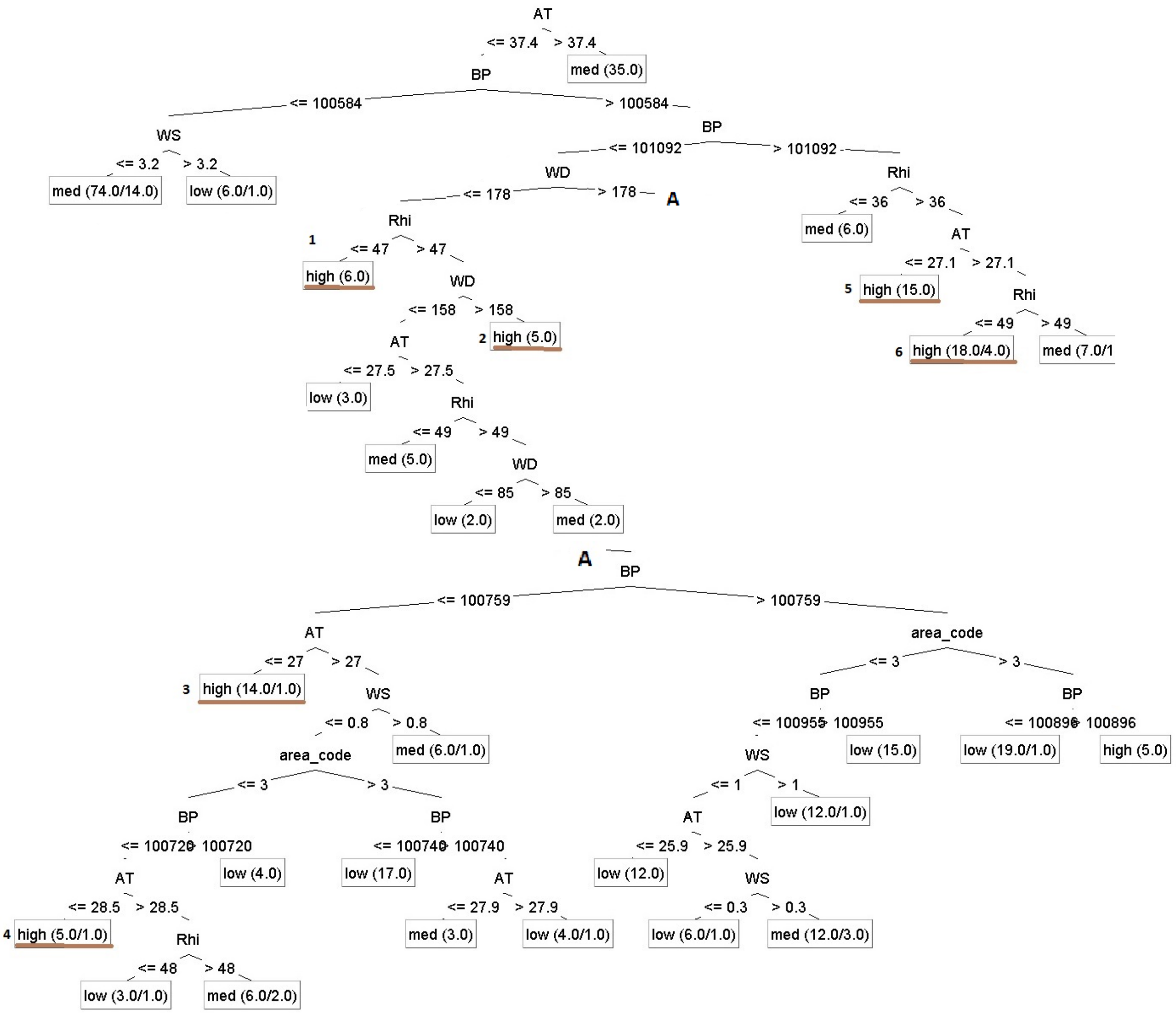

Fig. 13. J48 tree created for UTHM locations 3 (ORICC) and 4 (library). The six OLhigh(>0.5MG/M^3) are underlined. See fig 9a for rules for class **high**.

ACKNOWLEDGMENTS

The authors acknowledge the Faculty of Engineering staff who lent the E-Sampler equipment for this work.

REFERENCES

[1] K.-S. Patrycja, W. Marta , P. Cezary, N.-Ś. Ewa, K. Anna and W.-S. Jolanta, "Occupational asthma caused by samba (Triplochiton scleroxylon) wood dust in a professional maker of wooden models of airplanes:A case study," International journal of occupational medicine and environmental health ISSN 1896-494X vol. 27 no3, pp512-519, 2014.

[2] X. Simon, D. Bémer, S. Chazel and D. Thomas, "Downstream particle puffs emitted during pulse-jet cleaning of a baghouse wood dust collector: Influence of operating conditions and filter surface treatment," Powder Technology Volume 261, July 2014,

DOI: 10.1016/j.powtec.2014.04.028, p. 61–70, 2014.

[3] E. D. Bruschweiler, N. B. Hopf, P. Wild1, C. K. Huynh, M. Fenech, P. Thomas, M. Hor, N. Charriere, D. Savova-Bianchi and B. Danuser, "Workers exposed to wood dust have an increased micronucleus frequency in nasal and buccal cells: results from a pilot study," Mutagenesis vol. 29 no. 3, no. Published by Oxford University Press on behalf of the UK Environmental Mutagen Society. pp 201–207, 2014.

[4] T. W. B. Thomas L. Bean, "Wood Dust Exposure Hazards AEX-595.1-2006 (Revised)," Ohio State University Extension, Food, Agricultural and Biological Engineering, 590 Woody Hayes Dr., Columbus, Ohio 43210, [Online]. Available: ENVIRONMENT AND HEALTH. [Accessed 22 August 2014].

[5] Met One Instruments, "E-SAMPLER Particulate Monitor Operation Manual E-SAMPLER-9800 Manual Rev J [1].doc," Met One Instruments, Inc, http://www.metone.com/documents/E-SAMPLER_Brochure.pdf. Accessed 13 6 2013

[6] Machine Learning Group, "Weka 3: Data Mining Software in Java," University of Waikato, New Zealand, [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/index.html. [Accessed 2014].

[7] J. R. Quinlan, "Improved use of continuous attributes in c4.5.," Journal of Artificial Intelligence Research, vol. 4, pp. 77-90, 1996.

[8] WEKA, "Class MLPClassifier,"

http://weka.sourceforge.net/doc.packages/multiLayerPerceptrons/weka/classifiers/functions/MLPClassifier.html. Accessed 22 August 2014