SUPPLEMENTAL INFORMATION

TABLE A I

HTL CONDITIONS AND GLUCOSE OUTPUT AS MEASURED BY HPLC AT DIFFERENT RETENTION TIMES FOR RAW AND PRETREATED WOOD RESIDUE.

Sample number	Temperature / ° C	Pressure / bar	Retention time / seconds	Glucose / wt%
Raw				
1	210	2	60	0
2	210	2	180	0
3	210	2	360	0
4	210	2	540	0
5	210	2	720	0
6	210	2	900	0
7	210	60	60	0
8	210	60	180	0
9	210	60	360	0.537
10	210	60	540	0.499
11	210	60	720	0.420
12	210	60	900	0.412
Alkaline pretreatment				
13	210	2	60	0
14	210	2	180	0
15	210	2	360	0
16	210	2	540	0
17	210	2	720	0
18	210	2	900	0
19	210	60	60	0.164
20	210	60	180	0.847
21	210	60	360	0.679
22	210	60	540	0.629
23	210	60	720	0.535
24	210	60	900	0.405

TABLE A II

HTL CONDITIONS AND GLUCOSE OUTPUT AS MEASURED BY HPLC AT DIFFERENT RETENTION TIMES FOR PROCESSED COTTON.

Sample number	Temperature / ° C	Pressure / bar	Retention time / seconds	Glucose / wt%
1	130 to 140	133	60	0
2	150 to 160	133	60	0
3	160 to 170	133	60	0
4	170 to 180	133	60	0
5	180 to 190	133	60	0
6	190 to 200	133	60	0
7	200 to 210	133	60	0.14
8	210 to 220	133	60	0.42
9	220 to 228	133	60	0.8
10	230 to 236	133	60	1.39
11	210	111	180	1.11
12	210	111	360	0.94
13	210	111	540	0.85
14	210	111	720	0.76
15	210	111	900	0.71
16	220	110	60	1.99
17	220	110	180	2.57
18	220	110	360	1.91
19	220	110	540	0
20	220	110	720	0
21	220	110	900	0
22	230	123	60	3.24
23	230	123	180	2.68
24	230	123	360	0
25	230	123	540	0
26	230	123	720	0
27	230	123	900	0
28	240	114	60	0.91
29	240	114	180	1.6
30	240	114	360	1.33
31	240	114	540	1.17
32	240	114	720	1.01
33	240	114	900	0.93
34	250	108	60	0
35	250	108	180	0
36	250	108	360	0
37	250	108	540	0
38	250	108	720	0
39	250	108	900	0

TABLE A III

DISSOLVED GLUCOSE IN DISTILLED WATER AND THE CONCENTRATIONS USED TO CREATE ATR-FTIR SPECTRUM DATASET.

Sample number	Glucose / wt%
1	0.011
2	0.9
3	0.9
4	0.9
5	0.45
6	0.45
7	0.45
8	1.8
9	1.8
10	1.8
11	1.125
12	1.125
13	1.125
14	2.25
15	2.25
16	2.25
17	3.6
18	3.6
19	3.6
20	4.5
21	4.5
22	4.5
23	5
24	5
25	5
26	5.5
27	5.5
28	5.5
29	6.5
30	6.5
31	6.5
32	7.5
33	7.5
34	7.5
35	8.5
36	8.5
37	8.5
38	9
39	9
40	9

TABLE A IV

TESTING THE CLASSIFIER STABILITY ON DIFFERENT DATASETS PERMUTATION AND PROCEDURES PRIOR TO SELECTING THE SUPPORT VECTOR MACHINE CLASSIFIER AND DEEPENING THE STUDY. FEATURE ENGINEERING INCLUDED THE STATISTICAL FEATURES: MEAN (M); STANDARD DEVIATION (ST), VARIANCE (V), SKEWNESS (SK), KURTOSIS (K).

Test ID	**Ada Boost Classifier / Accuracy (%)**	**GradientBoosting Classifier / Accuracy (%)**	**Random Forest Classifier / Accuracy (%)**	**K Neighbors Classifier / Accuracy (%)**	**Support Vector Machine Classifier / Accuracy (%)**	**Logistic Regression / Accuracy (%)**	**Dataset and procedure applied**
1	62.50	50.00	62.50	62.50	62.50	75.00	C
2	93.80	88.00	100.00	100.00	100.00	100.00	C + DG
3	92.30	84.60	61.50	76.90	84.60	84.60	W + C
4	90.50	71.40	90.50	90.50	81.00	85.70	W + C + DG
5	96.88	90.62	75.00	87.50	90.62	90.62	C + GAN_C
6	84.31	86.27	74.51	84.31	82.35	86.27	W + GAN_W + C + GAN_C
7	84.34	86.75	78.31	89.16	89.16	87.95	W + GAN_W + C + GAN_C + DG+ GAN_DG
8	90.48	90.48	80.95	85.71	90.48	76.19	(W + C + DG) + (M + St + V + Sk + K)
9	80.95	85.71	90.48	80.95	90.48	90.48	(W + C + DG) + (M + St + V + Sk + K) + UMAP (n_neighbours=65)
10	95.16	90.32	90.32	90.32	95.16	96.77	Synthetic data only (W + C + DG) + (M + St + V + Sk + K)
11	91.57	96.39	90.36	92.70	97.59	93.98	(W + C + DG) + GAN + (M + St + V + Sk + K)

TABLE V

SVC ACCURACIES FOR TEST AND TRAIN DATASETS AND THEIR STANDARD DEVIATIONS (SD). ABBREVIATIONS STAND FOR W – WOOD, C – COTTON, DG – DISSOLVED GLUCOSE.

Dataset Abbreviation	Accuracy test / %	SD	Accuracy train / %	SD
W	92.00	10.33	86.32	2.72
C	80.00	12.08	94.19	7.87
DG	91.25	13.24	98.44	3.38
W+GAN_W	84.50	8.32	98.16	2.34
C+GAN_C	86.25	4.47	99.60	0.43
DG+GAN_DG	89.38	3.02	98.98	2.68
W+C	80.00	9.03	88.80	7.44
W+DG	89.23	7.43	97.65	2.23
C+DG	80.63	11.20	95.08	5.26
W+GAN_W+C+GAN_C	87.84	3.04	99.45	0.76
W+GAN_W+DG+GAN_DG	90.58	3.79	99.75	0.26
C+GAN_C+DG+GAN_DG	90.00	1.98	99.33	0.82
W+C+DG	84.29	8.41	95.49	5.95
W+C+DG+GANl	87.65	5.23	99.20	0.71
W+GAN_W+C+GAN_C+DG+GAN_DG	90.84	3.37	99.24	1.35

TABLE VI

SELECTION OF THE MOST PERFORMANT DATASETS BASED ON THE ACCURACY VALUE FROM THE ABLATION STUDY. AVERAGE STANDARD DEVIATIONS (SD), PRECISION AND RECALL ARE ALSO PRESENTED.

Dataset permutation	Accuracy / %	SD	Precision	Recall
Hybrid datasets with hand-crafted features from ablation study – A1				
M	91.80	2.88	0.9737	0.8981
M+St	91.56	2.78	0.9792	0.8869
M+St+Sk	92.40	2.12	0.9762	0.9091

Hybrid datasets with hand-crafted features from ablation study – A2 –with UMAP on				
M+St+V+K	91.08	2.73	0.9427	0.9135
St+V+K	90.84	2.67	0.9289	0.9150
St+V+Sk	90.96	3.64	0.9237	0.9254
Hybrid datasets with hand-crafted features from ablation study – A3				
M+St+K	93.01	2.39	0.9469	0.9320
M+St+Sk+K	91.68	2.00	0.9251	0.9265
M+V+K	**93.49**	**2.35**	**0.9588**	**0.9313**
St+Sk	91.56	3.85	0.9369	0.9228
St+Sk+K	92.28	2.28	0.9190	0.9421
St+V	91.92	2.84	0.9377	0.9196
Hybrid datasets with hand-crafted features from ablation study – A4 with UMAP on				
M+Sk	81.56	4.50	0.8873	0.8070
M+Sk+K	82.28	4.58	0.8804	0.8080
M+St	82.04	3.47	0.8582	0.8356
Sk+K	82.40	2.28	0.8880	0.8129
St+Sk+K	81.92	4.71	0.8663	0.8286
M	83.01	4.19	0.8790	0.8210

MATERIALS AND METHODS

A. Experimental Machine Set Up

The HTL system encompassed a continuous flow delivered by a high-pressure pump. It was equipped with two heater systems, one for the flowing media, which reduced the temperature difference before entering the reactor. The second heater was around the reactor chamber and it was directly controlling the biomass decomposition temperature. A series of gauges, valves and electronic instruments controlled and monitored the pressure and temperature.

B. Biomass Samples

1) **Samples Dataset A: wood**

Raw and alkaline pretreated *Dipetrocarpus caudatus* wood dust was used as a model for lignocellulosic biomass material. The wood waste was sourced from the Brunei Timber Sawmill (Tutong, Brunei Darussalam). 1.8 grams of material was loaded in the reactor cell for each experimental run. The tests were carried at one constant temperature of 210 °C and two different pressure conditions, 2 bars and 60 bars (Table A I, Supplemental information).

2) **Samples Dataset B: cotton**

Cotton was processed under HTL as a substitute replicating lignocellulosic biomass to obtain fermentable sugars. For each test, a weight of 0.40 grams of cotton was loaded into the HTL's cylindrical reactor. The resulting dataset included 5 distinctive batches. A ramp up study, in which the aqueous phase was collected between 130 °C to 236 °C for a period of 60 seconds, was also included in the dataset. The other 4 individual batches at 210, 220, 230 and 240 °C constant temperature maintained by the PID controller were collected for a period of 60 seconds every other 180 seconds (Table A II, Supplemental information).

3) **Samples Dataset C: dissolved glucose**

Dilute compounds in aqueous solutions have complex and broad spectra, which complicates the distinguishing of individual compounds. Water presence in the sample can also interfere with the transmittance spectrum. Approaches to solve this problem include spectra correction or modification of the analytical methodology. However, this requires additional data processing that is variable with each sample. The dataset generated from the HTL was small, totalling 63 samples, of which 21 were above the 50% NERs. This meant that the rest of the samples had a high-water content. Additionally, apart from the reduced sample numbers, training GANs only on the underrepresented classes can lead to mode collapse. The dissolved glucose spectrums represented the target material and were introduced as a counter-measure. The glucose solutions were used to build the dataset from spectrums by ATR-FTIR analysis. D-(+)-Glucose with a purity ≧98.0% was purchased from Nacalai Tesque. 0.90 grams of glucose (180 g/mol) was added to 10 mL of deionized water, and subsequent dilutions were produced at lower concentrations

(Table A III, Supplemental information). Dissolved glucose dataset contained 36 out of 39 samples above the threshold required to attain a NER bigger than 50%. The datasets showed distinctive features under ATR-FTIR, especially around the finger-print region between 800 cm^{-1} to 1200 cm^{-1} (Fig. A I) No solid glucose spectrums were used for training the machine learning models.

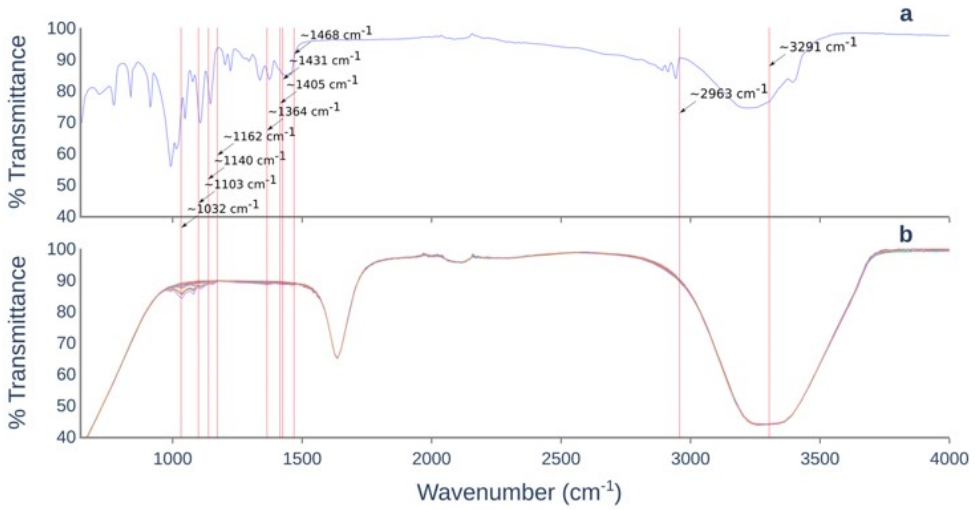


Fig. A I (**a**) Solid glucose - average of 5 spectrums. (**b**) Dissolved glucose in distilled water at various concentrations (see Table A III in Supplemental information).

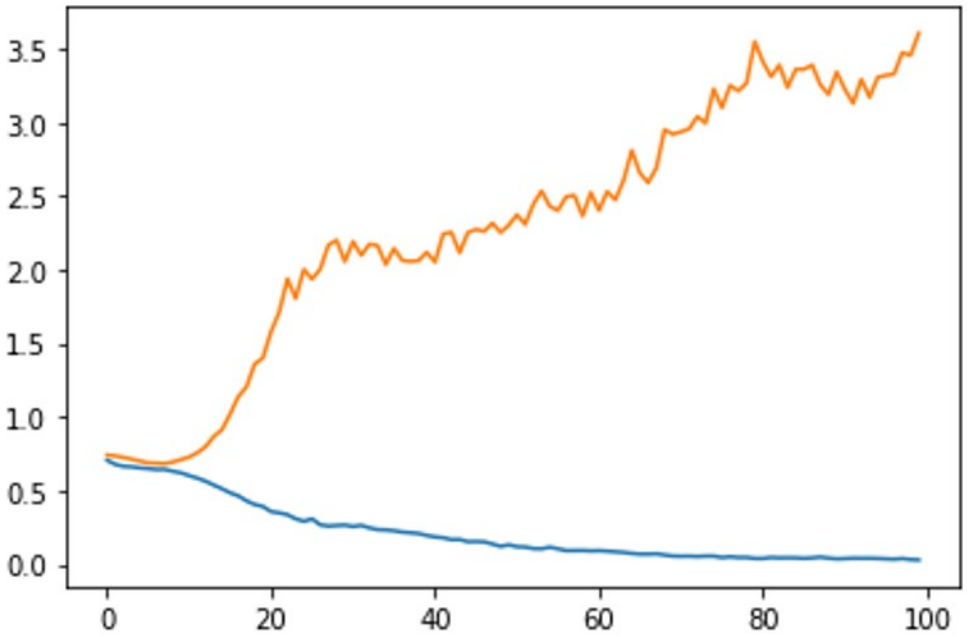
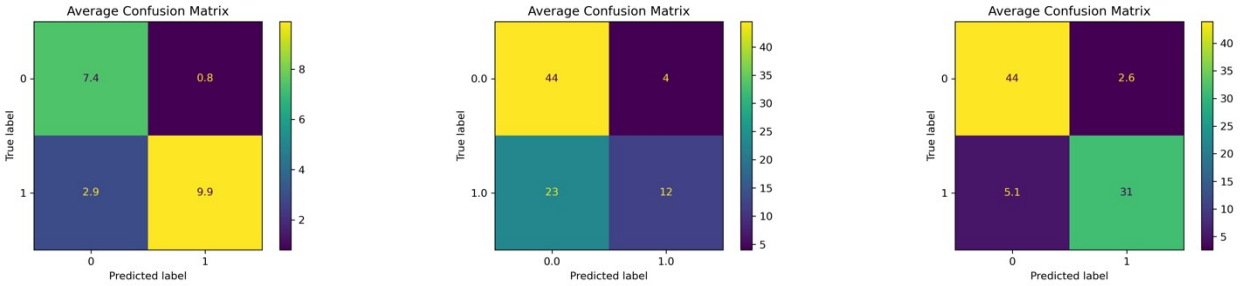


Fig. A II GAN loss curve; x-axis represent the number of epochs; y-axis represent the loss values for the generator and discriminator. Blue - generator loss; Orange - discriminator loss. Steady-state convergence is achieved, by the smooth decreasing curve of the generator, and the discriminator loss increasing.



(a) (b) (c)

Fig. A III Average confusion matrices over 10 SVC classifying iterations for (**a**) W+C+DG; (**b**) W+C+DG+GAN (Posterior GAN); (**c**) W+GAN_W+C+GAN_C+DG+GAN_DG (Interstitial GAN).

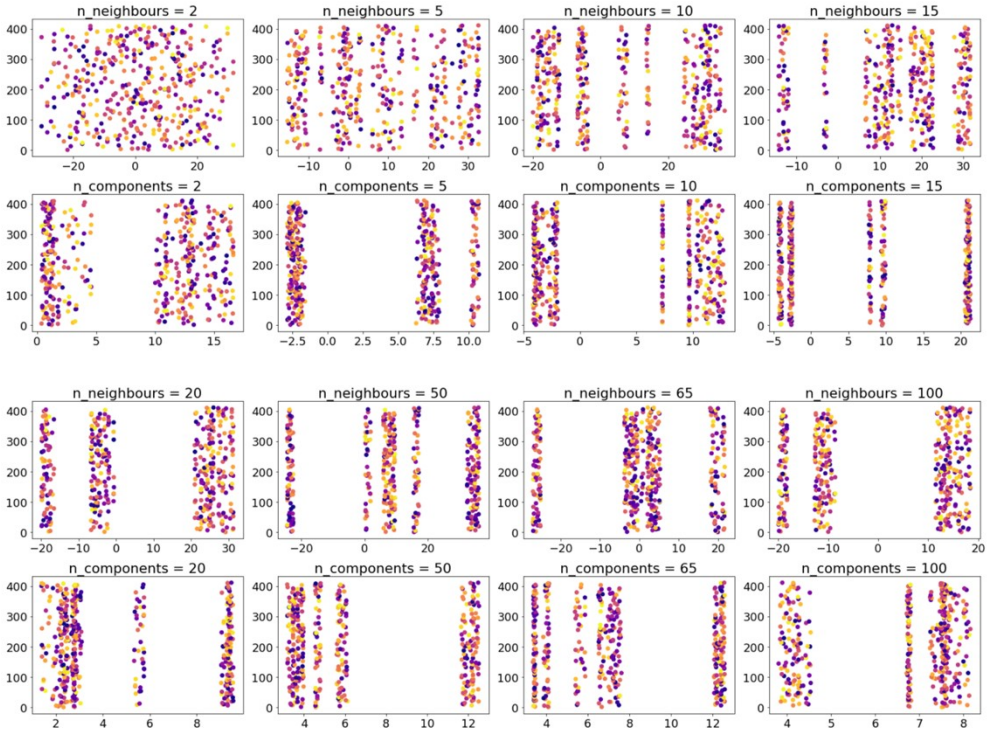


Figure A IV UMAP hyperparameters selection example for W + C + DG + GAN + features. Plot comprising n_neigbours and n_components: 2,5,10,15,20,50,65,100.