Classification of Asbestos on Multispectral Data

Marwa Al Hajriy

Supervisor: Dr. Riccardo Rastelli

Asbestos

- Fibrous minerals
- Properties: Durability, heat & fire resistance, cheap
- Construction building materials include
  roof tiles, cement, insulation

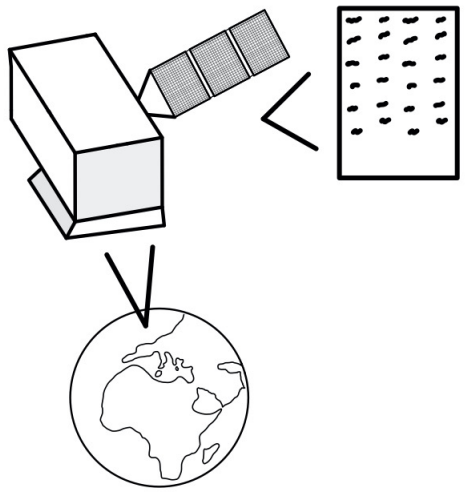
https://www.danmarque.co.uk/is-asbestos-roofing-harmful/

https://www.mesothelioma.com/asbestos-exposure/products/asbestos-sheets/

Diseases associated with Asbestos Exposure

- Malignant Mesothelioma
- Lung Cancer
- Cancer of the digestive system (stomach & oesophagus)
- International Agency of Research on Cancer (IARC) classified Asbestos Containing Materials (ACM) as Group 1 carcinogen in 1977
- Banning legislations 1980s

Satellite Data - Sentinel 2



- Spatial Resolution: smallest size in m per pixel that can be captured
- Spectral Resolution (SR): range of different light wavelengths detected

  Sentinel 2: Visible, Visible & Near Infrared (VNIR), Shortwave Infrared (SWIR)
- Central Wavelength (CW): The average wavelength in nm detected by each band

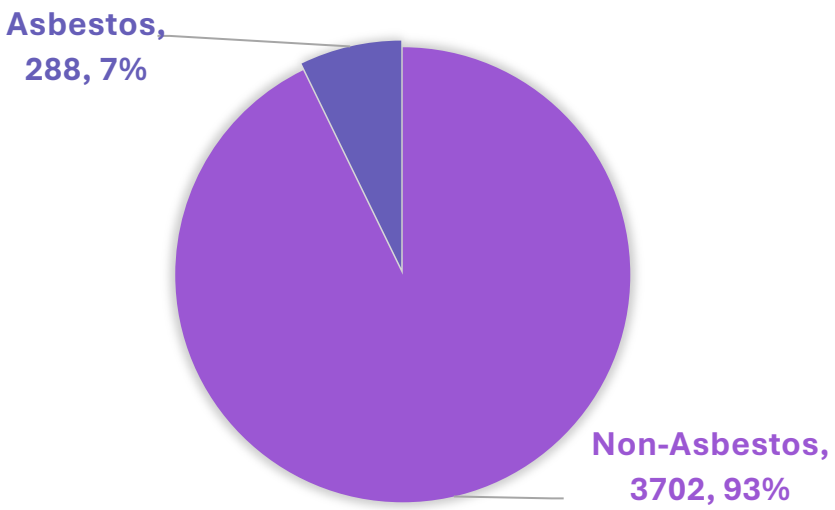
10m Spatial Resolution – RGB bands, Plants & Soil				
Band	B2	B3	B4	B8
SR CW	Blue 490nm	Green 560nm	Red 665nm	NIR 842nm

60m Spatial Resolution – Atmospheric Aerosol		
Band	B1	B9
SR CW	Blue 443nm	VNIR 945nm

20m Spatial Resolution – Ice, Cloud & Plant stress						
Band	B5	B6	B7	B8a	B11	B12
SR CW	Red Edge 705nm	Red Edge 740nm	Red Edge 783nm	NIR 865nm	SWIR 1 1610nm	SWIR 2 2190nm

Preprocessing

Data



Asbestos : Non-Asbestos

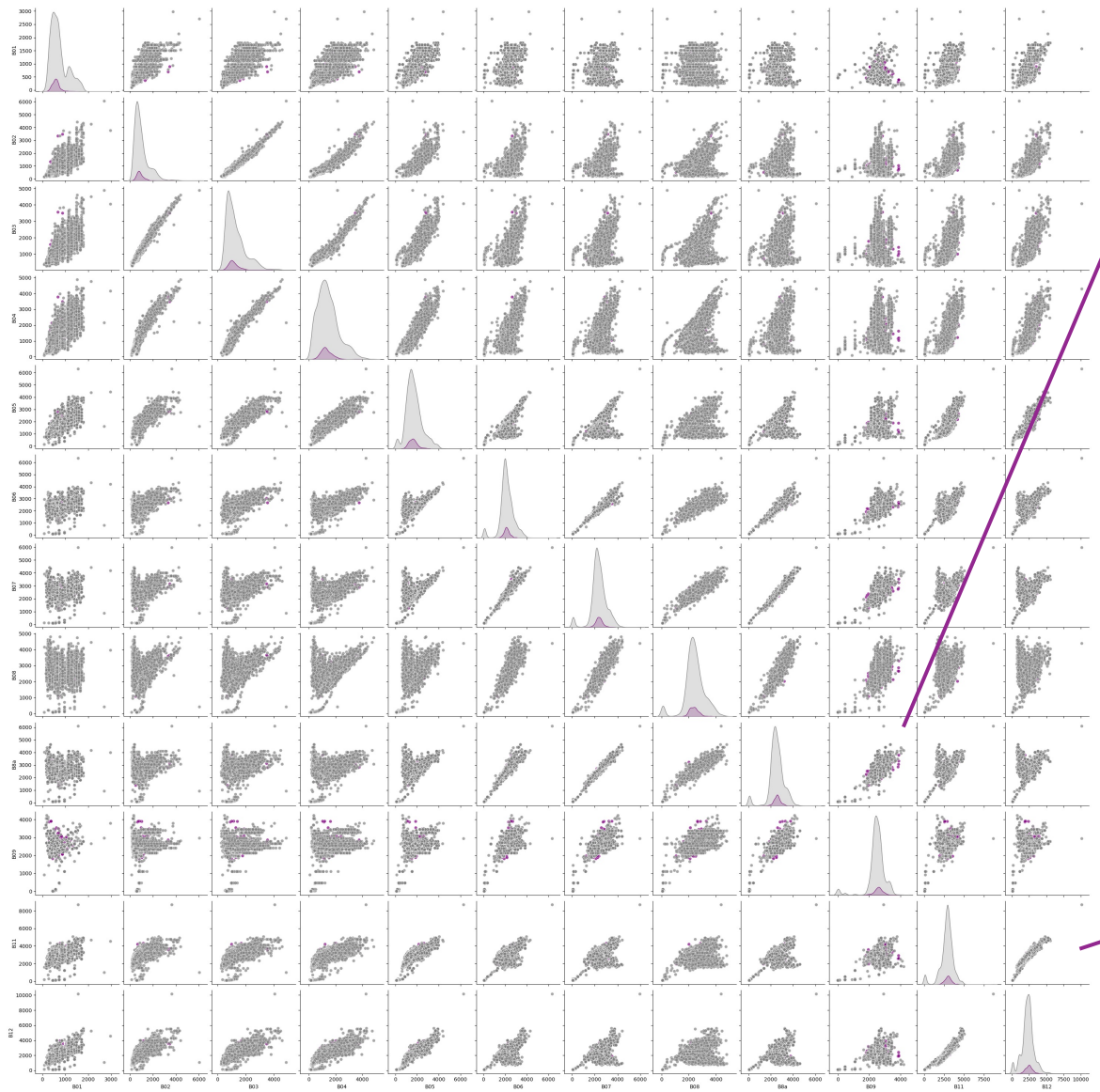
1:13

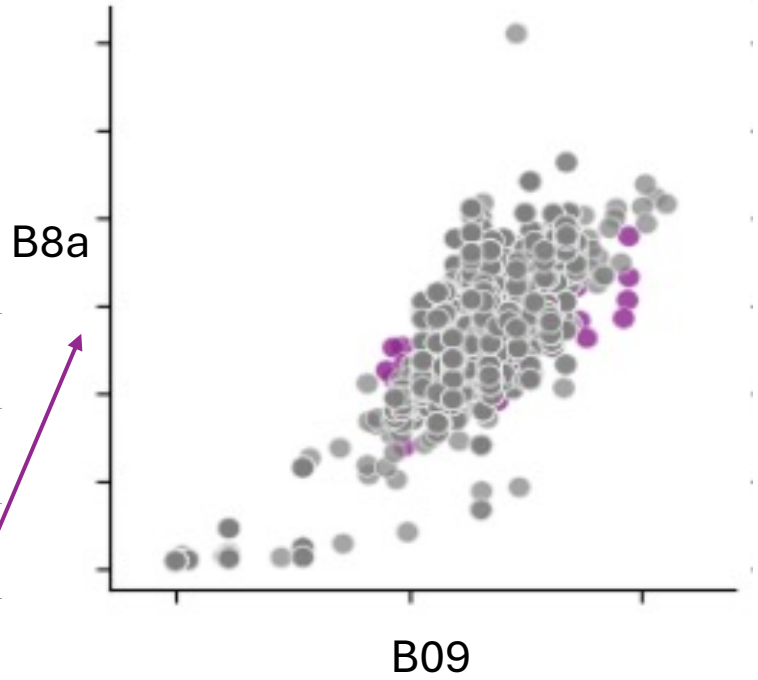
	B01	B02	B03	B04	B05	B06	B07	B08	B8a	B09	B11	B12	asbestos
0	551	792	1122	1262	1552	2301	2579	2526	2739	2592	2991	2558	1
1	533	1154	1542	1670	1602	1988	2241	2723	2335	2402	2684	2064	1
2	453	1270	1404	1712	1865	2139	2263	2159	2530	2464	3368	2921	1
3	718	981	1278	1498	1730	1962	2011	2308	2425	2388	3555	2893	1
4	582	765	976	1140	1294	1619	1781	2008	1987	2169	3011	2570	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
3985	292	315	493	318	764	2301	2959	3041	3397	2931	1798	918	0
3986	292	241	451	432	991	2066	2432	2690	2729	2931	2061	1286	0
3987	292	557	734	833	991	2066	2432	2740	2729	2931	2061	1286	0
3988	165	343	433	358	750	2407	3162	2500	3328	3684	1920	1037	0
3989	165	273	444	345	750	2407	3162	3759	3328	3684	1920	1037	0

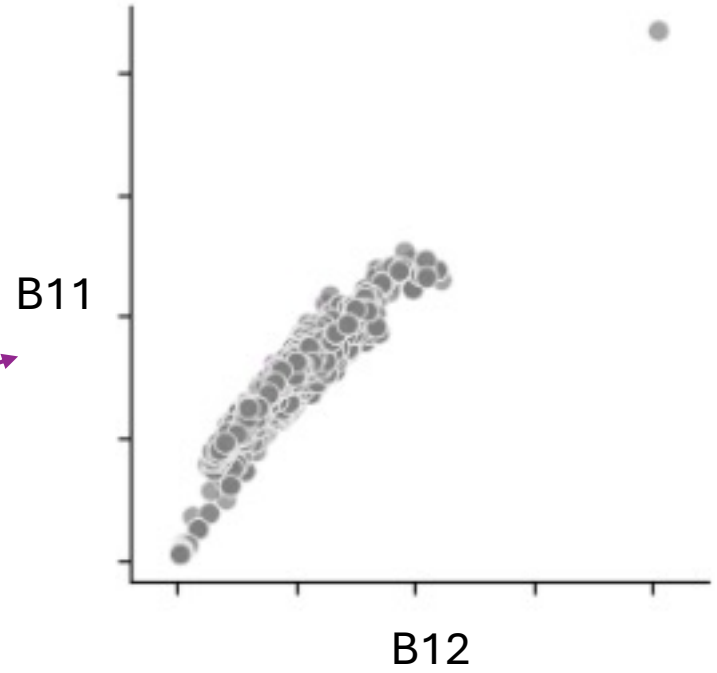
3990 rows × 13 columns

Data Visualization

Pairplot


B01
B02
B03
B04
B05
B06
B07
B08
B8a
B09
B11
B12
asbestos
0
1

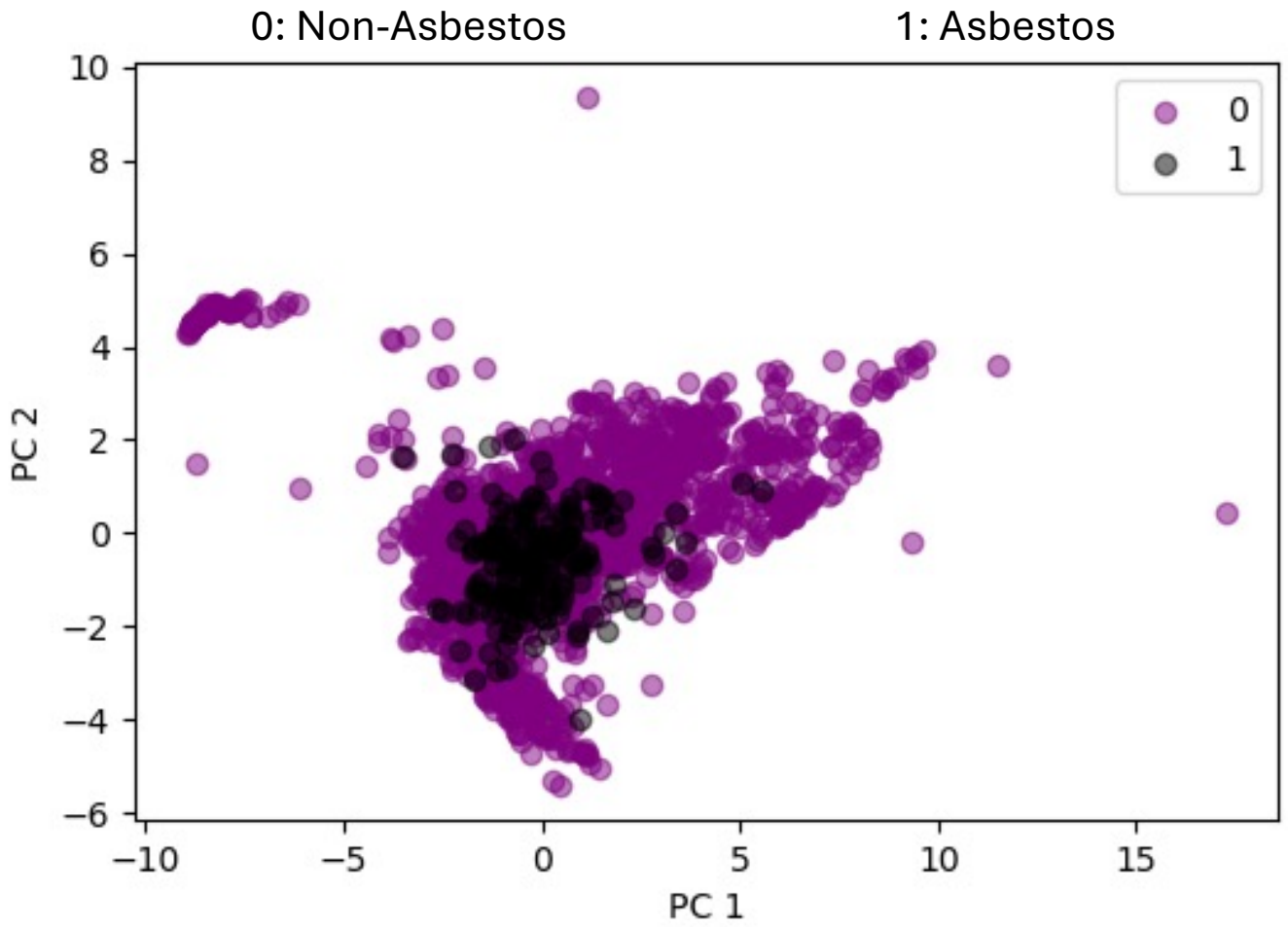

B8a
B09


B11
B12

Principle Component Analysis (PCA)

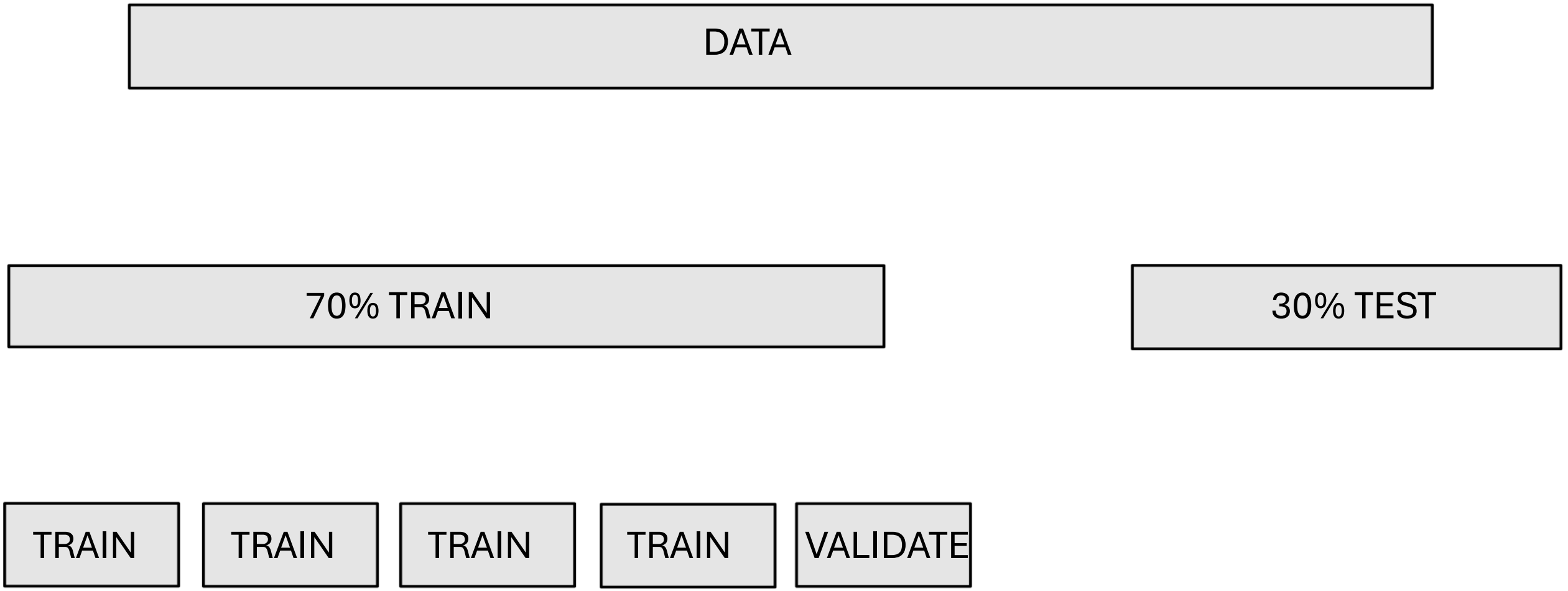
- Total: 87%

- PC1 66% of variance
- PC2 21% of variance



Splitting data & Cross Validation (CV)

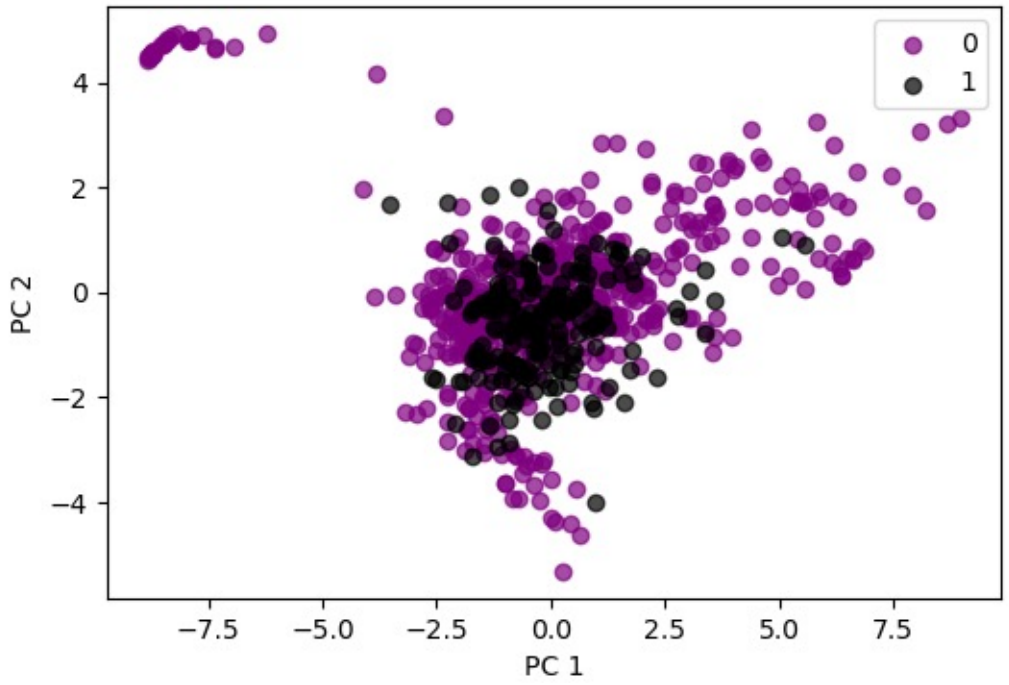
Stratified 5 fold CV

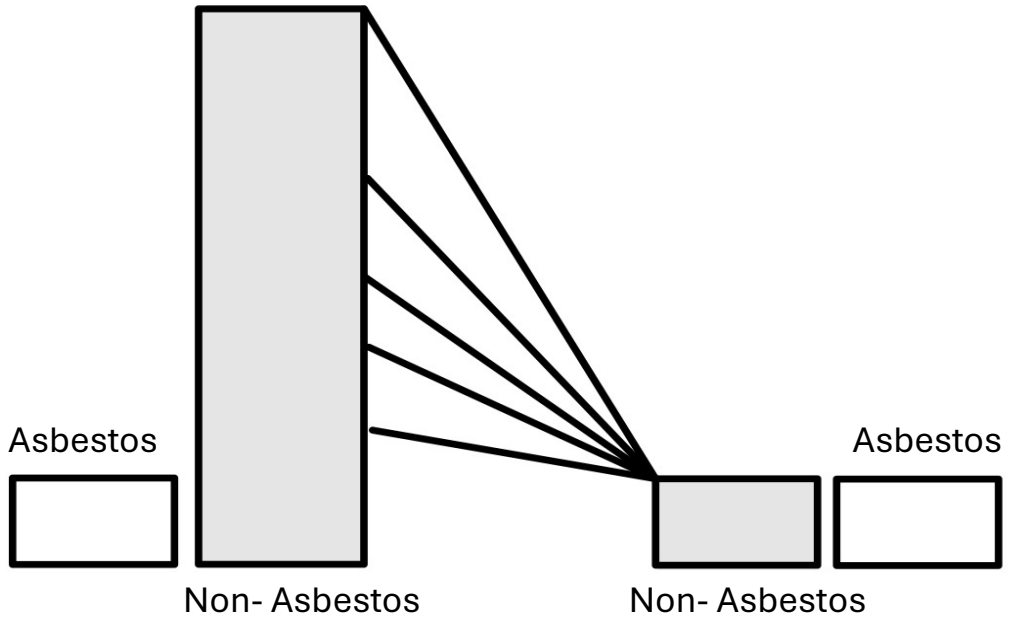


Handling Imbalance

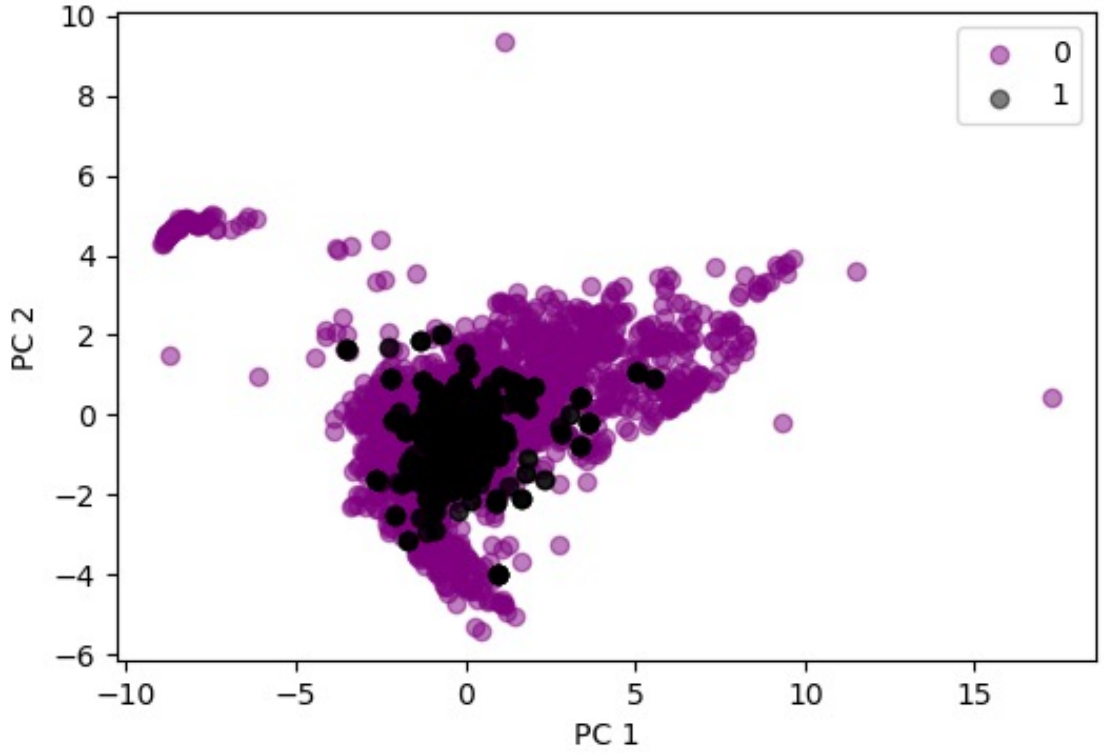
Undersampling

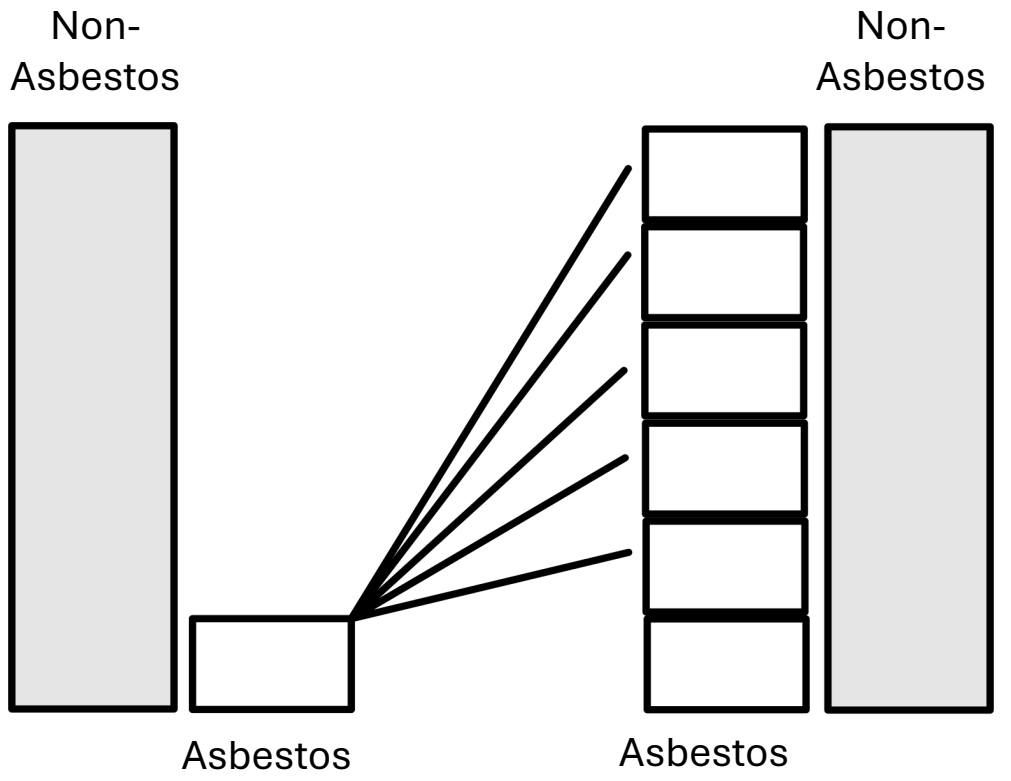
- Majority Class: Non-Asbestos
- Minority Class: Asbestos





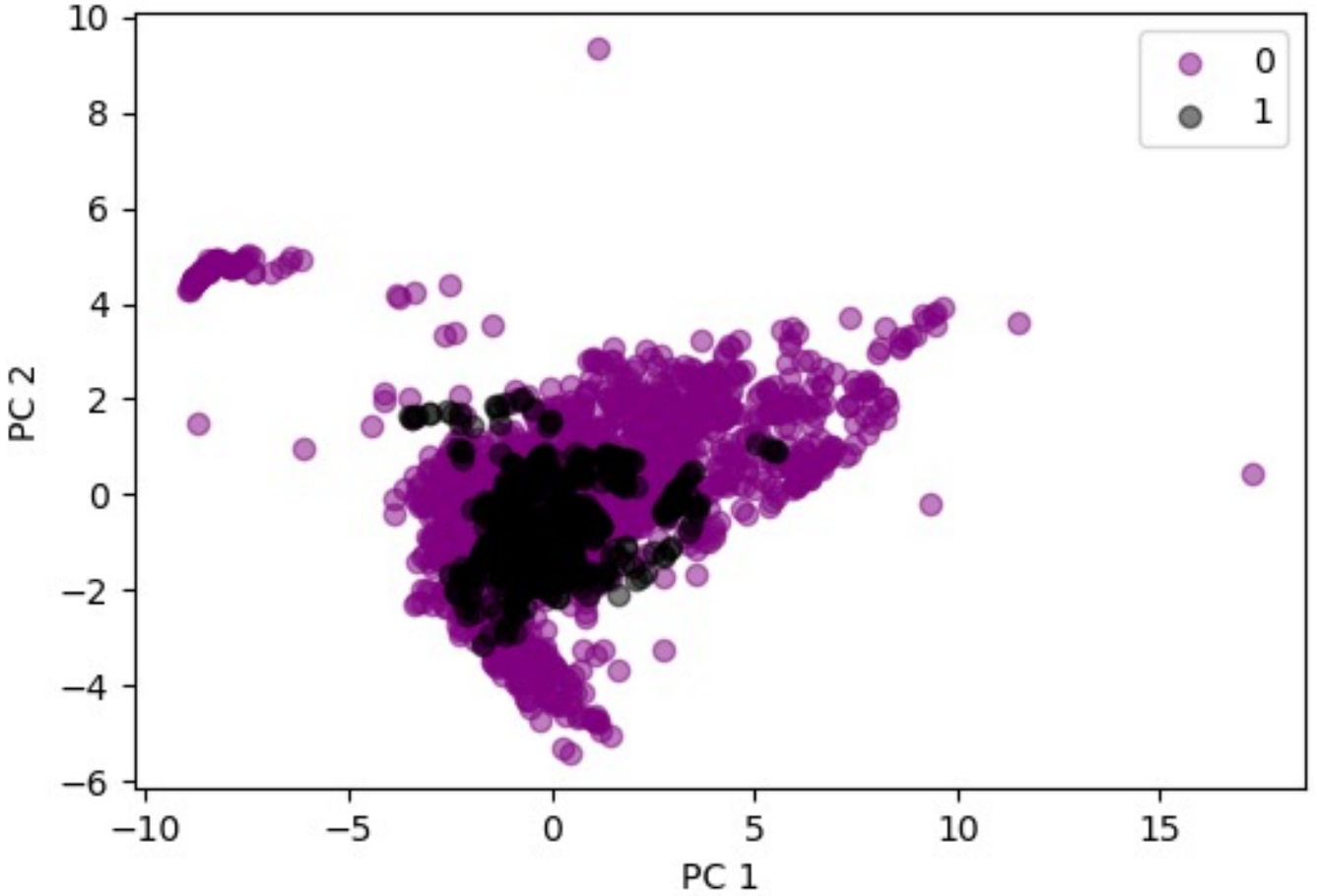
Asbestos : Non-Asbesos
1:3





Oversampling

Combination of Oversampling and Undersampling



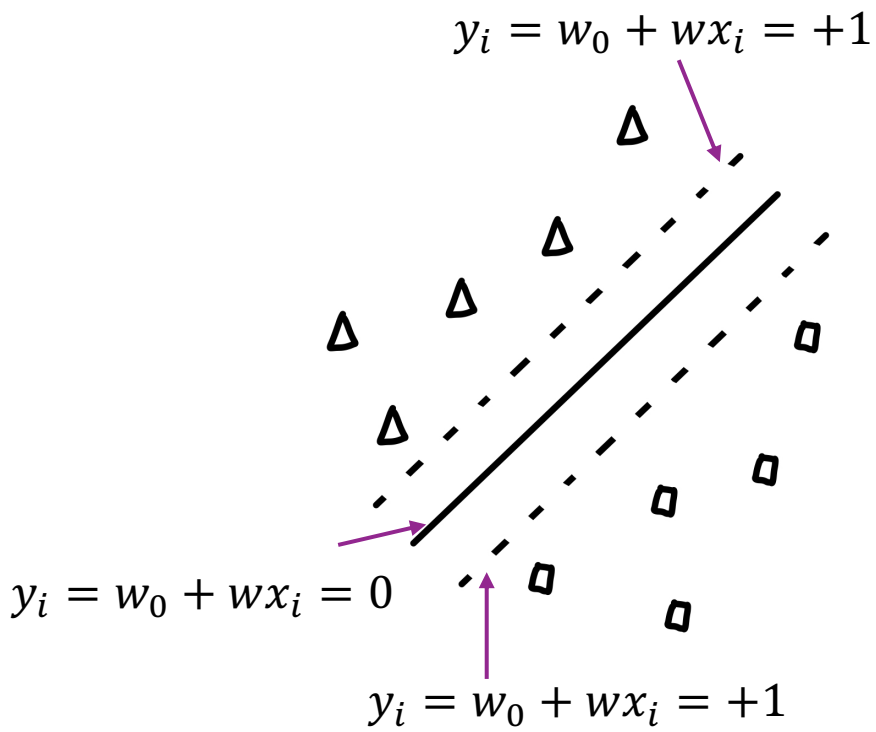
Models

Support Vector Machines (SVM)

- Finds optimum separating hyperplane by maximizing margin
- Kernel trick: kernel function enables nonlinear classification
- Optimization:

$$\min_{\{w,w_0\}} \frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{i=1}^{N}\xi_i$$

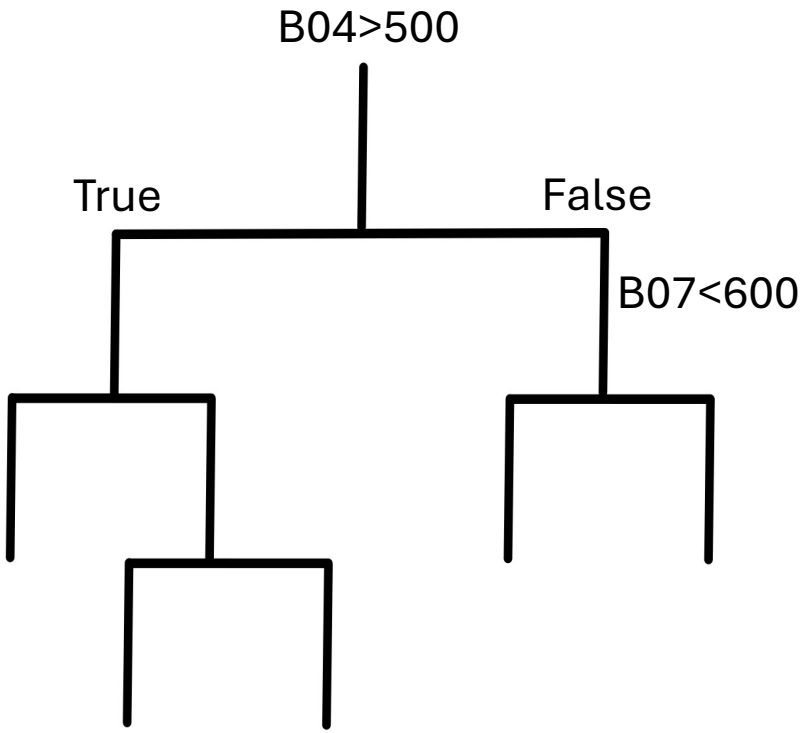
$$\text{s.t.} \quad y_i\left(w_0 + \sum_{i=1}^{N} w_i K(x,x_i)\right) \geq 1-\xi_i \qquad \xi_i \geq 0\ \forall i$$



C: cost hyperparameter, penalizes misclassification
ξ_i: slack variable, allows misclassification
$K(x,x_i)$: kernel function

Radial Basis Function $\quad K(x,x_i) = \exp\left(-\frac{\left|\left|x-x_i\right|\right|^2}{2\sigma^2}\right)$

Random Forest (RF)



- Decision trees perform classification through true or false conditions

- Training involves the minimization of Gini Impurity:

$$G = \sum_{i=1}^{c} p_i(1 - p_i)$$

where i=1,..c is the number of categories

p_i is the proportion of data points in category i

- RF is an ensemble method combines multiple Decision Trees & average out the results

- Less sensitive to multicollinearity

Artificial Neural Network (ANN)

- Interconnected neurons that mimic human brain

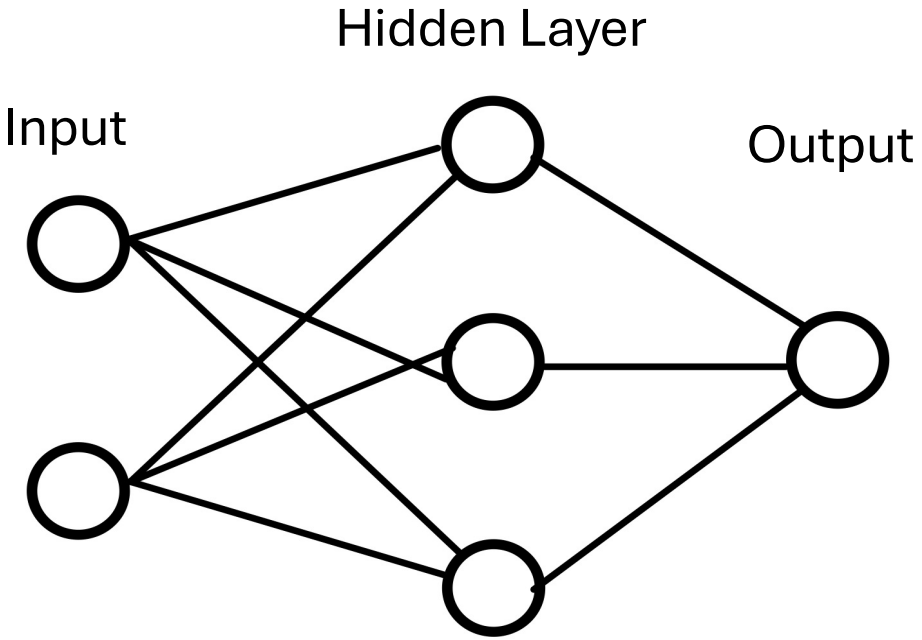
- Weights between neurons are modified until the loss function is minimized

$$\hat{y}_i = ReLU(wx_i + w_0)$$

$$ReLU(z) = \max(0, z)$$

Cross Entropy loss

$$L(y_i, \hat{y}_i) = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$



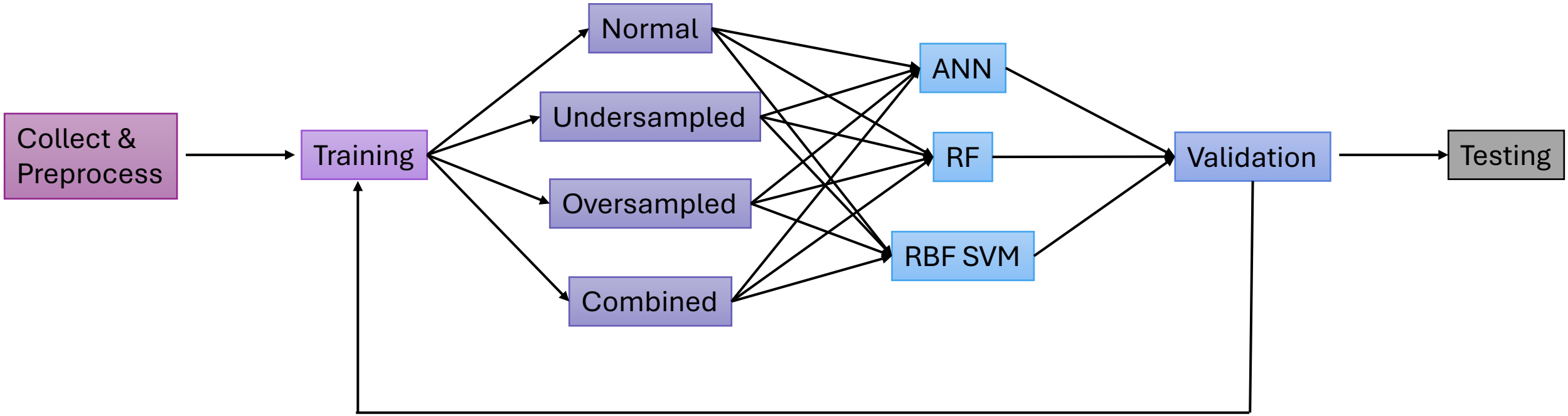
- Works well with complex nonlinear patterns

Performance Metrics

- Recall= $\frac{TP}{TP+FN}$
- Precision=$\frac{TP}{TP+FP}$
- F1 score=2 $\frac{Precision \times Recall}{Precision+Recall}$

		Predicted	
		0	1
True	0	**TN true negatives: predicted location does not have Asbestos and it truly doesn't have Asbestos**	**FP false positives: predicted location has Asbestos when it doesn't have Asbestos**
	1	**FN false negative: predicted location doesn't have Asbestos when it does have Asbestos**	**TP true positives: predicted location has Asbestos, and it truly has Asbestos**

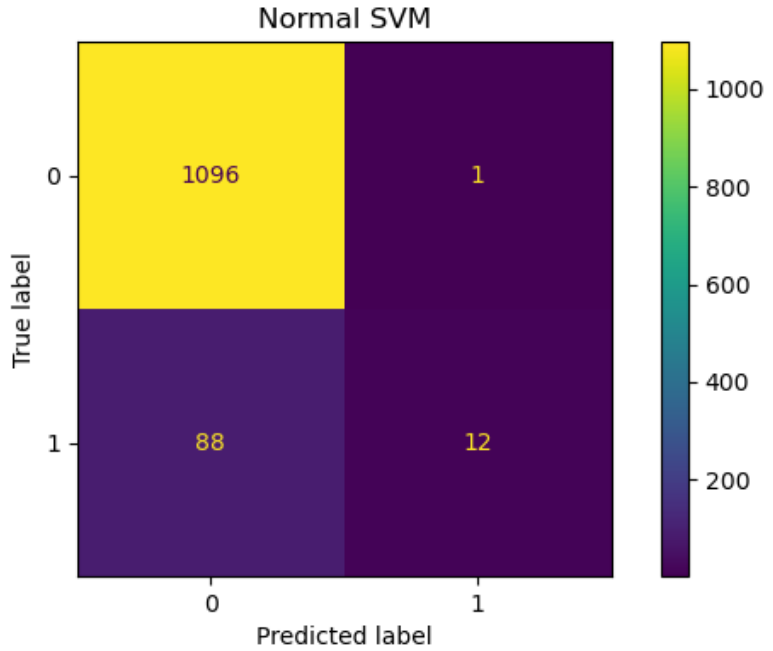
Flowchart

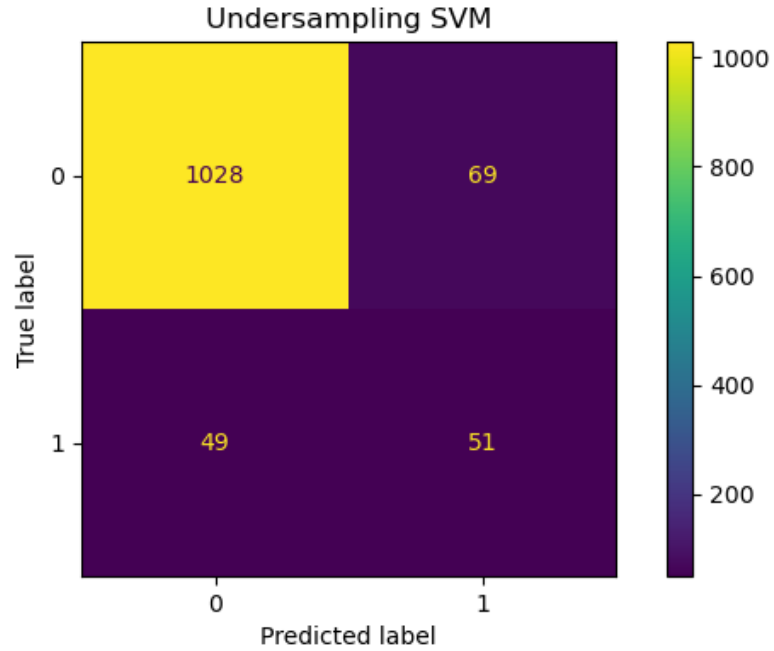

Collect & Preprocess
Training
Normal
Undersampled
Oversampled
Combined
ANN
RF
RBF SVM
Validation
Testing

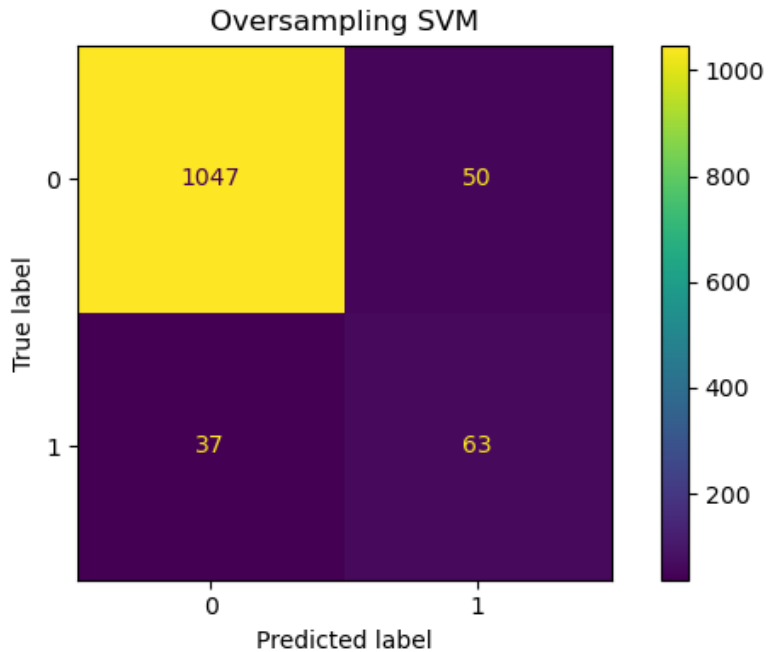
Results

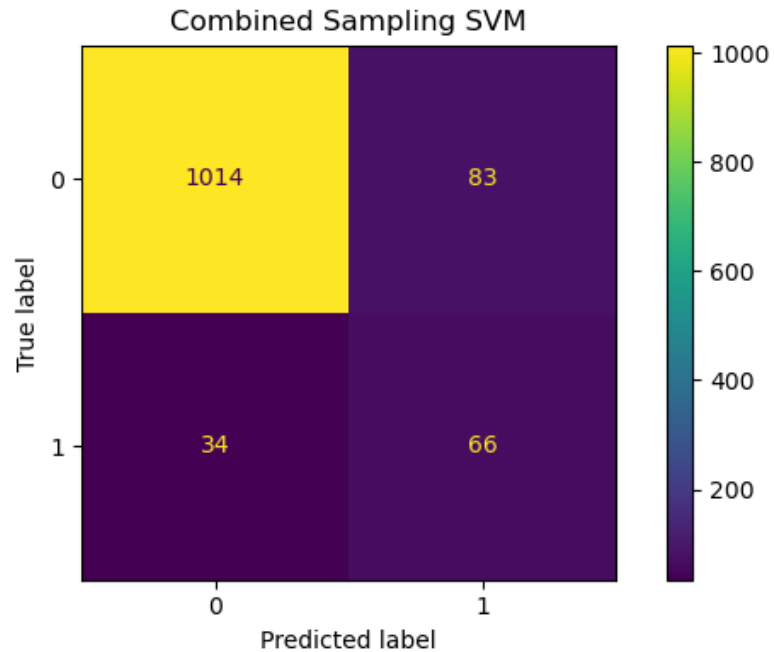
RBF SVM

	Training Results			Validation Results			Testing Results		
	Mean f1 (%)	Mean Recall (%)	Mean Precision (%)	Mean f1 (%)	Mean Recall (%)	Mean Precision (%)	F1 (%)	Recall (%)	Precision (%)
Normal	13.8	7.4	100	5.1	2.7	66.7	21.2	12	92.3
Undersampled	52.9	69.7	42.8	45.4	59.1	37.1	46.4	51	42.5
Oversampled	61.4	75.7	51.7	50.6	62.3	42.9	59.5	63	55.7
Combined	55.6	80	42.6	47.1	68.1	36.2	53	66	44.3


Normal SVM
1096
1
88
12
0
1
True label
Predicted label
1000
800
600
400
200

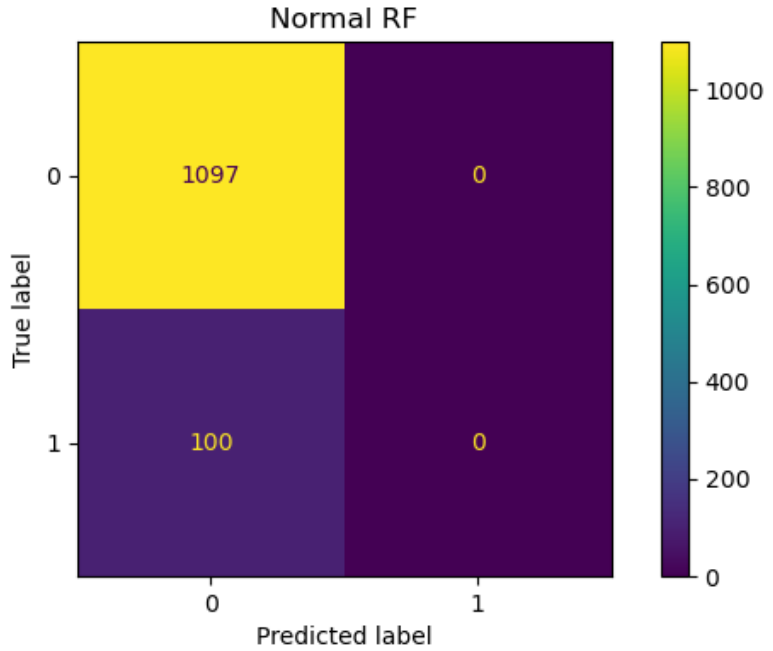

Undersampling SVM
1028
69
49
51
0
1
True label
Predicted label
1000
800
600
400
200

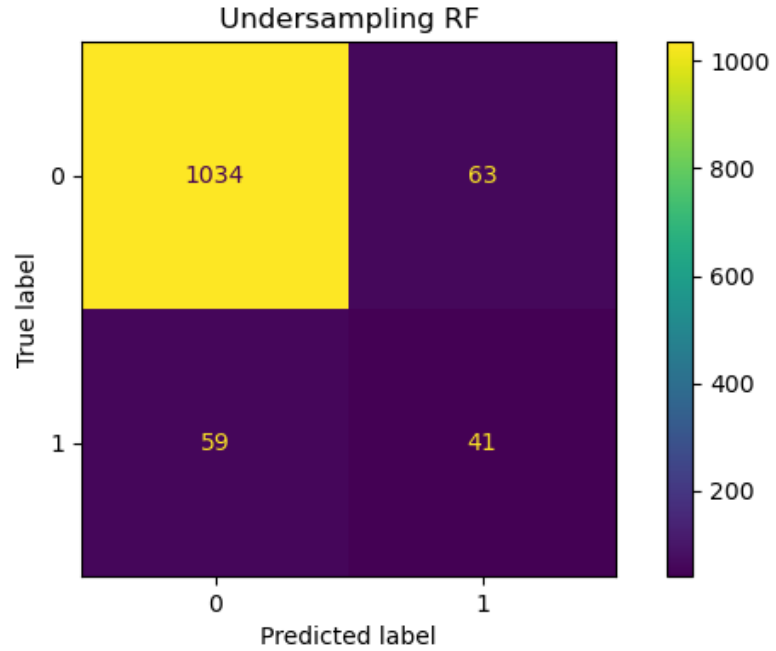

Oversampling SVM
1047
50
37
63
0
1
True label
Predicted label
1000
800
600
400
200

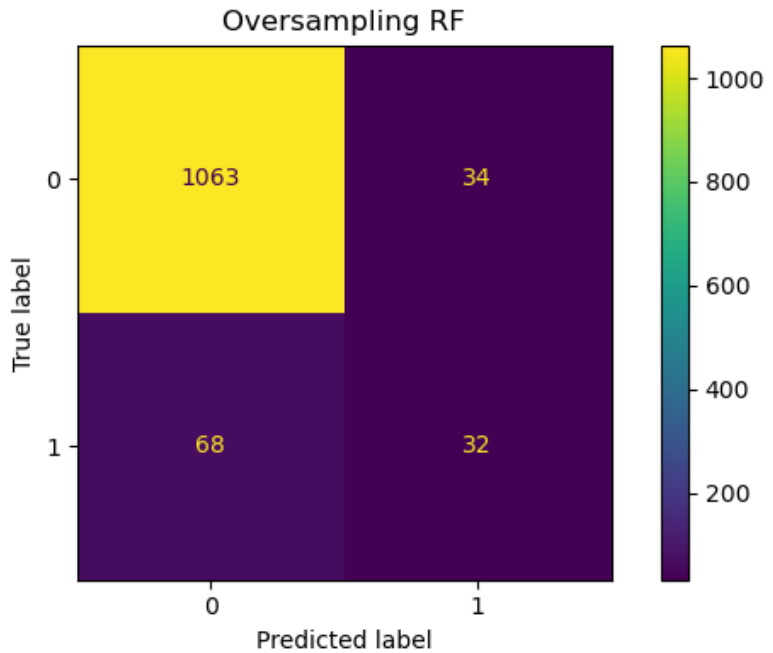

Combined Sampling SVM
1014
83
34
66
0
1
True label
Predicted label
1000
800
600
400
200

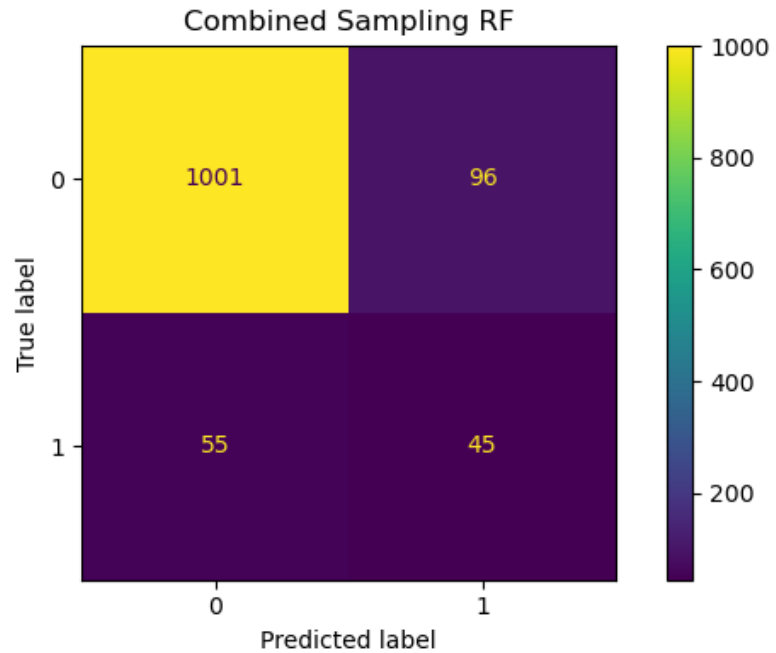
Random Forest

	Training Results			Validation Results			Testing Results		
	Mean f1 (%)	Mean Recall (%)	Mean Precision (%)	Mean f1 (%)	Mean Recall (%)	Mean Precision (%)	F1 (%)	Recall (%)	Precision (%)
Normal	0.2	0.13	20	0	0	0	0	0	0
Undersampled	49.9	56.8	44.5	34.5	28.8	31.2	40.2	41	39.4
Oversampled	53.3	52.8	54.2	30.5	27.1	35.1	38.5	32	48.5
Combined	45.2	63.4	35.2	33.9	46.8	26.6	37.3	45	31.9


Normal RF
1097
0
100
0
0
1
True label
Predicted label
1000
800
600
400
200
0

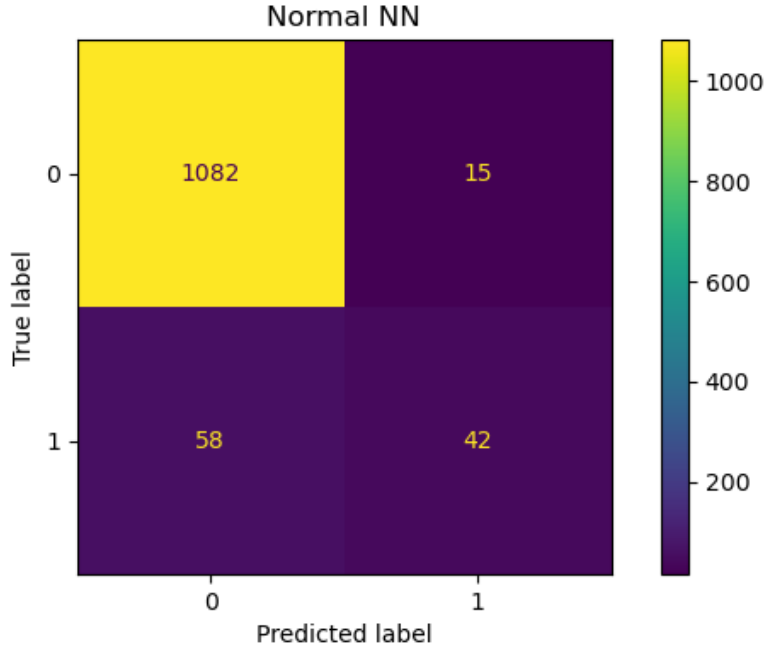

Undersampling RF
1034
63
59
41
0
1
True label
Predicted label
1000
800
600
400
200

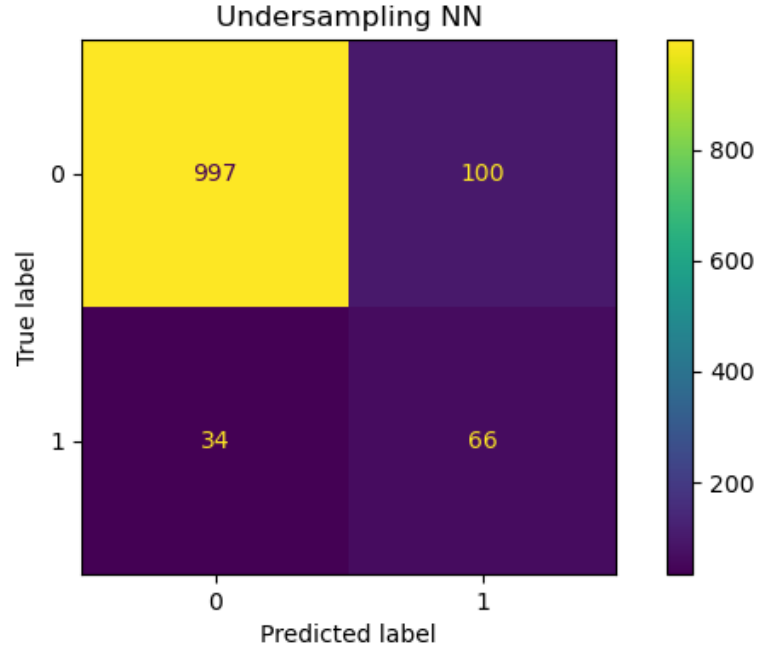

Oversampling RF
1063
34
68
32
0
1
True label
Predicted label
1000
800
600
400
200

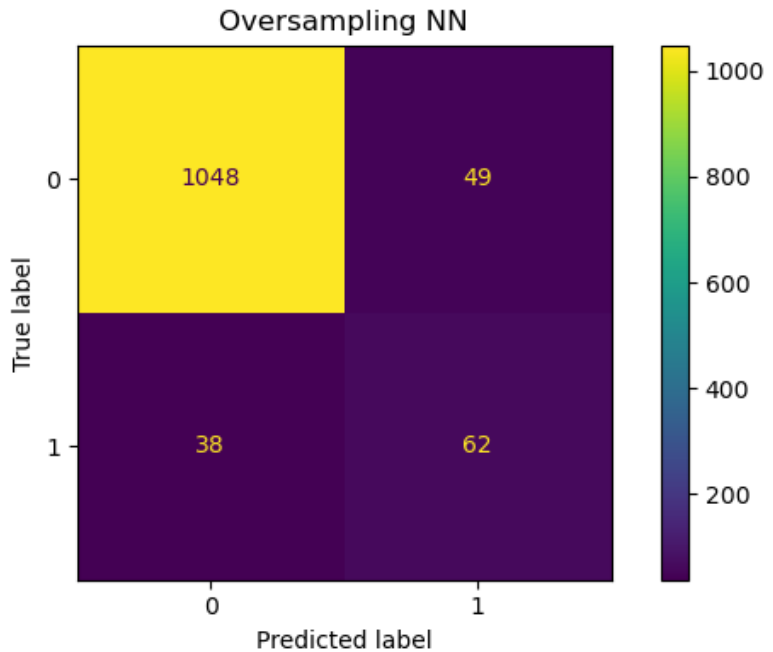

Combined Sampling RF
1001
96
55
45
0
1
True label
Predicted label
1000
800
600
400
200

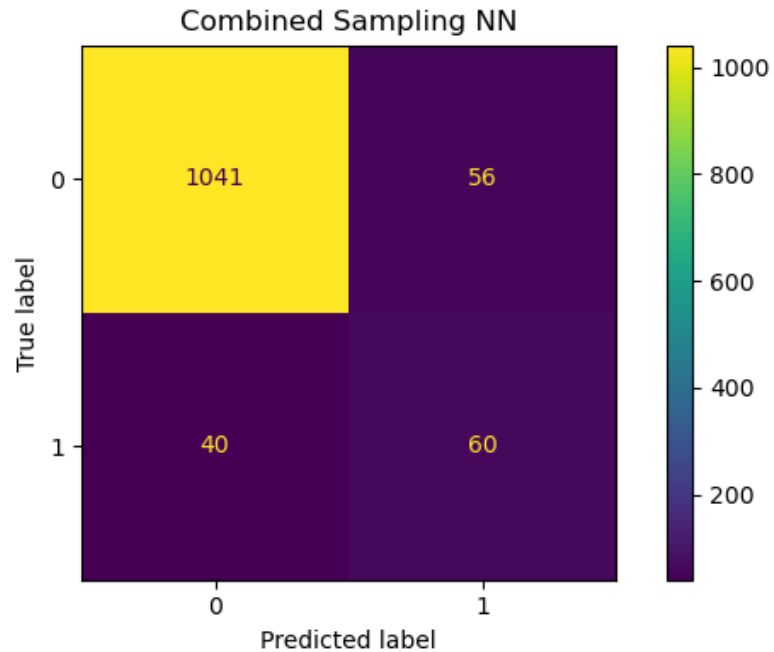
Artificial Neural Network

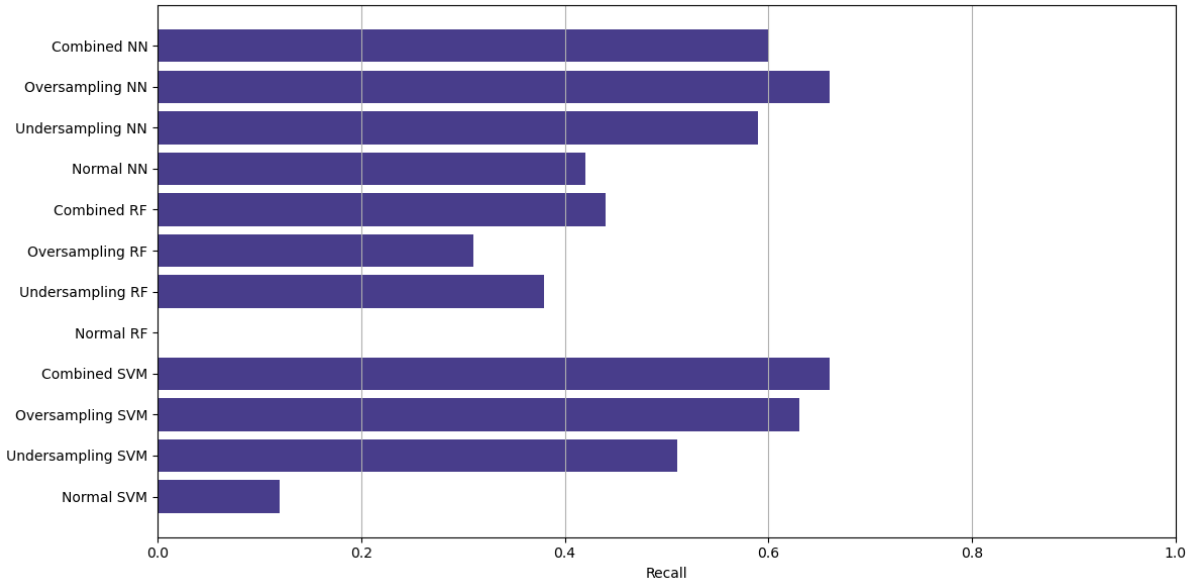
	Training Results			Validation Results			Testing Results		
	Mean f1 (%)	Mean Recall (%)	Mean Precision (%)	Mean f1 (%)	Mean Recall (%)	Mean Precision (%)	F1 (%)	Recall (%)	Precision (%)
Normal	65.4	54.3	82.5	46.9	37.7	62.3	53.5	42	73.7
Undersampled	51.7	74.1	39.9	45.7	67.6	34.8	49.6	66	39.7
Oversampled	76.6	91.1	66.1	49.9	58	44	58.8	62	55.8
Combined	64.8	83.8	52.9	51	67.1	41.4	55.5	60	51.7

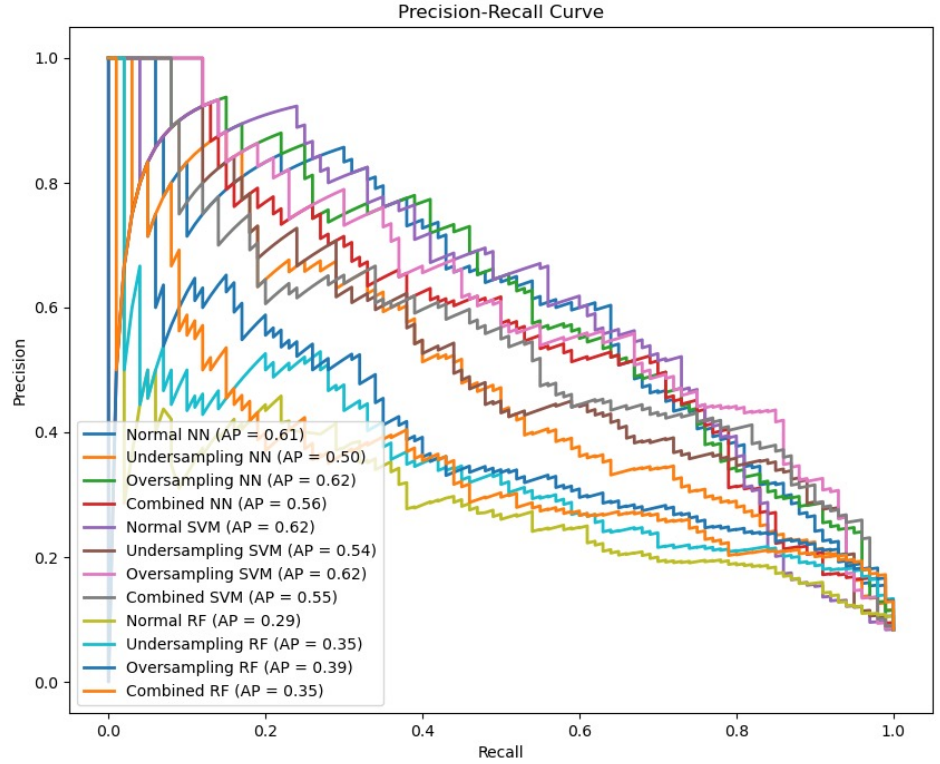

Normal NN
1082
15
58
42
0
1
True label
Predicted label
1000
800
600
400
200

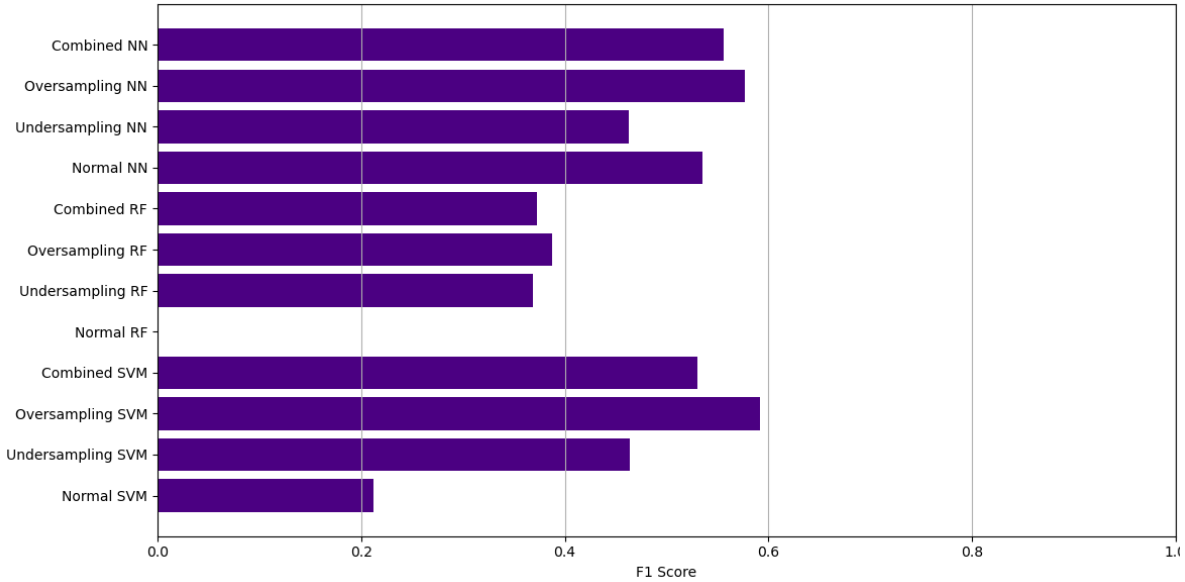

Undersampling NN
997
100
34
66
0
1
True label
Predicted label
800
600
400
200

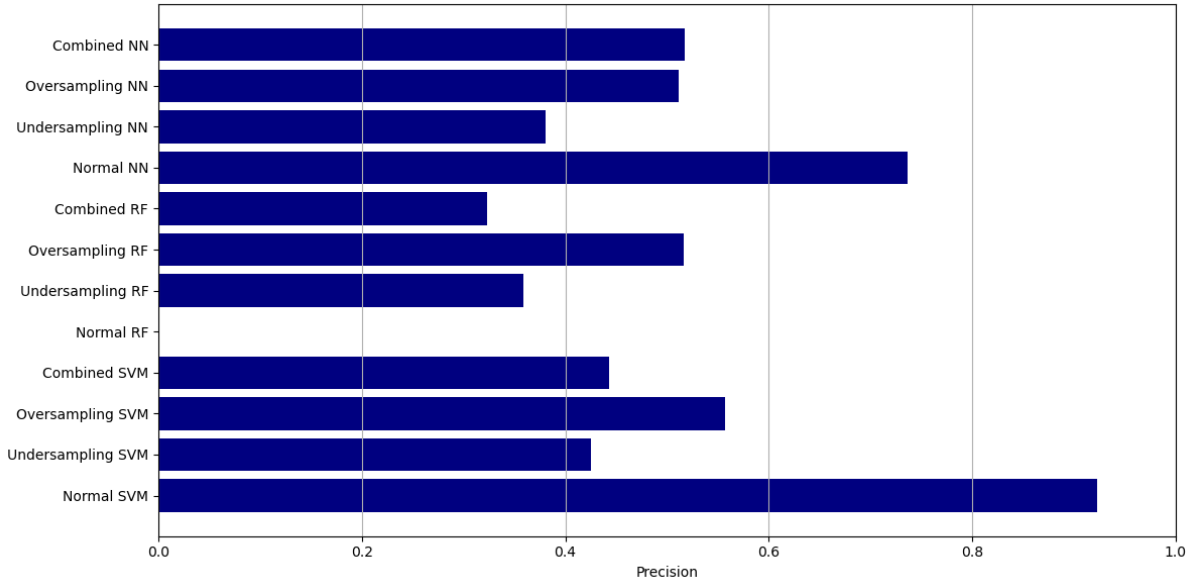

Oversampling NN
1048
49
38
62
0
1
True label
Predicted label
1000
800
600
400
200


Combined Sampling NN
1041
56
40
60
0
1
True label
Predicted label
1000
800
600
400
200


Combined NN
Oversampling NN
Undersampling NN
Normal NN
Combined RF
Oversampling RF
Undersampling RF
Normal RF
Combined SVM
Oversampling SVM
Undersampling SVM
Normal SVM
0.0
0.2
0.4
0.6
0.8
1.0
Recall


Precision-Recall Curve
Precision
Recall
Normal NN (AP = 0.61)
Undersampling NN (AP = 0.50)
Oversampling NN (AP = 0.62)
Combined NN (AP = 0.56)
Normal SVM (AP = 0.62)
Undersampling SVM (AP = 0.54)
Oversampling SVM (AP = 0.62)
Combined SVM (AP = 0.55)
Normal RF (AP = 0.29)
Undersampling RF (AP = 0.35)
Oversampling RF (AP = 0.39)
Combined RF (AP = 0.35)


Combined NN
Oversampling NN
Undersampling NN
Normal NN
Combined RF
Oversampling RF
Undersampling RF
Normal RF
Combined SVM
Oversampling SVM
Undersampling SVM
Normal SVM
0.0
0.2
0.4
0.6
0.8
F1 Score


Combined NN
Oversampling NN
Undersampling NN
Normal NN
Combined RF
Oversampling RF
Undersampling RF
Normal RF
Combined SVM
Oversampling SVM
Undersampling SVM
Normal SVM
0.0
0.2
0.4
0.6
0.8
1.0
Precision

Conclusion

- Oversampling yield highest metrics
- Suboptimal results: Oversampled SVM(f1: 59.5%)
- Limitations: Imbalance – Features – Hyperparameter Tuning

[1] Im S., Youn Kw., Shin D., Lee Mj., Choi SJ. Review of carcinogenicity of asbestos and proposal of approval standards of an occupational cancer caused by asbestos in Korea [online]. 2015; 27(1):34. [Accessed 9 July 2024]. Available from < https://doi.org/10.1186%2Fs40557-015-0080-1>

[2] P.Thives.L, Chisi E., Thives Júnior JJ., Vieira AS. Is asbestos still a problem in the world? A current review [online]. 2022; 319. [Accessed 9 July 2024]. Available from < https://doi.org/10.1016/j.jenvman.2022.115716>

[3] Sentinel-2 Bands. [online]. sentinelhub [Accessed 11 July 2024]. Available from: https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/bands/

[4] S2 Mission. [online]. PROGRAMME OF THE EUROPEAN UNION, Copernicus: SentiWiki; [Accessed 11 July 2024]. Available from: https://sentiwiki.copernicus.eu/web/s2-mission#Spatial-Resolution

[5] Hastie T., Tibshirani T., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Stanford, California: Springer; 2017.