EFFECTIVE WATER QUALITY MONITORING USING DATA FUSION

Department of Information Technology
SSN College of Engineering, Kalavakkam - 603 110
Academic Year 2015 - 2016

STUDENT PROJECT BY

1. AISHWARYA A.
   Reg No. : 312212205002

2. ARUN SHUNMUGAM J.
   Reg No. : 312212205014

3. DEEPAK S.
   Reg No. : 312212205023

PROJECT GUIDES

Ms. S. Mohanavalli
Assistant Professor

Ms. Srividya
Assistant Professor

INTRODUCTION

- A robust **Real-Time Water Quality Monitoring System** to detect and identify the quality of river water dynamically
- A combination of **Data Analytics, Data Fusion and Machine Learning** techniques
- Real-time Data from river water and surroundings obtained and processed to alert the user when water detected non-potable or non-usable
- System classifies water usage according to different needs
- Affordable by a common man
- Suited to Indian circumstances

ORIGIN AND IMPORTANCE

- Water - depleting resource
- Classify water, rather waste it (Know your water)
- High correlation between water quality and human health
- Water-borne diseases due to improper discharge of industrial wastes, Domestic sewage
- Water is polluted. already existing systems - not suitable, not affordable, not real time
- Only preliminary research is being done regarding application of Data Fusion in Environmental Science
- Contribution to Environment

BACKGROUND

- Major pollutants: Asbestos, Lead, Mercury, Nitrates, Phosphates, Sulphur, Oils, Petrochemicals
- Industries involved: complex organic chemical industries, electric power plants, food industry, iron and steel industry, mines and quarries, nuclear industry, pulp and paper industry, water treatment
- Causes: chemicals, pathogens, solids and emulsions, hydraulic fracturing, unregulated industrial waste discharge, lacks sufficient treatment capacity in India, growing population, unrelenting urbanization
- Effects: elevated temperature, discoloration, increased turbidity, toxic wastes increase, oxygen depletion, affecting plants, fishes' gills, waterborne diseases in humans

OBJECTIVE

- To identify the quality of water
- To classify the usage of water according to the various uses
- To alert the user whenever quality deteriorated
- To assess the nature and extent of pollution control needed in different water bodies
- To evaluate water quality trend over a period of time
- To understand the environmental fate
- To assess the fitness of water for different uses

LITERATURE SURVEY

- CPCB's system NWMP monitors around 54 parameters but in a static manner
- CPCB tests waters on a yearly or a half yearly basis
- It plans to bring in a real-time feature to its system in the near future
- 9 core parameters and 5 different classes of usage - Drinking, Outdoor bathing, Irrigation, Industrial cooling, Controlled Waste Disposal

- Air and soil quality parameters not included in existing systems
- Prevalent systems use multi sensor data fusion for monitoring using wireless sensor networks
- Existing architectures use local and central fusion modules for same quality parameter

- Feature Extraction is used for Dimensionality Reduction
- It is a predecessor step to classification
- Decision Support System integrates feature extraction and classification processes
- Feature extraction helps get desired results with parameters reduced

- Machine learning helps achieve automation
- It also helps achieve dynamic performance
- There is only preliminary research done using data analytics and machine learning in water quality monitoring
- This is an initiative to smart water use accustomed to Indian circumstances

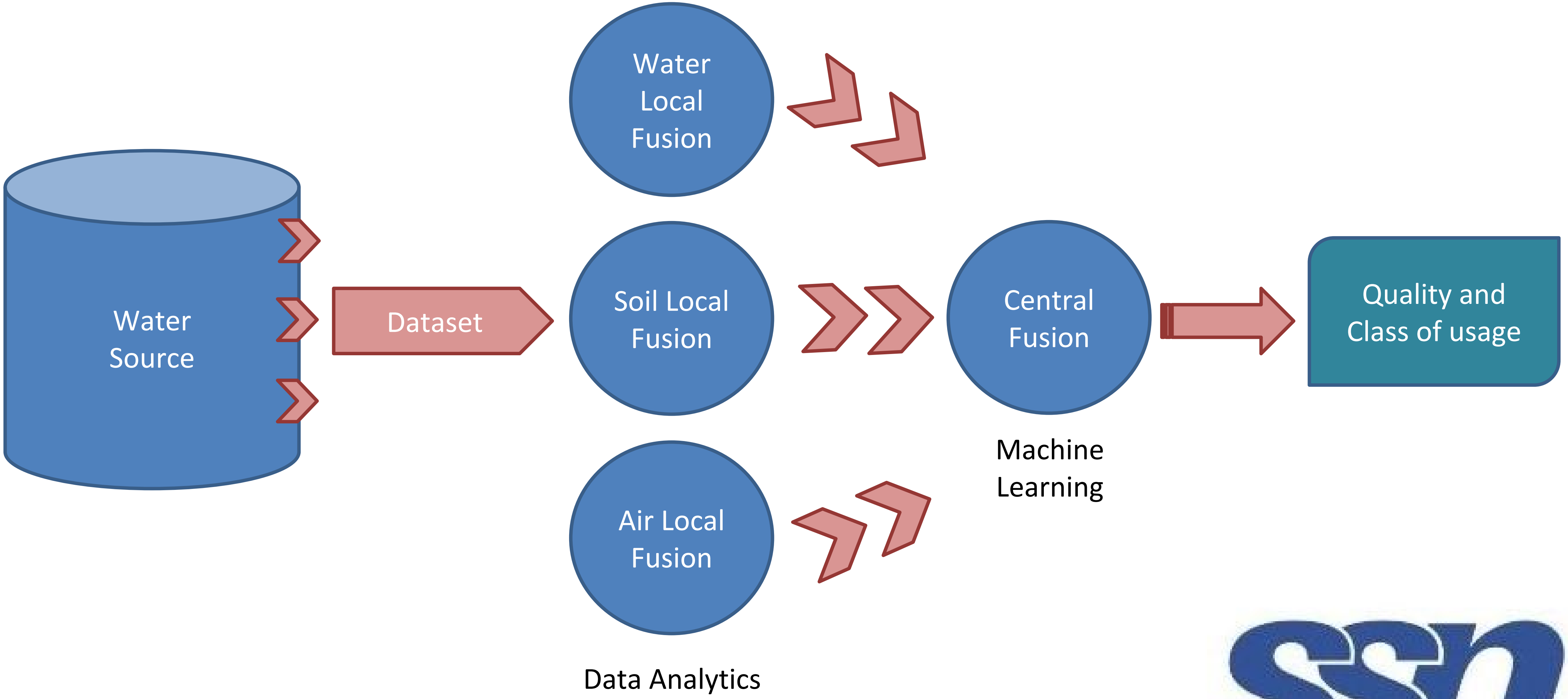
SOFTWARE REQUIREMENTS

- **RStudio Desktop 0.99.893**
  - Powerful and productive IDE for R
  - https://www.rstudio.com/products/rstudio/download/
- **R-3.2.4 for Windows (32/64 bit)**
  - https://cran.r-project.org/bin/windows/base/

R Packages needed :

- stats
- cluster
- caret
- e1071
- FSelector
- mlbench
- randomForest
- rpart

Overall System Design

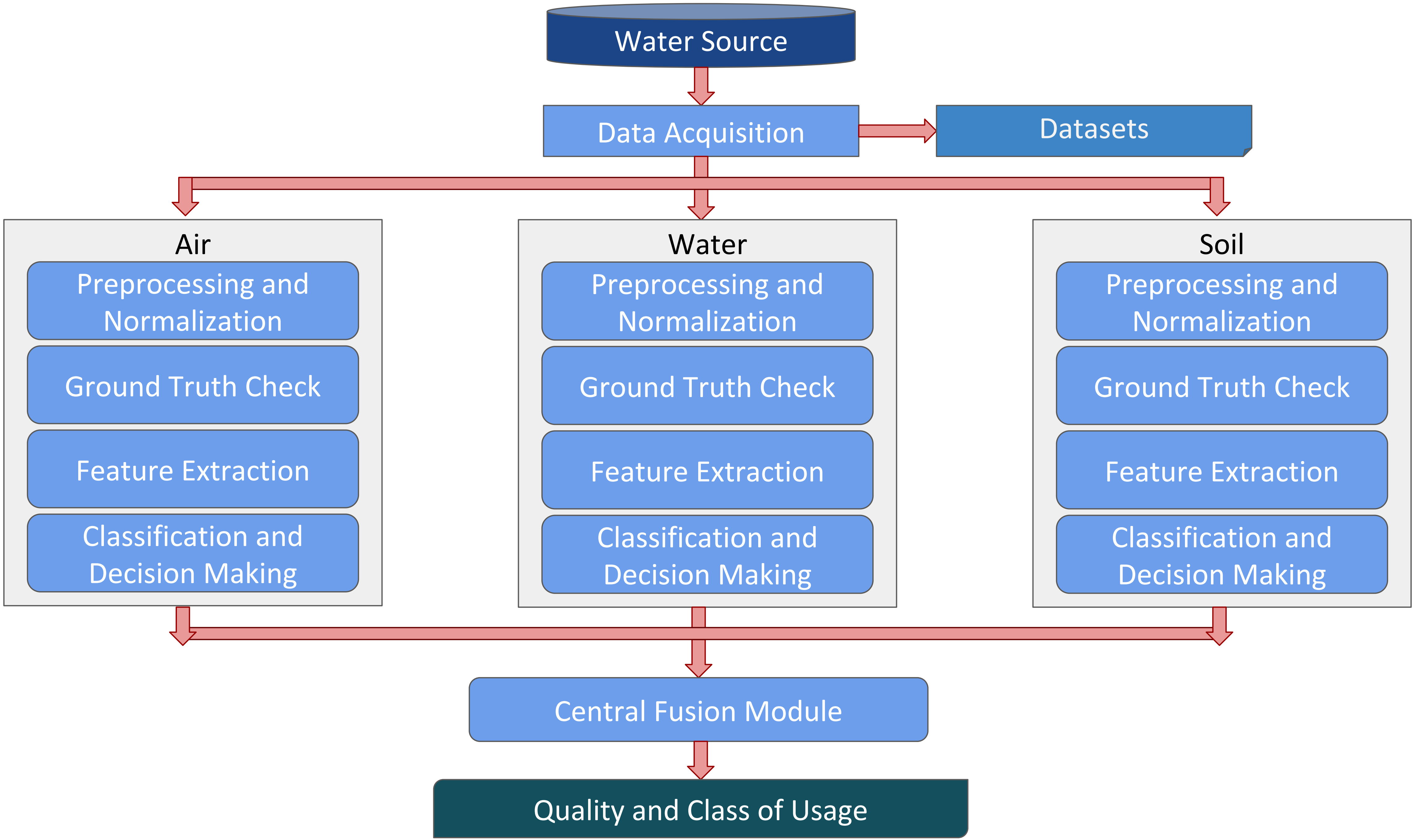


OVERALL SYSTEM DESIGN

- Water, air and soil quality parameters acquired from river water
- Acquired data is split and sent to individual Local fusion modules
- Each local fusion module observes values from water, air and soil, processes this data and gives individual decisions of the quality
- Each local module decides the quality of their data individually
- Individual decisions are sent to Central Fusion module
- This module decides the final overall quality of water
- This module suggests the appropriate class of usage

SYSTEM ARCHITECTURE

- Level 1 : Data Acquisition
- Level 2 : Local Fusion Module
- Level 2.1 : Preprocessing and Normalization
- Level 2.2 : Ground Truth Check
- Level 2.3 : Feature Extraction
- Level 2.4 : Classification and Decision Making
- Level 3 : Central Fusion Module
- Output : Quality of water and Class of Usage

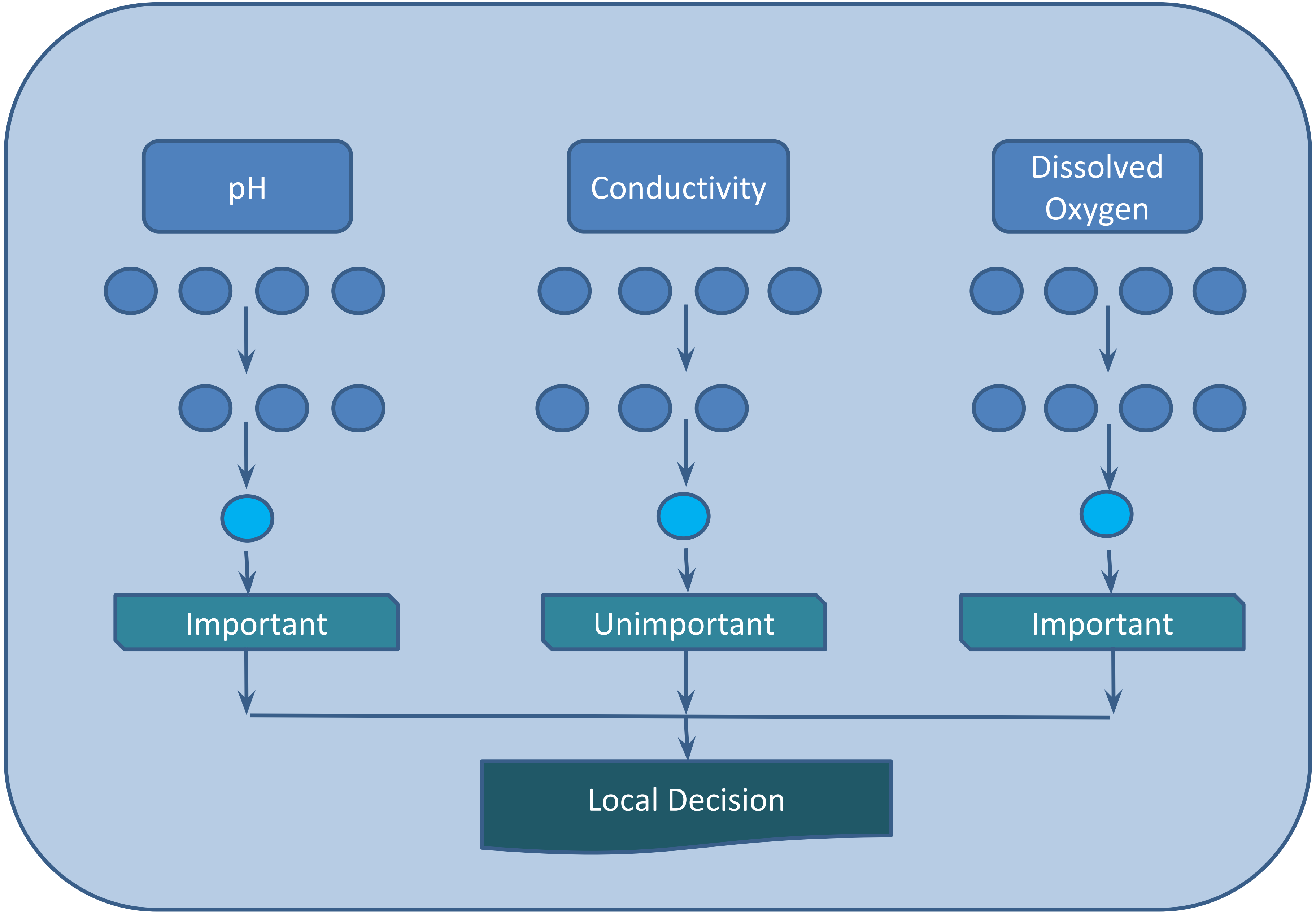


Water Source
Data Acquisition
Datasets
Air
Preprocessing and Normalization
Ground Truth Check
Feature Extraction
Classification and Decision Making
Water
Preprocessing and Normalization
Ground Truth Check
Feature Extraction
Classification and Decision Making
Soil
Preprocessing and Normalization
Ground Truth Check
Feature Extraction
Classification and Decision Making
Central Fusion Module
Quality and Class of Usage

LOCAL FUSION MODULE

- Parameter readings are obtained via datasets
- Records cleaned for missing or erroneous values using Preprocessing and Normalization
- Next, mean values determined using K-Means Clustering in water dataset to check with ground truth
- Voting system is used for air and soil parameters
- Feature Extraction used to cut down irrelevant features - Dimension Reduction - different methods for every dataset
- Refined dataset used to classify records using Naive Bayes and Support Vector Machine
- Local decision given by Decision Tree method

Local Fusion Module



Preprocessing and Normalization

Ground Truth Check

Feature Extraction

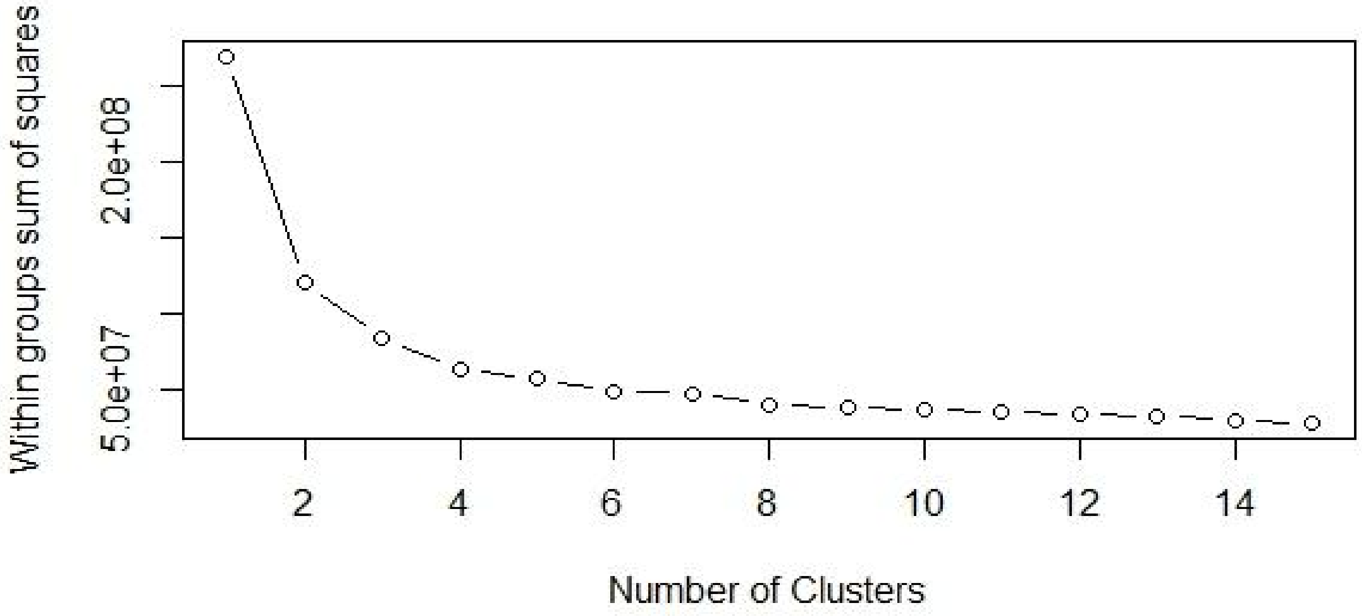
Classification

PREPROCESSING AND NORMALIZATION

- Records with missing or erroneous values identified and omitted
- Normalization done for every value

GROUND TRUTH CHECK - CLUSTERING

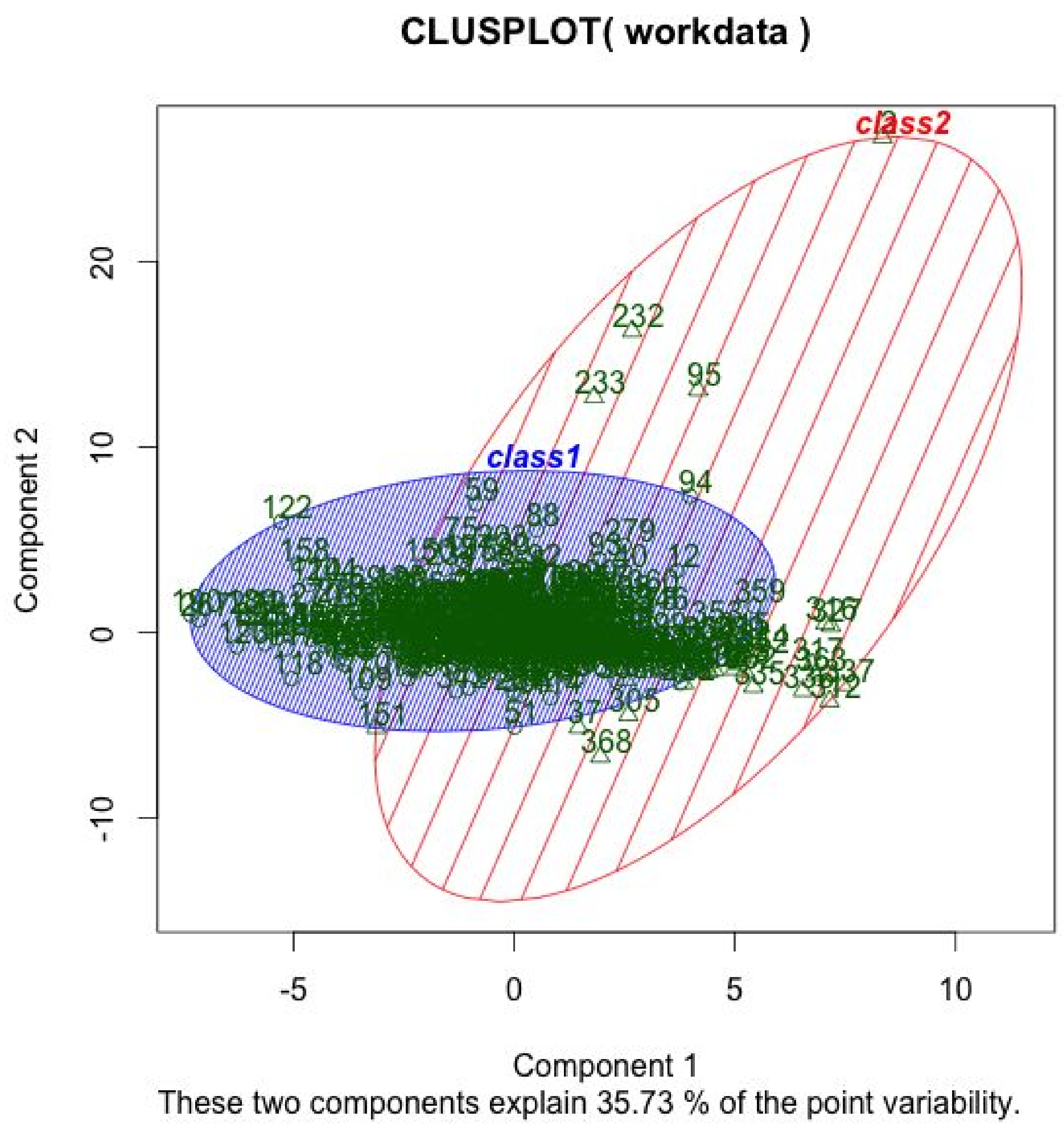
- Elbow Method used to determine optimal number of clusters for dataset
- K-Means Clustering used to categorize the data points
- Cluster labels are appended to the dataset



Elbow Method Optimal Clusters = 2 for **UCI Water Dataset**

CLUSTERING RESULTS FOR WATER DATASET

S.No	Name of the dataset	Number of Attributes	Number of clusters	Cluster means (For instance)
1	UCI Water Dataset	37	2 clusters Cluster1 - 271 entries Cluster2 – 109 entries	Cluster 1-pH (7.778544) Cluster 2-pH (7.927731)


CLUSPLOT(workdata)
class1
class2
Component 2
Component 1
20
10
0
-10
-5
0
5
10
These two components explain 35.73 % of the point variability.

FEATURE EXTRACTION

- Real-time data flows in at regular intervals
- Lesser control over data
- Clustering similarity between different number of attributes
- Currently using brute force method to extract the core attributes from all others
- Core attributes vary with data. Dataset 1 core attribute is pH but for dataset 2 core attribute is different. Difference due to diverse water sources
- When control over horizontal loading of data cannot be achieved, vertical control can be done
- Greatly reduces time & space and also improves processing efficiency

SELECTION METHODS

- Brute Force
- Chi-Squared
- Neural Networks with Feature Extraction
- Removal of Redundant Features by Correlation
- Recursive Feature Extraction (RFE) with Random Forest
- Random Forest
- Boruta
- Linear Vector Quantization (LVQ)
- Partial Least Squares Discriminant Analysis (PLSDA)

CLASSIFICATION METHODS

- Naive Bayes Classification
- Support Vector Machine (SVM)

EXTRACTED FEATURES - WATER

```
Chi squared feature selection

[1] "PH.D"     "DBO.D"    "PH.E"     "PH.P"     "DBO.P"    "SSV.E"    "DQO.D"    "RD.DQO.G"
 [9] "DBO.E"    "SSV.P"    "COND.S"   "COND.D"   "COND.E"   "DQO.E"    "COND.P"   "SED.D"
[17] "SED.P"    "SS.D"     "SED.E"    "ZN.E"


Random forest filter
 [1] "PH.D"     "DBO.P"    "DBO.D"    "PH.P"     "PH.E"     "SSV.E"    "DBO.E"    "DQO.D"
 [9] "SSV.P"    "RD.SS.G"  "DQO.E"    "SS.P"     "RD.DQO.G" "COND.P"   "SS.D"     "DBO.S"
[17] "SED.P"    "COND.E"   "RD.DBO.G" "RD.SS.P"


Automatic Feature Selection Methods using Recursive Feature Elimination (RFE) with Random Forest Algorithm

"PH.D"     "DBO.P"    "DBO.D"    "PH.P"     "PH.E"     "SSV.E"    "DQO.D"    "SSV.P"    "DBO.E"    "RD.DQO.G" "RD.SS.G"
"SS.P"     "DBO.S" "SS.S"     "DQO.E"    "SS.D"     "COND.E"   "SED.P"    "COND.S"   "COND.P"
```

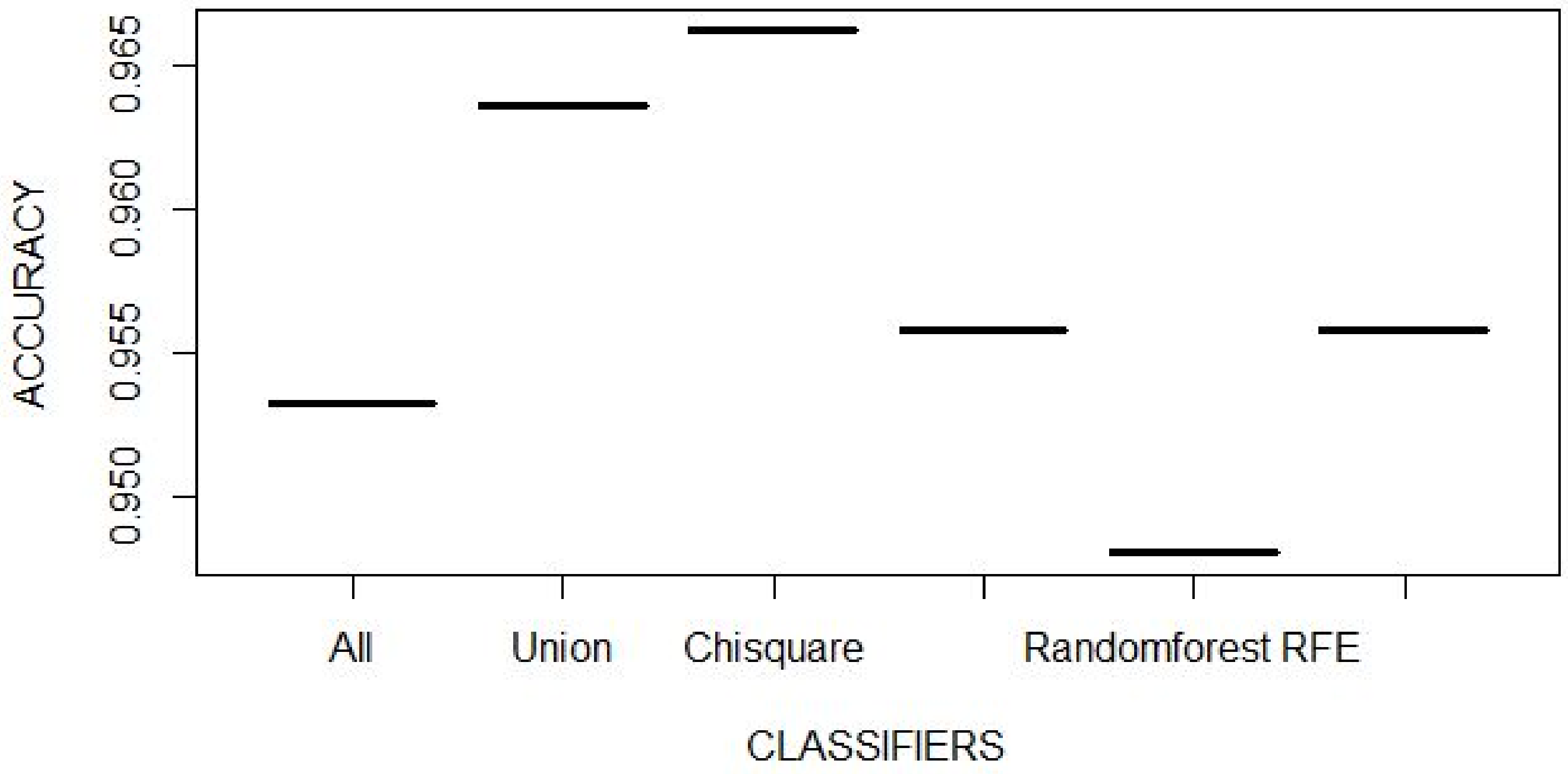
PERFORMANCE METRICS

- **Accuracy:** a description of systematic errors given by the formula (True Positives + True Negatives) / (Positives + Negatives)
- **Precision**: proportion of instances that are truly of a class divided by the total instances classified as that class
- **Recall**: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate)
- **F-Measure**: A combined measure for precision and recall calculated as 2 * Precision * Recall / (Precision + Recall)

RESULTS - WATER

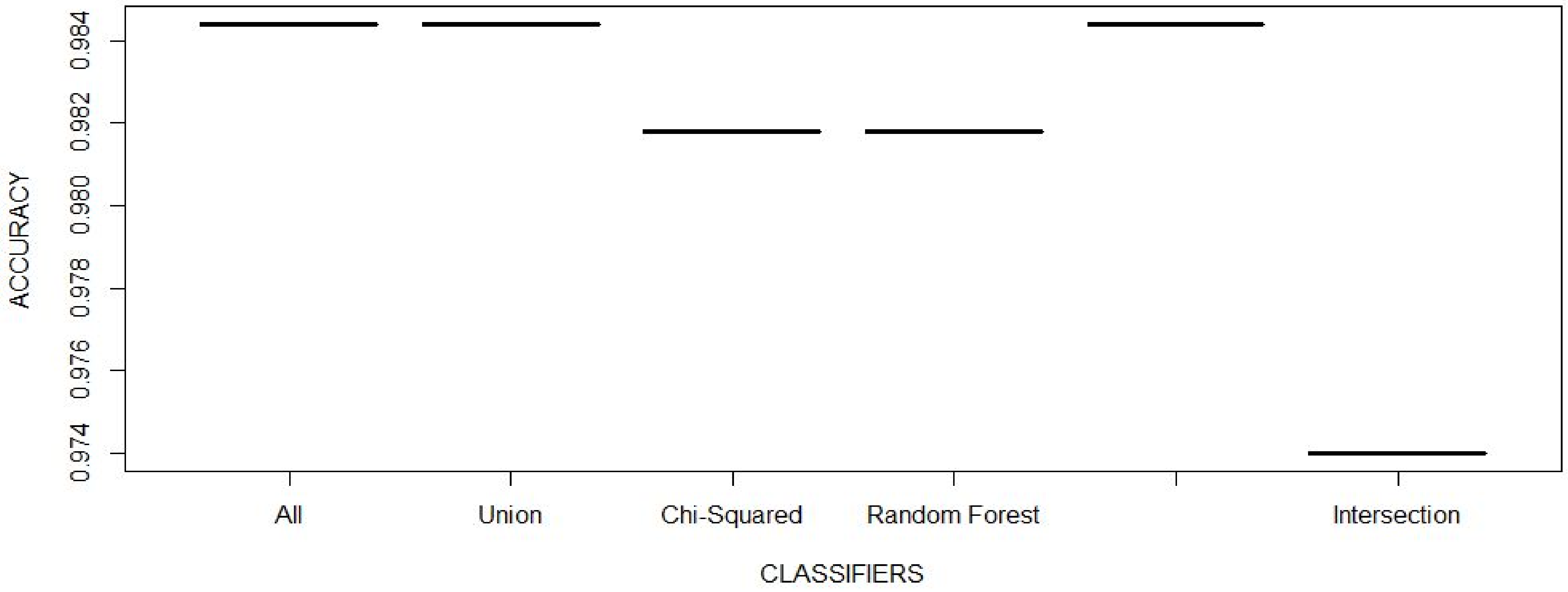
Classifier used : Naive Bayes

S.NO	FE METHODS	NO. OF PARAMETERS	ACCURACY	PRECISION	RECALL	F-SCORE
1	None	37	0.9532	0.9824	0.9654	0.9484
2	Union (RandomForest +ChiSquared)	25	0.9636	0.9818	0.9759	0.9581
3	ChiSquared	20	0.9662	0.9967	0.9627	0.9595
4	RandomForest	20	0.9558	0.9849	0.9646	0.9500
5	RFE with RandomForest	20	0.9480	0.9702	0.9702	0.9412
6	Intersection (RF + Chi)	16	0.9558	0.9967	0.9504	0.9472


ACCURACY
0.950
0.955
0.960
0.965
All
Union
Chisquare
Randomforest RFE
CLASSIFIERS

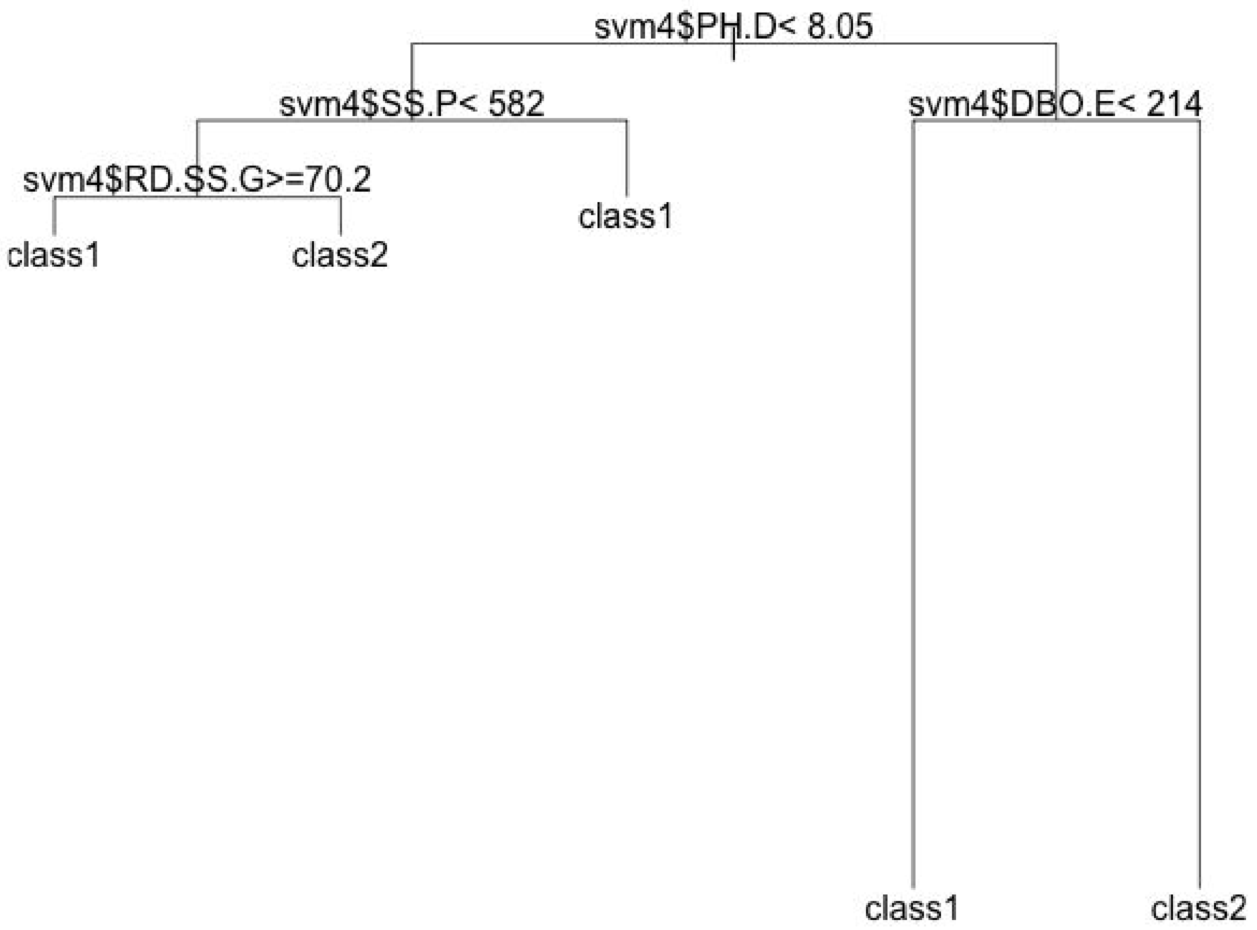
Classifier used : Support Vector Machine (SVM)

S.NO	FE METHODS	NO. OF PARAMETERS	ACCURACY	PRECISION	RECALL	F-SCORE
1	None	37	0.9844	0.9883	0.9941	0.9824
2	Union (RandomForest +ChiSquared)	25	0.9844	0.9883	0.9941	0.9824
3	ChiSquared	20	0.9818	0.9826	0.9970	0.9796
4	**RandomForest**	**20**	**0.9818**	**0.9854**	**0.9941**	**0.9795**
5	RFE with RandomForest	20	0.9844	0.9883	0.9941	0.9824
6	Intersection (RF + Chi)	16	0.9740	0.9768	0.9941	0.9710



ACCURACY
0.974
0.976
0.978
0.980
0.982
0.984
All
Union
Chi-Squared
Random Forest
Intersection
CLASSIFIERS

- After Feature Selection and classification, Decision Trees are used to derive the decision
- **Decision Tree Learning** gives the quality of water from the results of Classification



OTHER FUSION MODULES - AIR AND SOIL

- Similar to Water, cleaning and feature extraction done for Air and Soil
- Once cleaning done same feature extraction models are executed for air and soil
- Best methods out of those are used
- Voting System is used for Air and Soil parameters to attain the Ground Truth for Classification

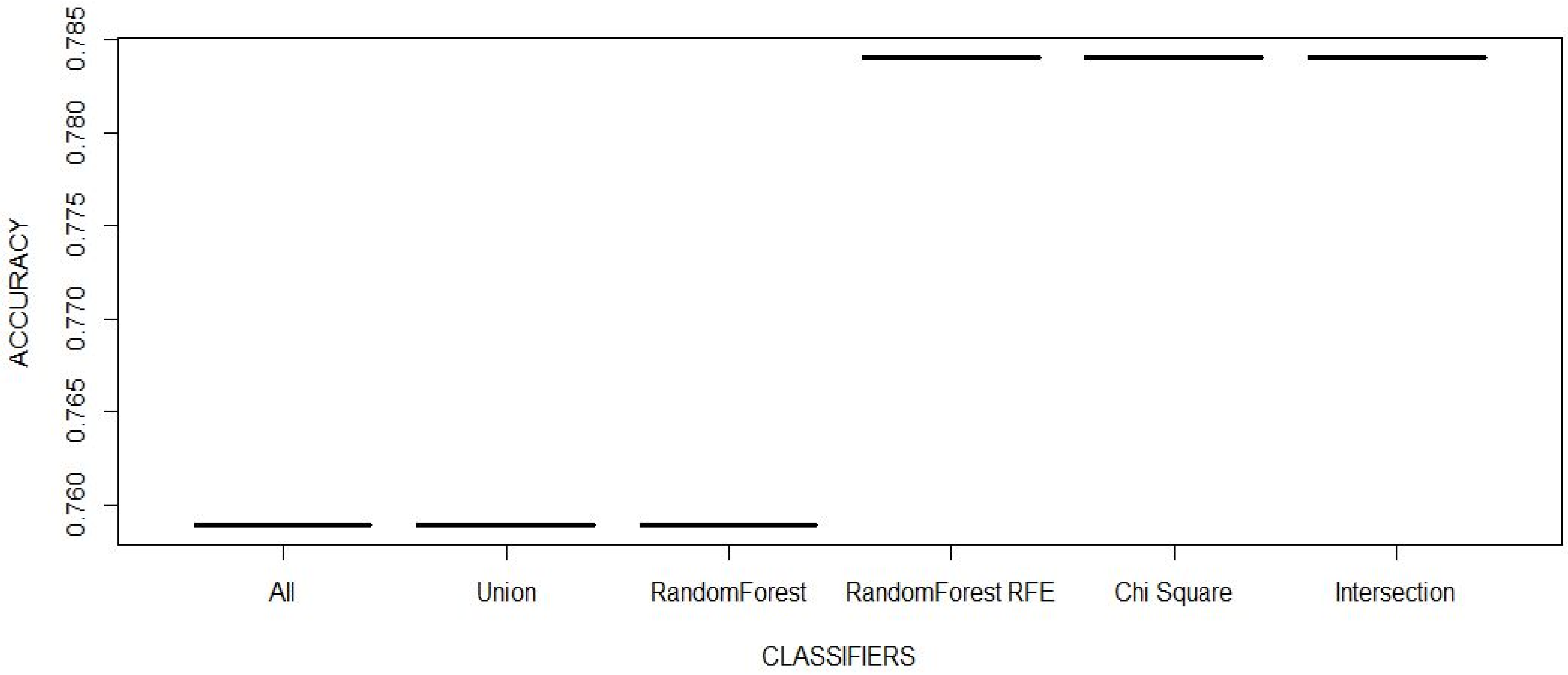
EXTRACTED FEATURES - AIR

```
Chi-square
[1] "O3"    "BA_P"  "PM2.5" "SO2"   "NO2"

Random Forest
[1] "O3"    "NO2"   "PM2.5" "BA_P"  "SO2"   "PM10"

Union
[1] "O3"    "BA_P"  "PM2.5" "SO2"   "NO2"   "PM10"

Intersection
[1] "O3"    "BA_P"  "PM2.5" "SO2"   "NO2"

Random Forest RFE
[1] "O3"    "PM2.5" "BA_P"  "NO2"   "SO2"
```

RESULTS - AIR

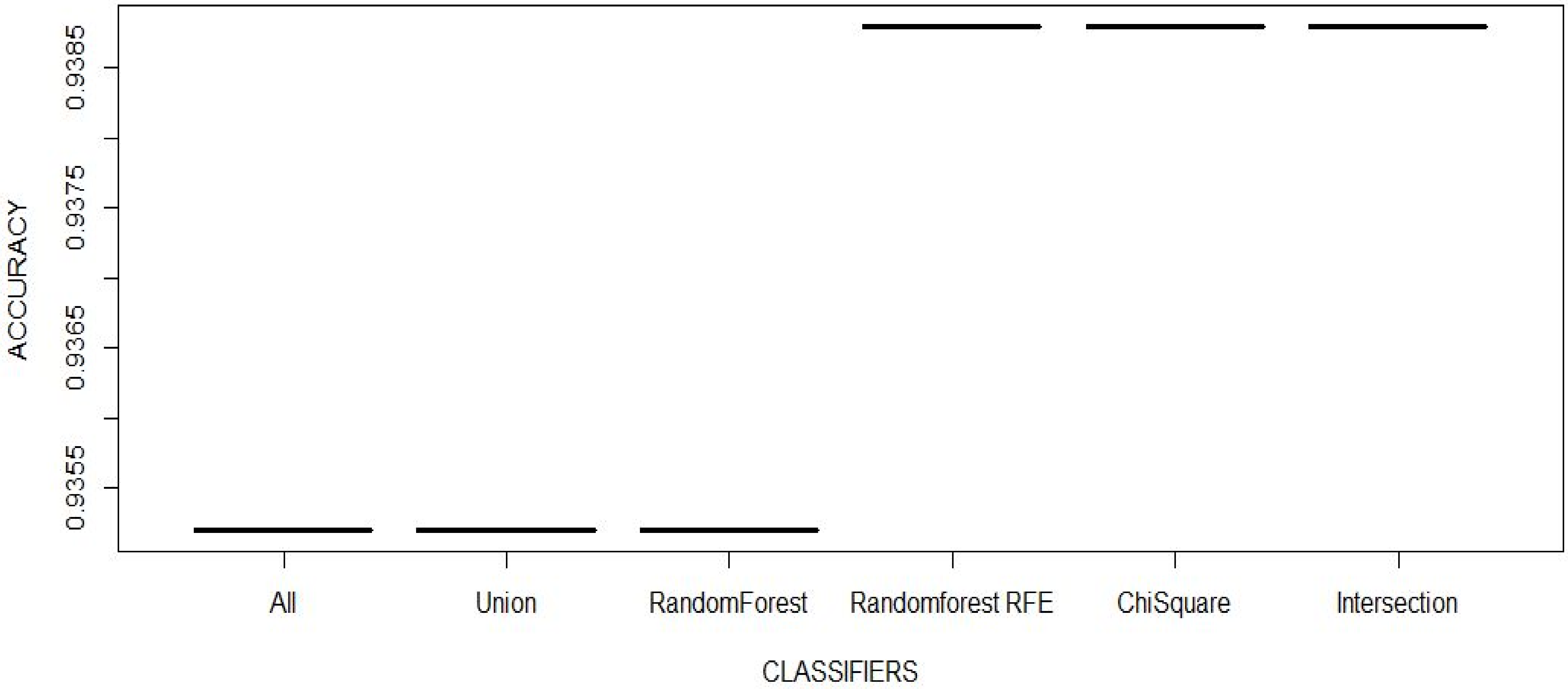
Classifier used : Naive Bayes

S.NO	FE METHODS	NO. OF PARAMETERS	ACCURACY	PRECISION	RECALL	F-SCORE
1	None	6	0.7589	0.8493	0.5254	0.4462
2	Union (RandomForest +ChiSquared)	6	0.7589	0.8493	0.5254	0.4462
3	ChiSquared	5	0.7841	0.8630	0.5575	0.4811
4	RandomForest	6	0.7589	0.8493	0.5254	0.4462
5	RFE with RandomForest	5	0.7841	0.8630	0.5575	0.4811
6	Intersection (RF + Chi)	5	0.7841	0.8630	0.5575	0.4811


0.785
0.780
0.775
0.770
0.765
0.760
ACCURACY
All
Union
RandomForest
RandomForest RFE
Chi Square
Intersection
CLASSIFIERS

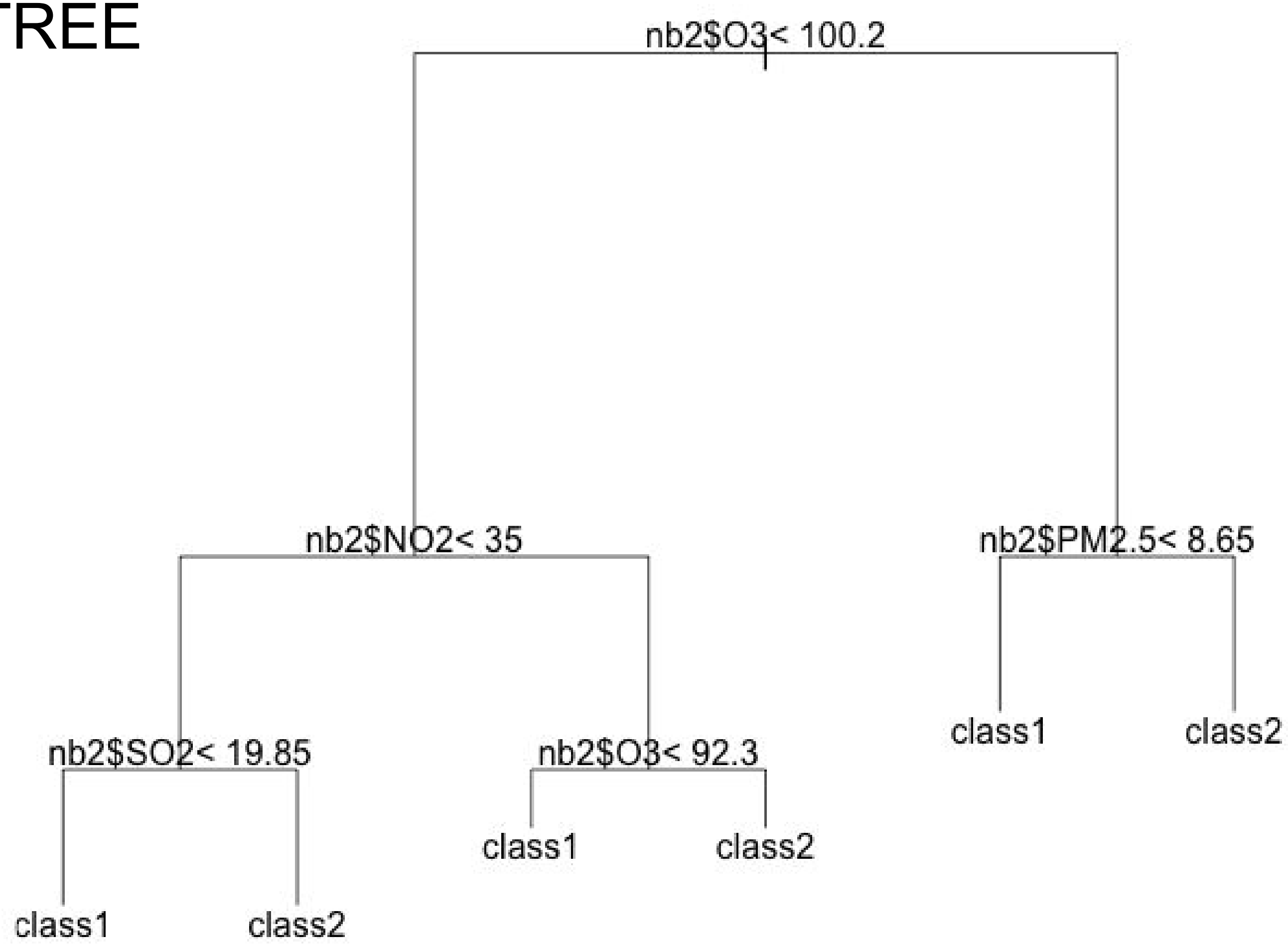
Classifier used : Support Vector Machine (SVM)

S.NO	FE METHODS	NO. OF PARAMETERS	ACCURACY	PRECISION	RECALL	F-SCORE
1	None	6	0.9352	0.8082	0.9365	0.7568
2	Union (RandomForest +ChiSquared)	6	0.9352	0.8082	0.9365	0.7568
3	**ChiSquared**	**5**	**0.9388**	**0.8219**	**0.9375**	**0.7705**
4	RandomForest	6	0.9352	0.8082	0.9365	0.7568
5	RFE with RandomForest	5	0.9388	0.8219	0.9375	0.7705
6	Intersection (RF + Chi)	5	0.9388	0.8219	0.9375	0.7705



ACCURACY
0.9355
0.9365
0.9375
0.9385
All
Union
RandomForest
Randomforest RFE
ChiSquare
Intersection
CLASSIFIERS

DECISION TREE - AIR



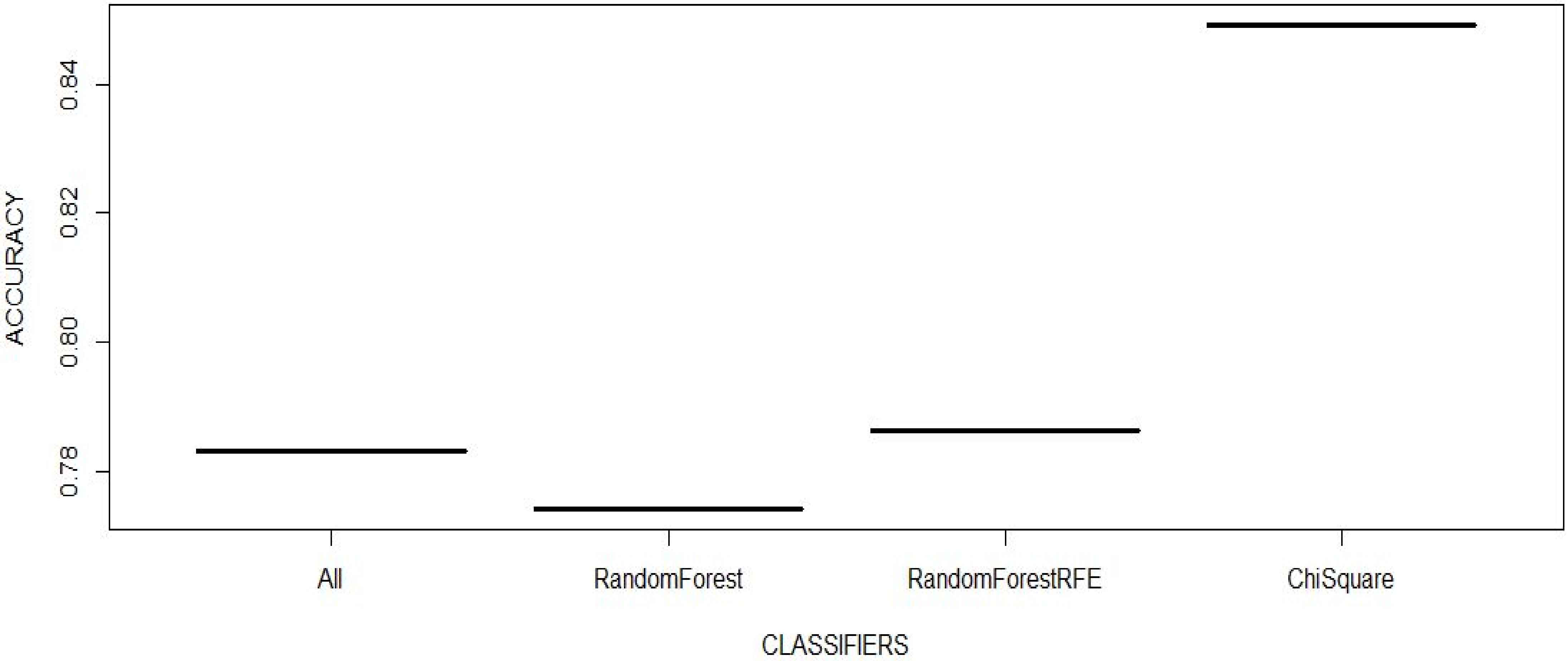
EXTRACTED FEATURES - SOIL

```
Chi-square
[1] "pH"            "CEC"     "Conductivity"   "NITROGEN"

Random Forest
[1] "pH"              "CEC"              "ESP"               "Conductivity"
[5] "NITROGEN"        "WATER.CAPACITY" "SATURATION"      "PHOSPHOROUS"
[9] "CARBON"

Random Forest RFE
[1] "pH"            "CEC"              "ESP"              "Conductivity"
[5] "NITROGEN"
```

RESULTS - SOIL

Classifier used : Naive Bayes

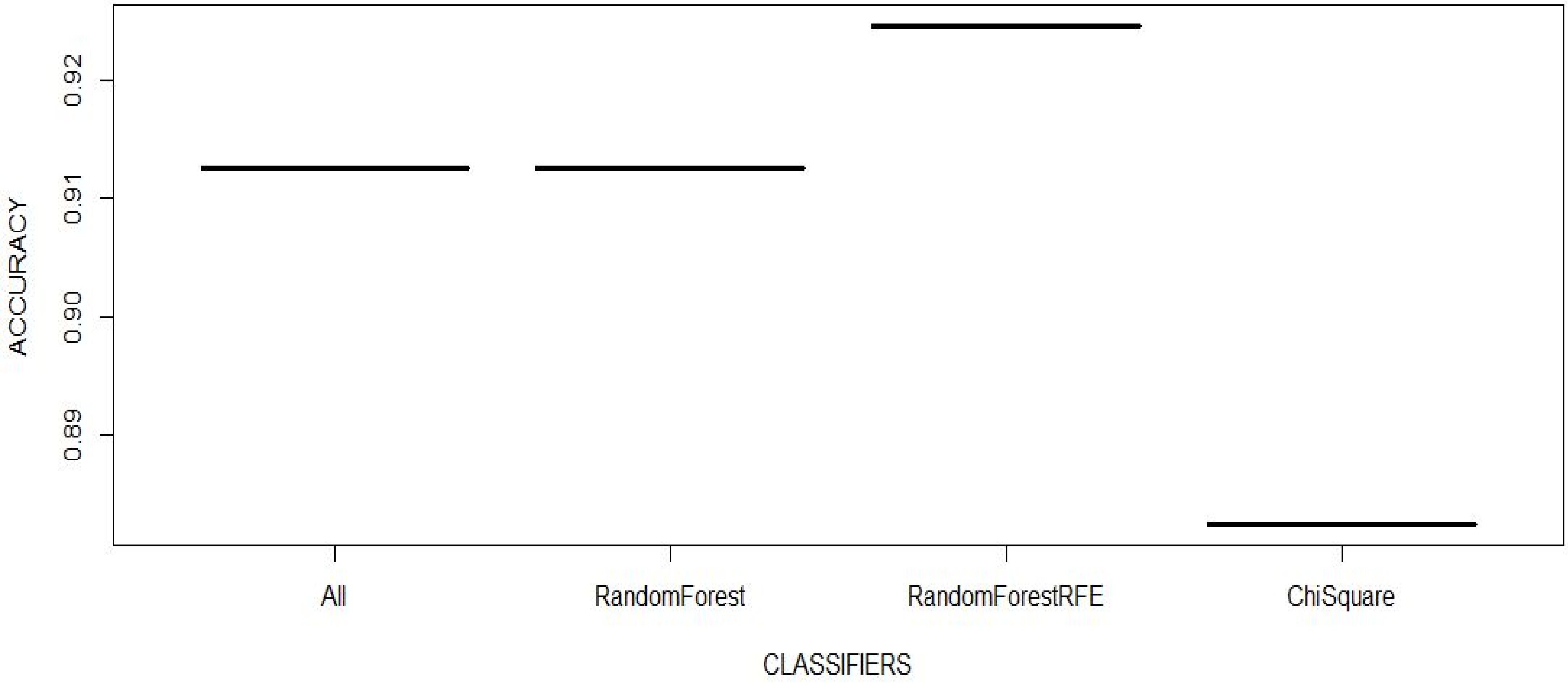
S.NO	FE METHODS	NO. OF PARAMETERS	ACCURACY	PRECISION	RECALL	F-SCORE
1	None	10	0.7831	0.8229	0.9186	0.7559
2	RandomForest	9	0.7740	0.8166	0.9147	0.7469
3	RFE with RandomForest	5	0.7861	0.8191	0.9302	0.7619
4	ChiSquared	4	0.8493	0.8489	0.9806	0.8324



ACCURACY
0.78
0.80
0.82
0.84
All
RandomForest
RandomForestRFE
ChiSquare
CLASSIFIERS

Classifier used : Support Vector Machine (SVM)

S.NO	FE METHODS	NO. OF PARAMETERS	ACCURACY	PRECISION	RECALL	F-SCORE
1	None	10	0.9126	0.8989	1	0.8989
2	RandomForest	9	0.9126	0.8989	1	0.8989
3	**RFE with RandomForest**	**5**	**0.9246**	**0.9175**	**0.9922**	**0.9104**
4	ChiSquared	4	0.8825	0.8711	0.9961	0.8677

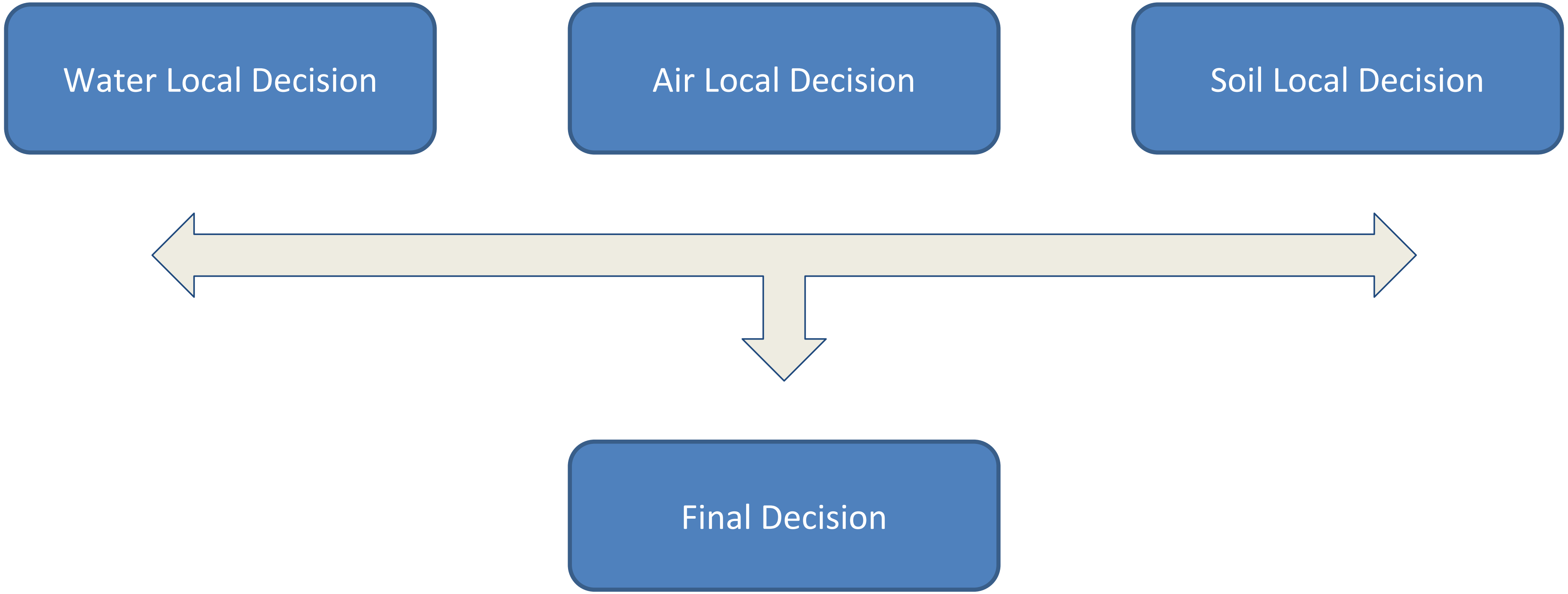


0.92
0.91
0.90
0.89
ACCURACY
All
RandomForest
RandomForestRFE
ChiSquare
CLASSIFIERS

DECISION TREE SOIL

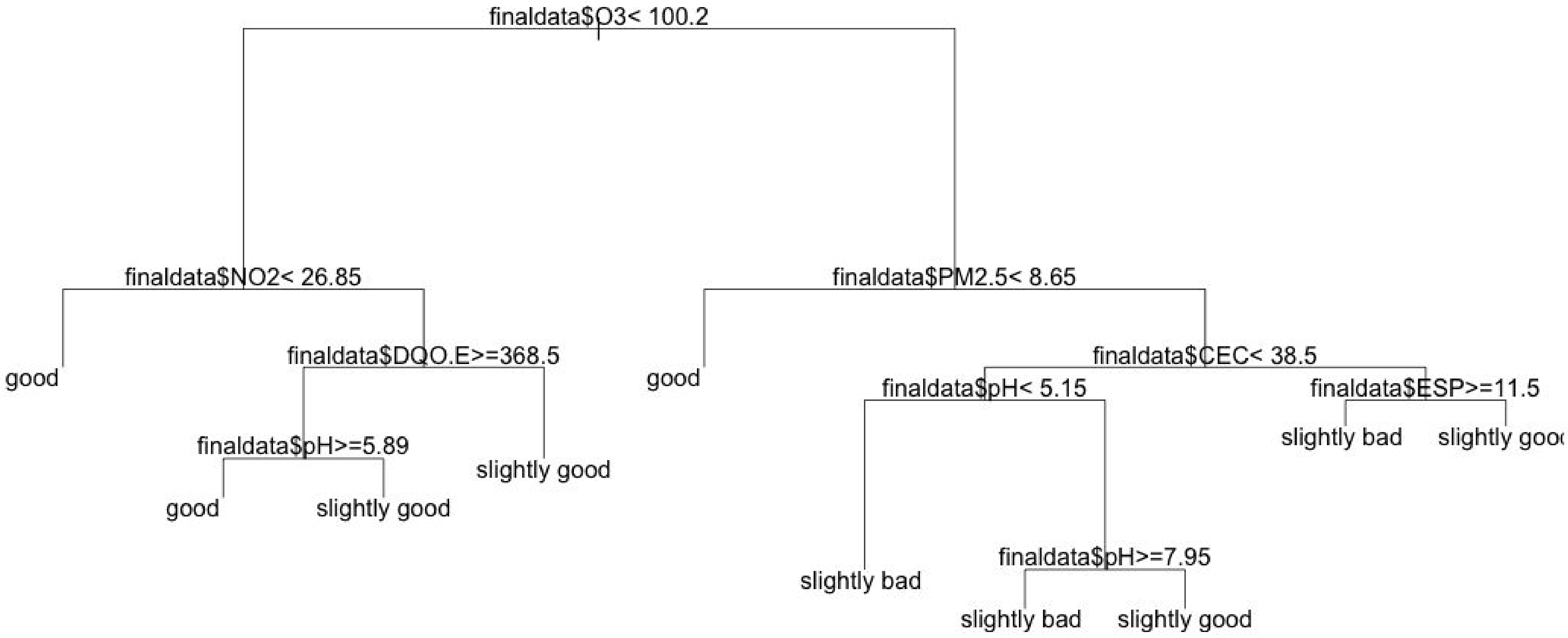
CENTRAL FUSION MODULE

- Local decisions or labels are obtained from the local modules after their respective Decision Trees
- Fused and synthesized using Decision Tree Methods to provide a Final Decision
- Based on the Final Decision, priorities are retrieved whether the water can be used for drinking or not
- Four labels are given : Good, Slightly Good, Slightly Bad, Bad


Water Local Decision
Air Local Decision
Soil Local Decision
Final Decision

COMBINATION OF LOCAL DECISIONS FOR FINAL DECISION

S.No	WATER	AIR	SOIL	FINAL
1	Bad	Bad	Bad	Bad
2	Bad	Bad	Good	Bad
3	Bad	Good	Bad	Bad
4	Bad	Good	Good	Bad
5	Good	Bad	Bad	Slightly Bad
6	Good	Bad	Good	Slightly Good
7	Good	Good	Bad	Slightly Good
8	Good	Good	Good	Good


finaldata$O3< 100.2
finaldata$NO2< 26.85
finaldata$PM2.5< 8.65
good
finaldata$DQO.E>=368.5
good
finaldata$CEC< 38.5
finaldata$pH< 5.15
finaldata$ESP>=11.5
finaldata$pH>=5.89
slightly good
slightly bad
slightly good
good
slightly good
finaldata$pH>=7.95
slightly bad
slightly bad
slightly good

SYSTEM PERFORMANCE

- Accuracy of water local fusion module - 98.18%
- Accuracy of air local fusion module - 93.88%
- Accuracy of soil local fusion module - 92.46%
- **Accuracy of overall water quality monitoring system - 95.25%**

ADVANTAGES OF SYSTEM

- Includes real-time feature
- Supports dynamic decision making
- Data Fusion of all possible parameters that affect water quality - includes air and soil
- Feature Extraction reduces dimension; hence faster processing and accurate results
- cost-effective and includes real-time compared to NWMP
- ensemble learning methods
- more suited to Indian circumstances

FUTURE SCOPE

- Usage of sensor network to acquire data values from river waters
- Testing waters at various zones of the river
- Many other feature extraction methods can be tried
- Can be implemented to give more extensive classes of usage regarding water quality
- Can be upgraded to support many other water resources such as lakes, oceans, etc.

REFERENCES

- Bhardwaj, R. M. "Water quality monitoring in India—achievements and constraints." IWG-Env, International Work Session on Water Statistics, Vienna (2005): 1-12.
- Byer, David, and Kenneth H. Carlson. "Expanded summary: Real-time detection of intentional chemical contamination in the distribution system."Journal (American Water Works Association) 97.7 (2005): 130-133.
- Domingos, Pedro. "A few useful things to know about machine learning. "Communications of the ACM 55.10 (2012): 78-87.
- Han, Jiawei, Micheline Kamber, and Jian Pei. “Data mining: concepts and techniques”. Elsevier, 2011.
- Karami, Ebrahim, Francis M. Bui, and Ha H. Nguyen. "Multisensor data fusion for water quality monitoring using wireless sensor networks." Communications and Electronics (ICCE), 2012 Fourth International Conference on. IEEE, 2012.
- Nakamura, Eduardo F., Antonio AF Loureiro, and Alejandro C. Frery. "Information fusion for wireless sensor networks: Methods, models, and classifications." ACM Computing Surveys (CSUR) 39.3 (2007): 9.
- Pechenizkiy, M., Puuronen, S. and Tsymbal, A., 2003. “Feature extraction for classification in the data mining process.”
- Smith, Richard A., Gregory E. Schwarz, and Richard B. Alexander. "Regional interpretation of water quality monitoring data." Water resources research 33.12 (1997): 2781-2798.

- Standard, Indian. "Drinking water-specification." 1st Revision, IS 10500 (1991). 47
- Stoianov, Ivan, et al. "Sensor networks for monitoring water supply and sewer systems: Lessons from Boston." Proceedings of the 8th Annual Water Distribution Systems Analysis Symposium. 2006.
- “5 Major Causes of Water Pollution in India” - http://www.yourarticlelibrary.com/water-pollution/5-major-causes-of-water-pollution-in-india/19764
- “Chi-Square Test of Independence” - http://www.stattrek.com/chi-square-test/independence.aspx?Tutorial=AP
- “Data Mining - Decision Tree Induction” - http://www.tutorialspoint.com/data_mining/dm_dti.htm
- “Decision Trees” - http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf
- “Feature Extraction Models” - http://topepo.github.io/caret/Feature_Extraction.html
- “Feature Selection with the Caret R Package” - http://machinelearningmastery.com/feature-selection-with-the-caret-r-package/
- “Naive Bayesian” - http://www.saedsayad.com/naive_bayesian.htm
- “National Water Quality Monitoring Programme” - www.cpcb.nic.in/divisionsofheadoffice/pams/NWMP.pdf
- “Pearson’s Chi-Square Test for Independence” - http://www.ling.upenn.edu/~clight/chisquared.htm
- “RandomForest” - https://en.wikipedia.org/wiki/Random_forest
- “Random Forests:some methodological insights” - http://arxiv.org/pdf/0811.3619.pdf
- “Support Vector Machines (SVM) Introductory Overview” - http://www.statsoft.com/textbook/support-vector-machines48
- “Support Vector Machines” - http://scikit-learn.org/stable/modules/svm.html
- “The k-means clustering algorithm” - http://cs229.stanford.edu/notes/cs229-notes7a.pdf

DATASETS

- “Water Treatment Plant Dataset” - http://archive.ics.uci.edu/ml/machine-learning-databases/water-treatment/
- “Air pollutant concentrations 2013” - http://www.eea.europa.eu/data-and-maps/data/air-pollutant-concentrations-at-station/pollutant-concentrations-by-city/air-pollutant-concentrations-2013-dataset-cities
- “ISRIC/WDC-Soil Dataset” - http://www.isric.org/content/download-form?dataset=SOTWIS_BR.zip

Thank You!