Agriculture data analytics

Team members	:	Dheeraj Raghavendra Purnachand Asish Varanasi
Mentors	:	Dr. Avinash Sharma Dr. P.Krishna Reddy

Motivation

- Agriculture plays a vital role in economy for developing countries like India.
- Rapid technological advancement - but not much focus on agricultural domain.
- Several newly emerging problems in agriculture.
- In 2012, the NCRB of India reported 13,754 farmer suicides.
- Hence, a need for guidance to the farmers.
- Esagu is one such platform.

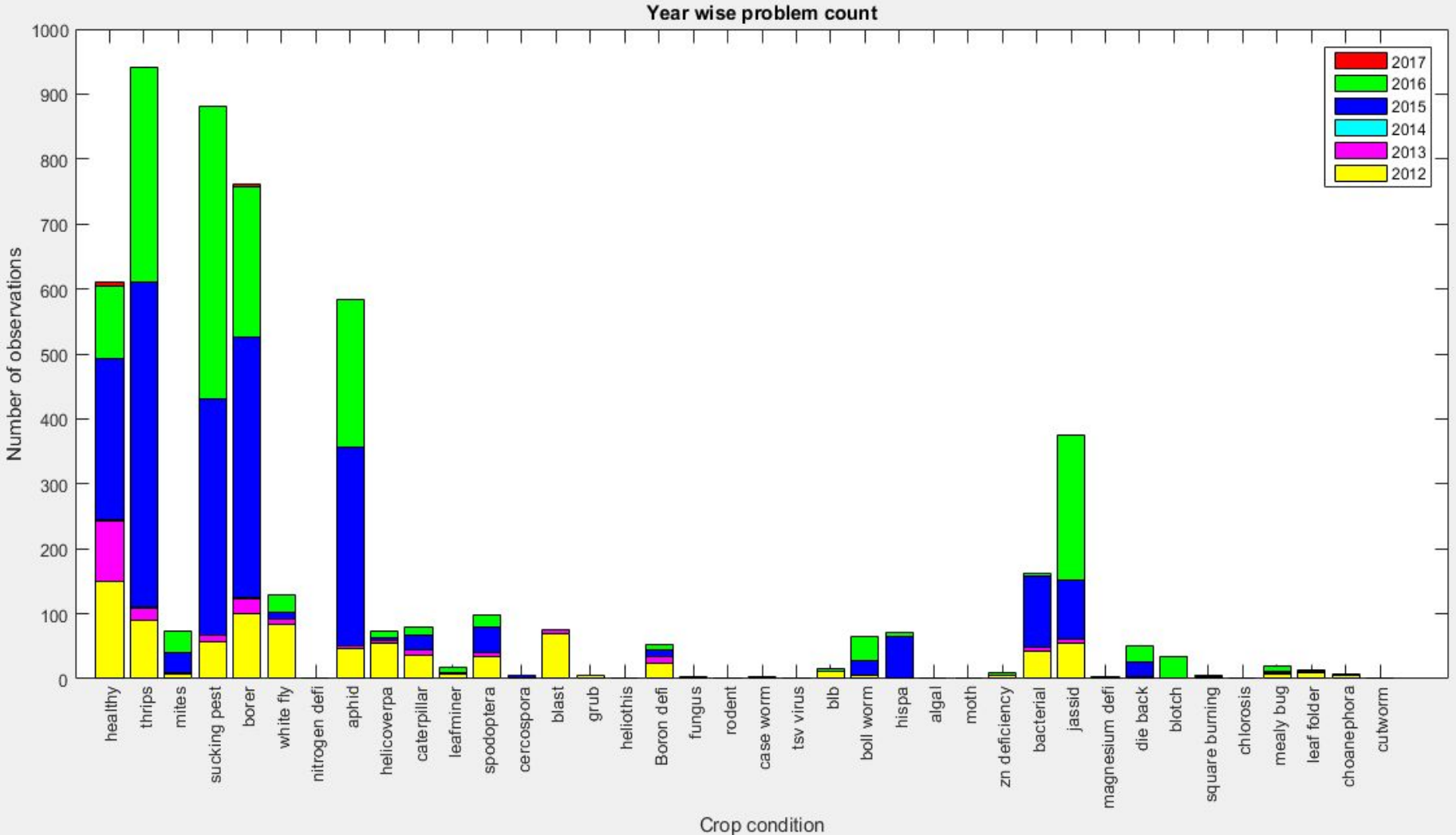
Key pain points

Recommendation of Diagnostics:

- Accuracy
- Timely advice
- Minimum required input
- Ease of communication
- Cost effectiveness
- Feedback system

Data

- Total of 54 crop species
- Keywords were manually searched to find possible diseases
- Filtering resulted in 39 diseases
- Only 19 diseases with overall count greater than 5.
- Farms are concentrated around Rajendranagar mandal.
- Weather is assumed to be same as that of rajendranagar mandal.
- Correlation with weather and disease prediction using the given data.


Year wise problem count
Number of observations
Crop condition
2017
2016
2015
2014
2013
2012
0
100
200
300
400
500
600
700
800
900
1000
healthy
thrips
mites
sucking pest
borer
white fly
nitrogen defi
aphid
helicoverpa
caterpillar
leafminer
spodoptera
cercospora
blast
grub
heliothis
Boron defi
fungus
rodent
case worm
tsv virus
blb
boll worm
hispa
algal
moth
zn deficiency
bacterial
jassid
magnesium defi
die back
blotch
square burning
chlorosis
mealy bug
leaf folder
choanephora
cutworm

Input format

- Weather data contains 8 weather parameters
- The following 3 more parameters are used along with weather parameters:
  - Crop species : Obtained from crop id
  - Season : Known
  - Age of the crop : Calculated using date of sowing
- Total of 11 input parameters

Models Implemented

The following models are implemented on the available data.

- Bayesian estimation.
- Neural networks.
- Support vector machine.

Bayesian Estimation

For each season, we first fix the crop species. Then for each species of crop, we find the following probabilities:

- Probabilities of occurrence for all the 19 problems i.e, p(problem) for given species of crop.
- Probability of weather parameter being at certain value, given a particular problem i.e, p(weather parameter/problem). We estimate this by finding the values of all the weather parameters corresponding to the dates of problem occurrence.
- Probability of the crop being at certain age, given a particular problem i.e, p (age of crop/problem).

Bayesian Estimation

Then, for each problem, we find this probability :

$$p(\text{problem}/[\text{temperature,humidity,age,other weather parameters}]) =$$

$$\frac{p([\text{temp,humidity,age,other parameters}]/\text{problem}) * p(\text{problem})}{p([\text{temp,humidity,age,other parameters}])}$$

$$= \frac{p(\text{temp}/pb)*p(\text{humidity}/pb)*p(\text{age}/pb)*p(\text{other}/pb)*p(pb)}{p(\text{temp})*p(\text{humidity})*p(\text{age})*p(\text{other parameters})}$$

(Assuming independence between different parameters)

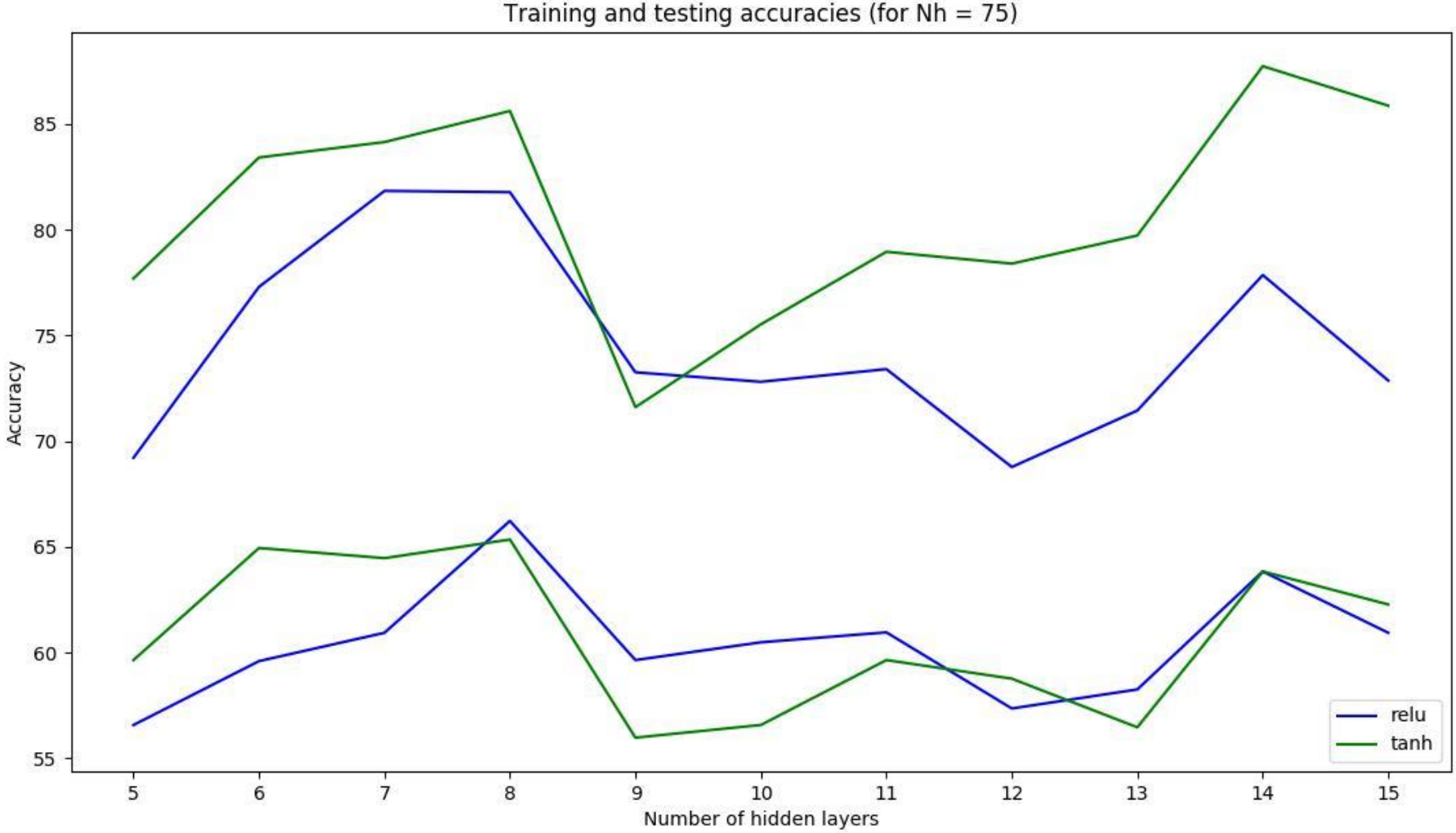
Bayesian results

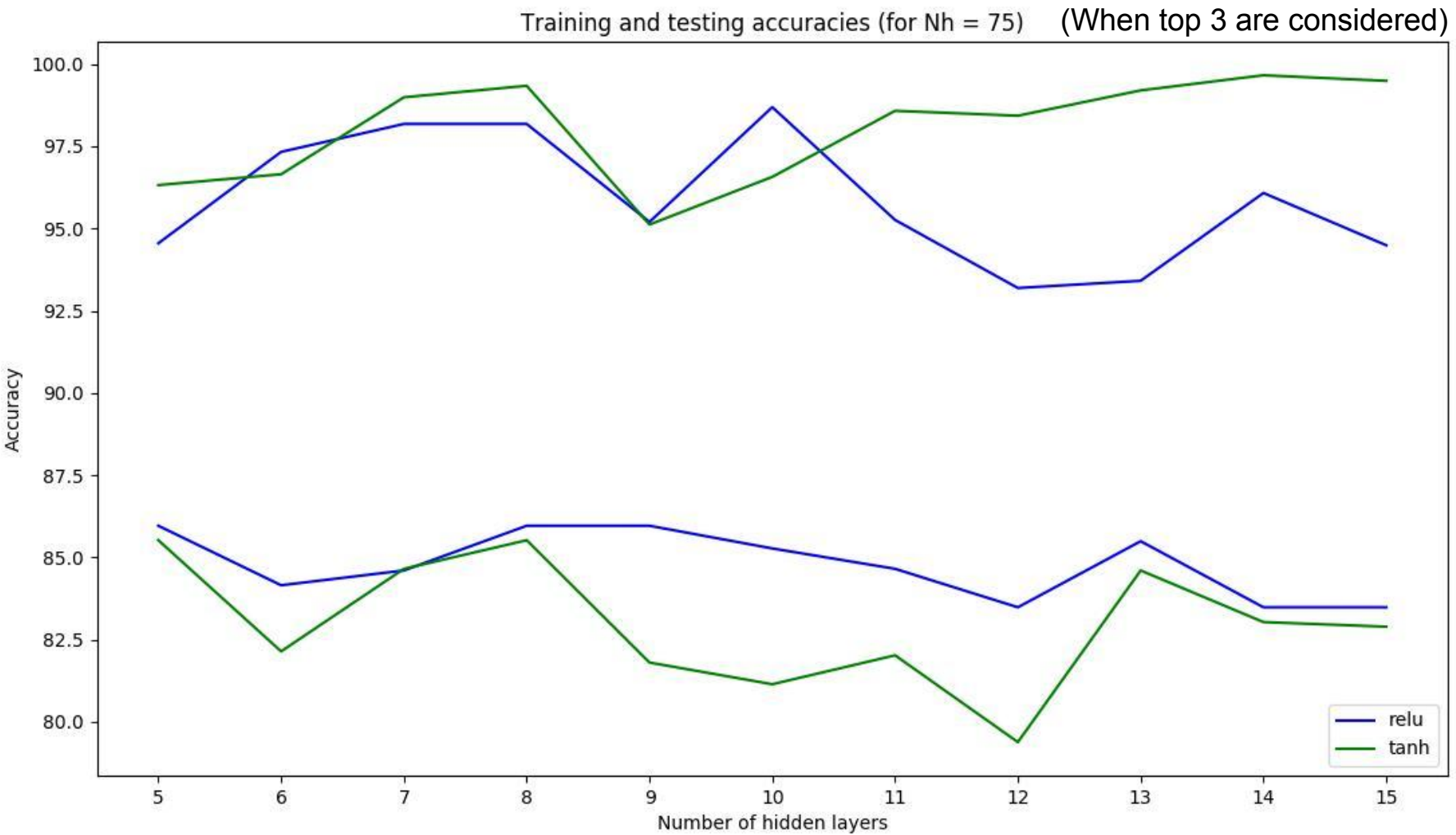
Accuracies	Fold 1	Fold 2	Fold 3
Top-1	53.22	52.91	57.65
Top-3	80.88	80.55	80.40

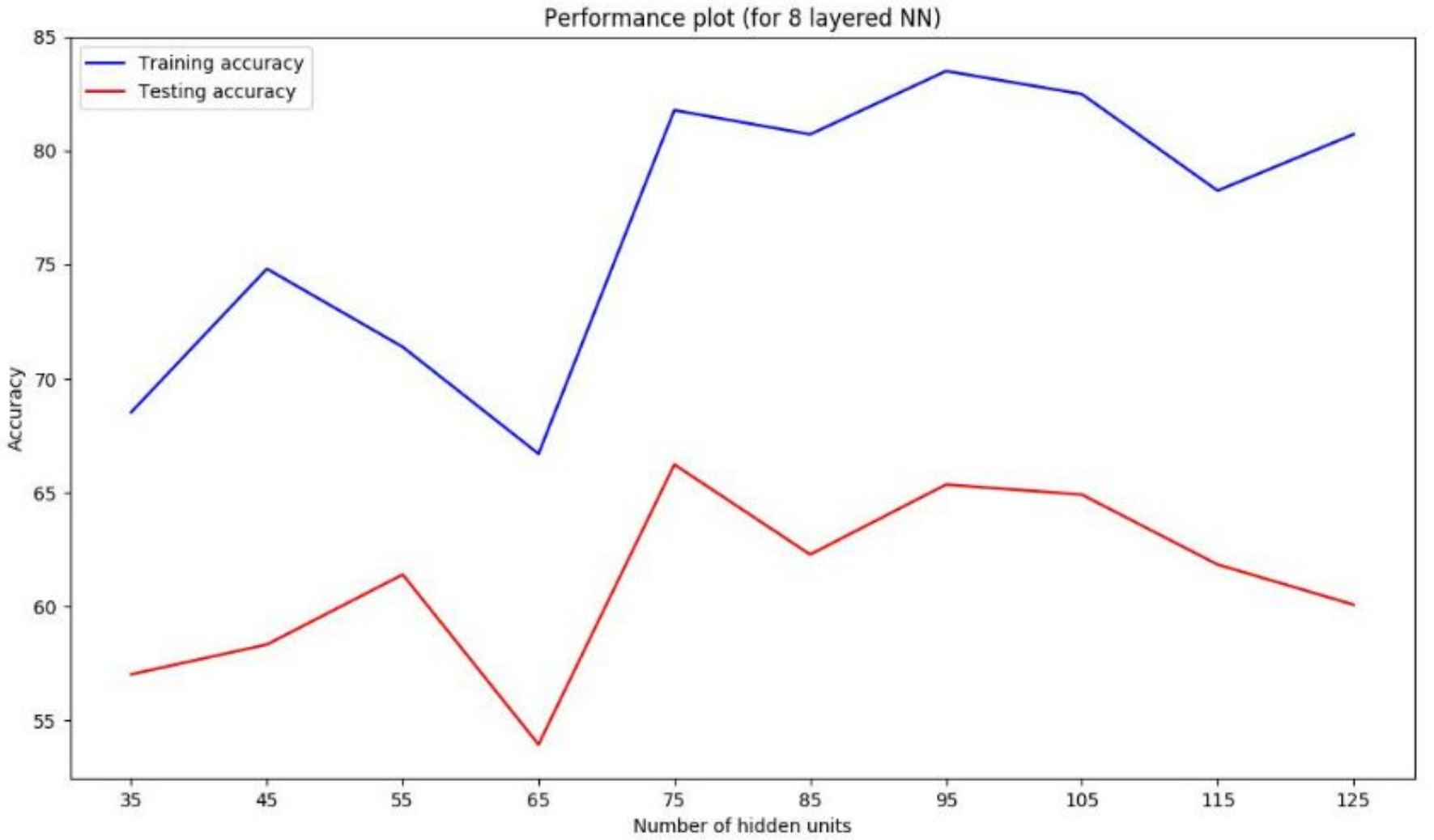
Type	Average accuracies
Top-1	54.6
Top-3	80.61

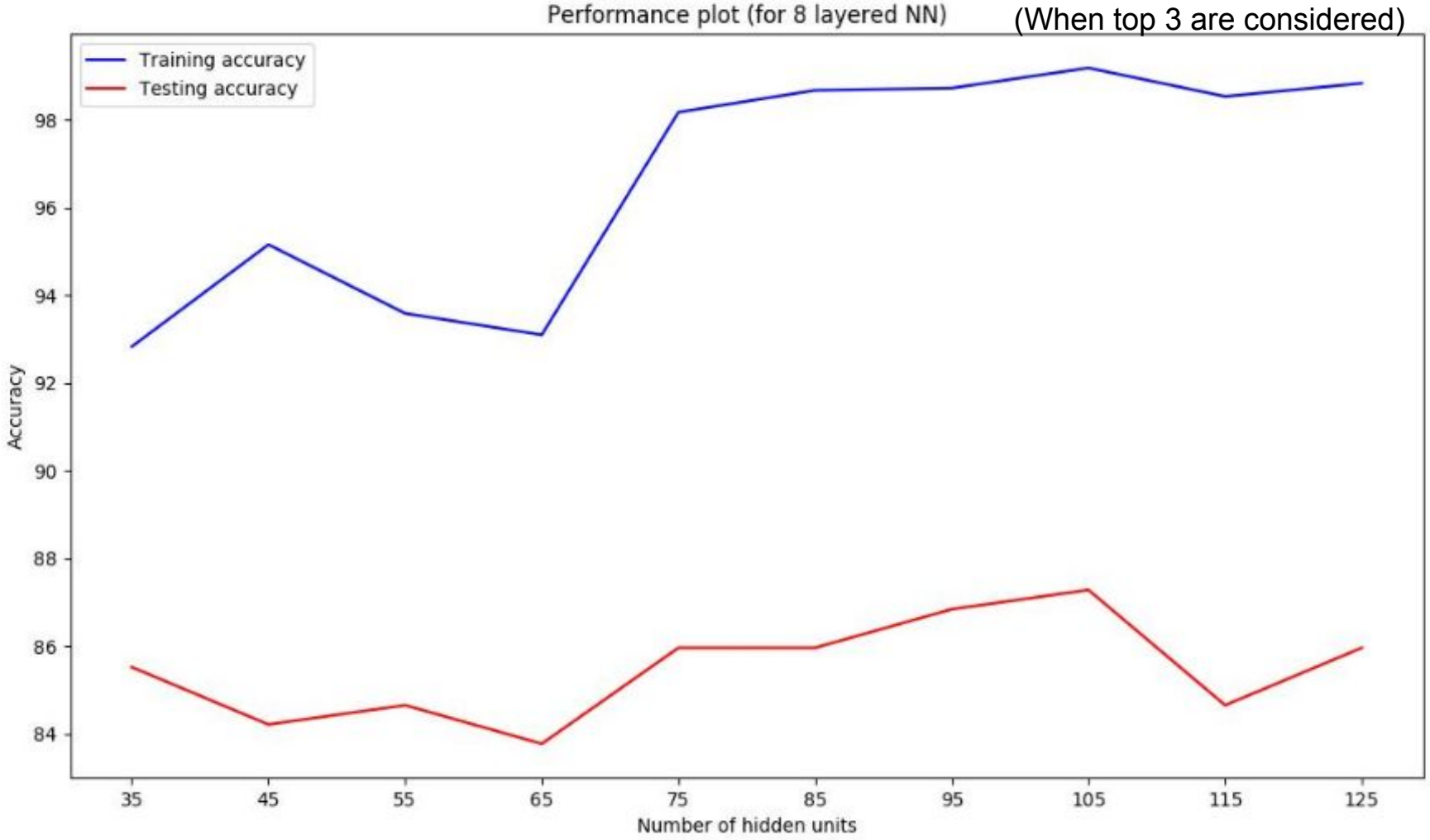
Neural Networks

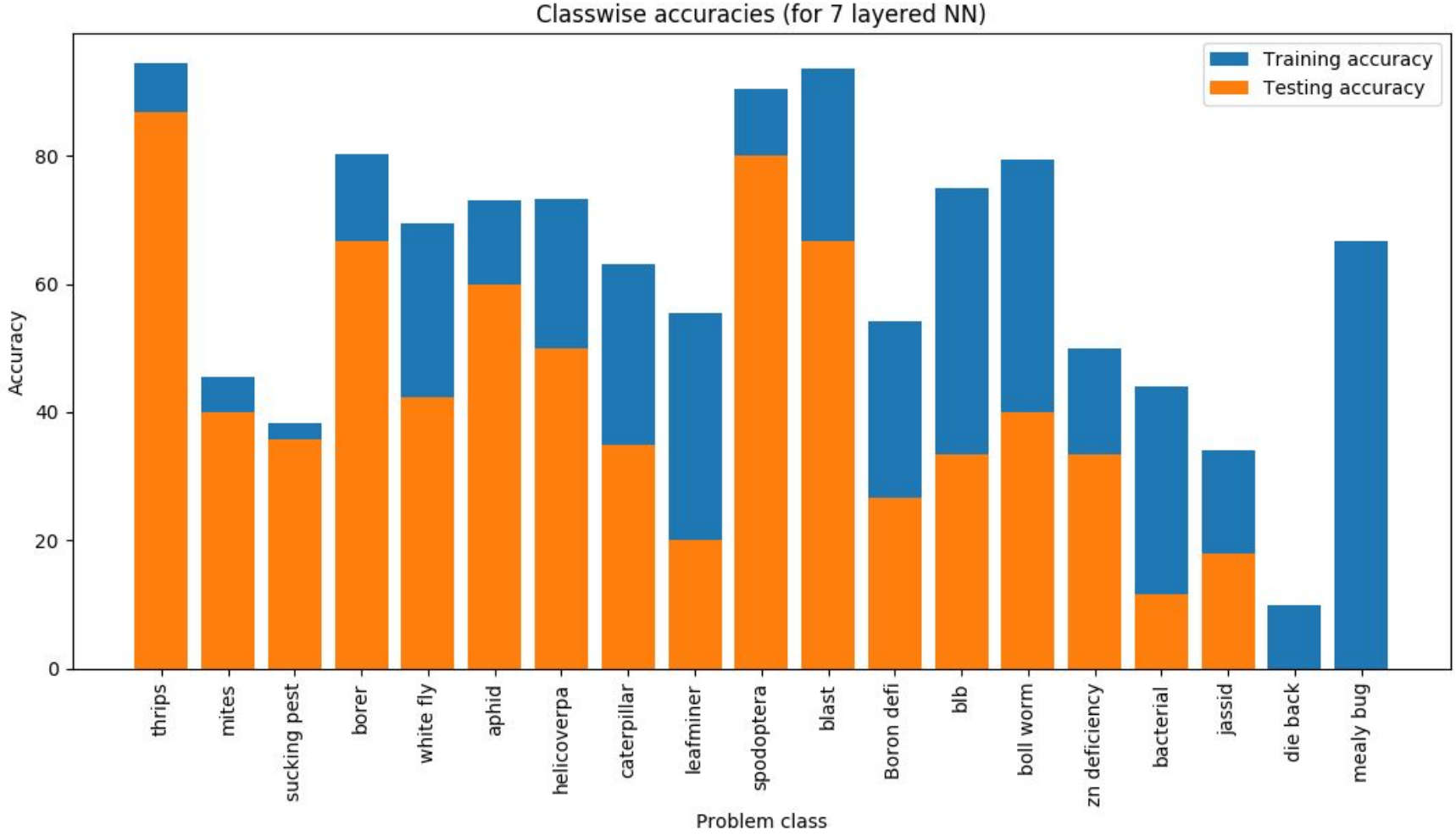
- Input has 11 parameters(3 crop parameters,8 weather parameters)
- Only 19 problems have more than 5 instances.
- Hence, Number of output neurons = 19.
- Implemented neural networks for varying number of hidden layers and different activation functions.
- Optimal value found at 8 intermediate layers and learning rate as 1, with relu activation function.
- Maximum accuracy obtained is 66.23%
- 85.96% when top 3 outputs are considered.


Training and testing accuracies (for Nh = 75)
Accuracy
Number of hidden layers
relu
tanh
85
80
75
70
65
60
55
5
6
7
8
9
10
11
12
13
14
15


Training and testing accuracies (for Nh = 75)
(When top 3 are considered)
100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5
80.0
Accuracy
5
6
7
8
9
10
11
12
13
14
15
Number of hidden layers
relu
tanh


Performance plot (for 8 layered NN)
Training accuracy
Testing accuracy
Accuracy
85
80
75
70
65
60
55
35
45
55
65
75
85
95
105
115
125
Number of hidden units


Performance plot (for 8 layered NN)
(When top 3 are considered)
Training accuracy
Testing accuracy
Accuracy
98
96
94
92
90
88
86
84
35
45
55
65
75
85
95
105
115
125
Number of hidden units

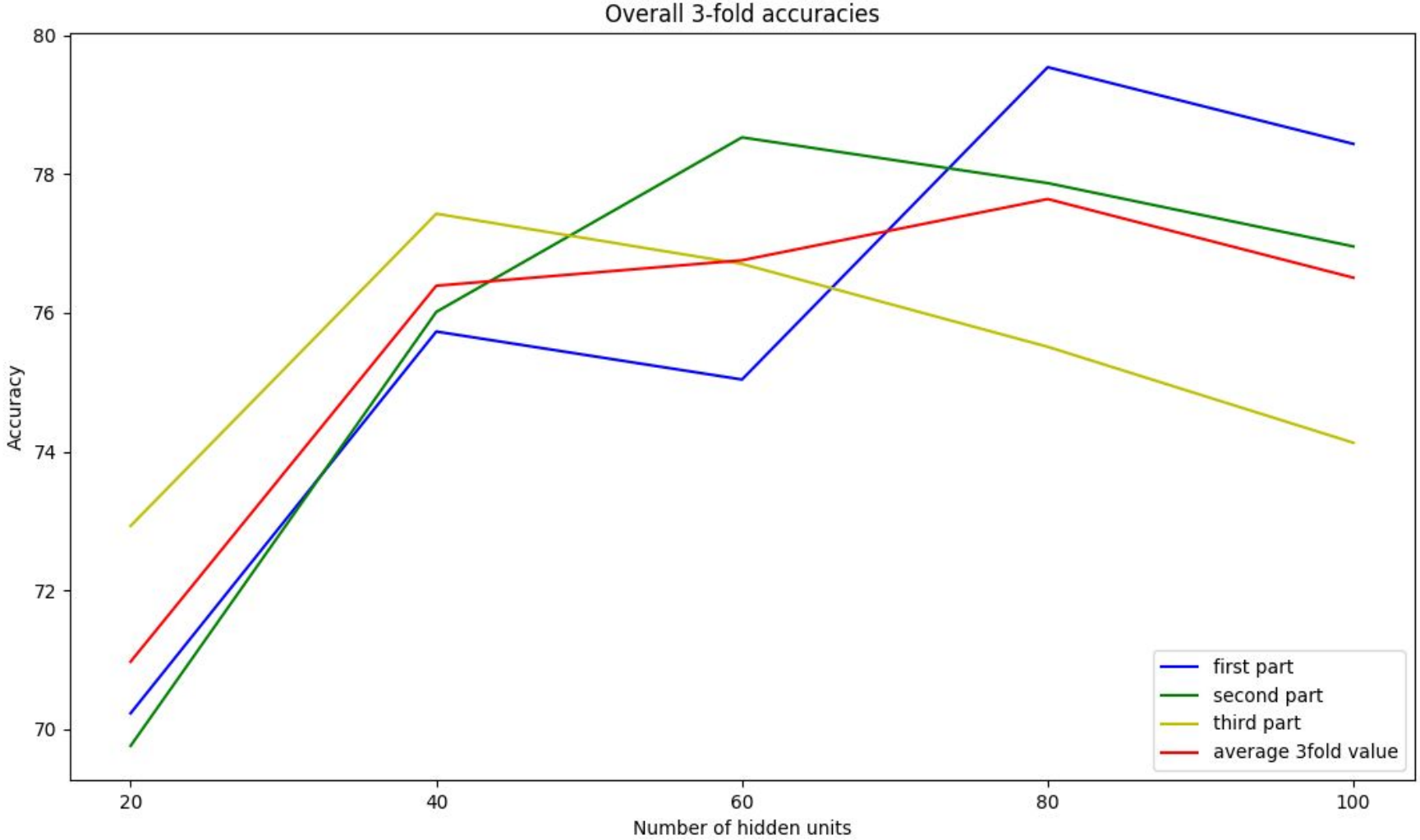

Classwise accuracies (for 7 layered NN)
Training accuracy
Testing accuracy
Accuracy
0
20
40
60
80
thrips
mites
sucking pest
borer
white fly
aphid
helicoverpa
caterpillar
leafminer
spodoptera
blast
Boron defi
blb
boll worm
zn deficiency
bacterial
jassid
die back
mealy bug
Problem class

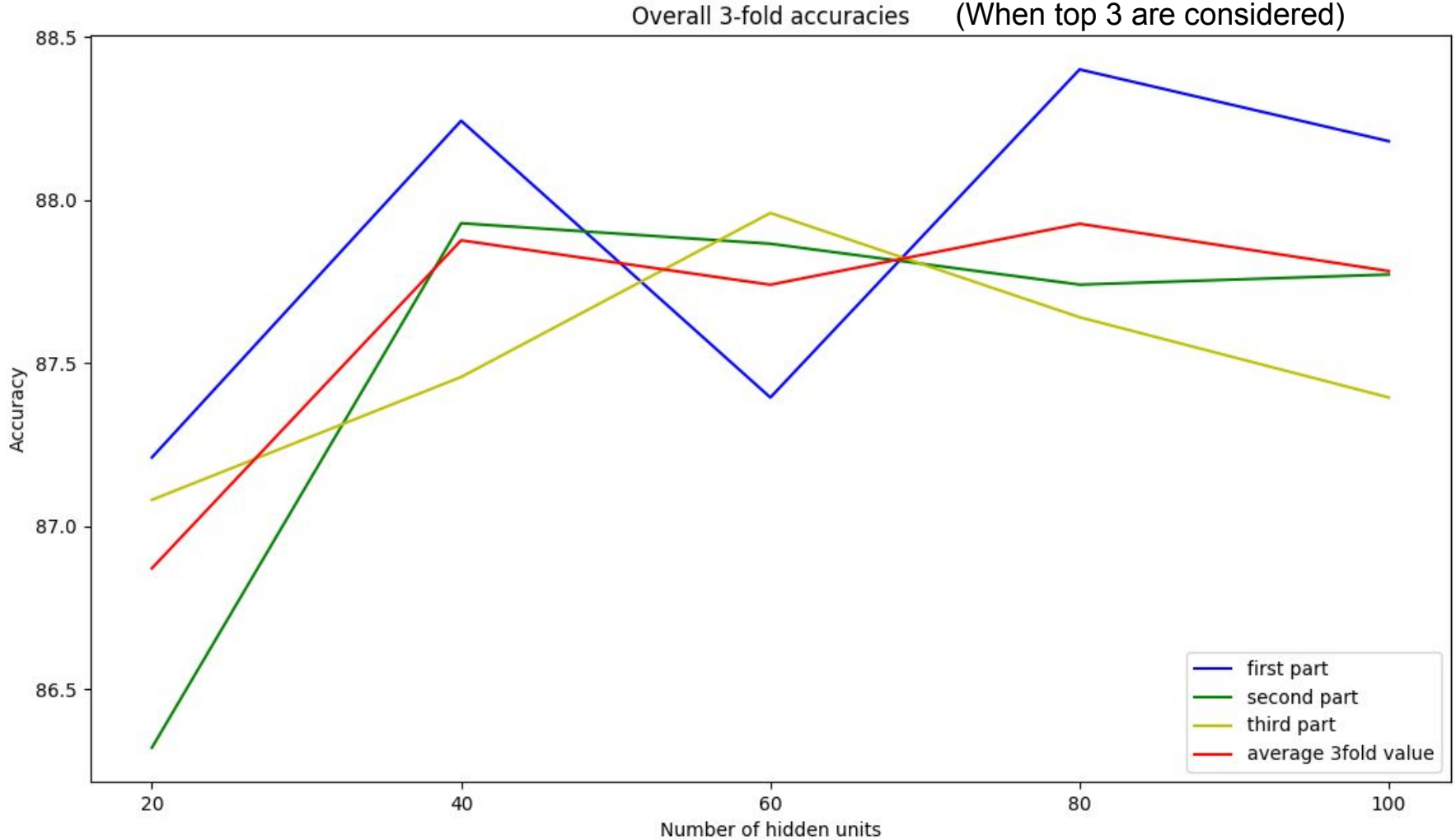
Meta classifier

- Calculated classwise accuracies for implemented neural networks
- Some classes have very less accuracies
- Hence,divided into 2 groups of sizes 12 and 7 each
- Testing accuracies of 72% for both the groups when trained separately
- Around 93% when top 3 are considered.
- Hence, different neural networks are trained for the 2 groups.

Meta classifier contd.

- Built a classifier to decide which class a particular input belongs to.
- Based on the class, the corresponding trained weights are used.
- Maximum accuracy obtained is 77.64%.
- 87.92% when top 3 outputs are considered.


Overall 3-fold accuracies
80
78
76
74
72
70
Accuracy
20
40
60
80
100
Number of hidden units
first part
second part
third part
average 3fold value


Overall 3-fold accuracies
(When top 3 are considered)
88.5
88.0
87.5
87.0
86.5
Accuracy
20
40
60
80
100
Number of hidden units
first part
second part
third part
average 3fold value

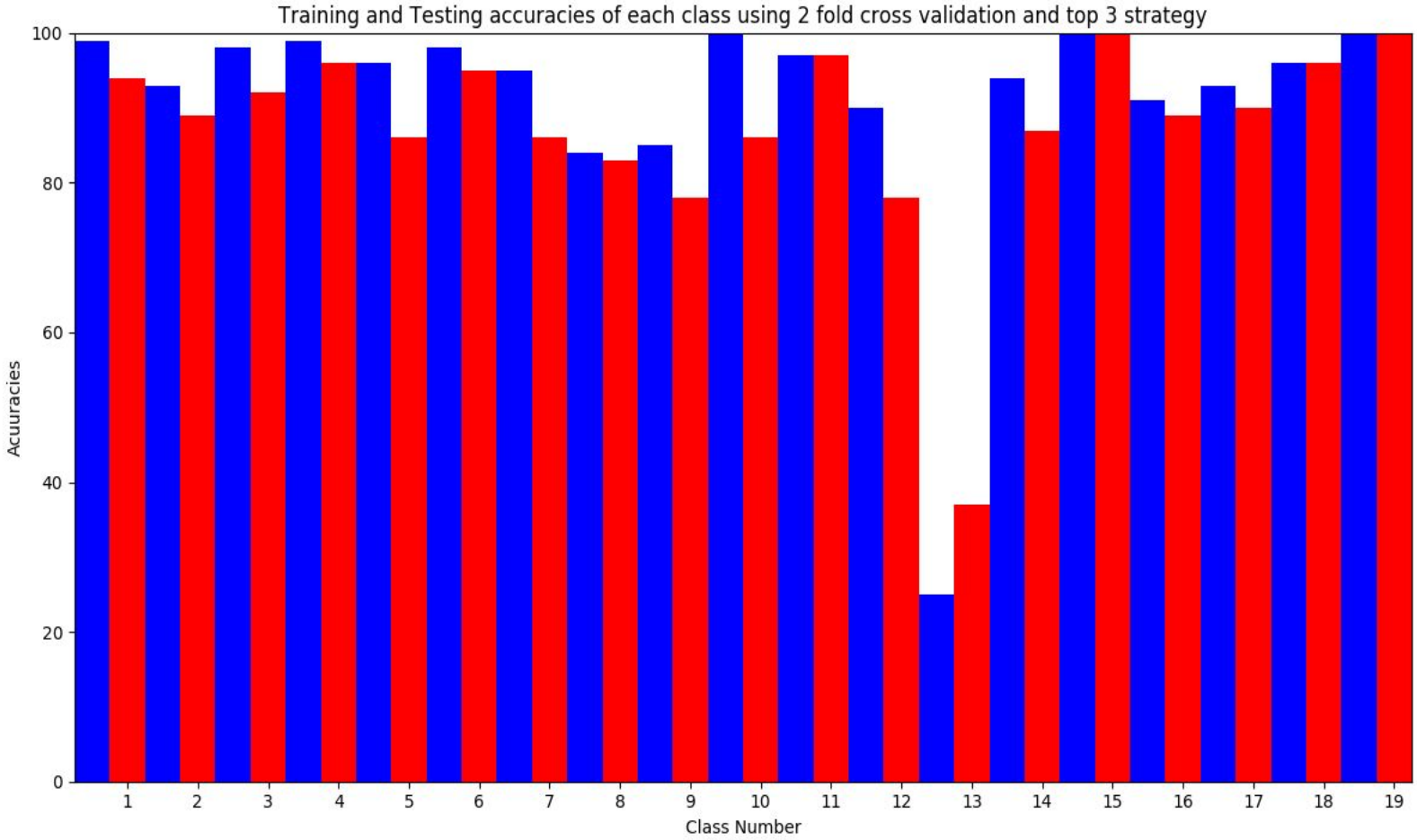
SVM

Approach 1

Simple Multiclass SVM

Tried varying all possible parameters like penalty term, kernel function, degree in polynomial and power in case of exponential kernel, maximum iterations and other parameters. But there is no much change in the accuracy obtained. So fixed the model with polynomial kernel with degree 5, penalty term as 1 and max iterations as 9000.

Also top 3 strategy has been used to compute the accuracies in all models.

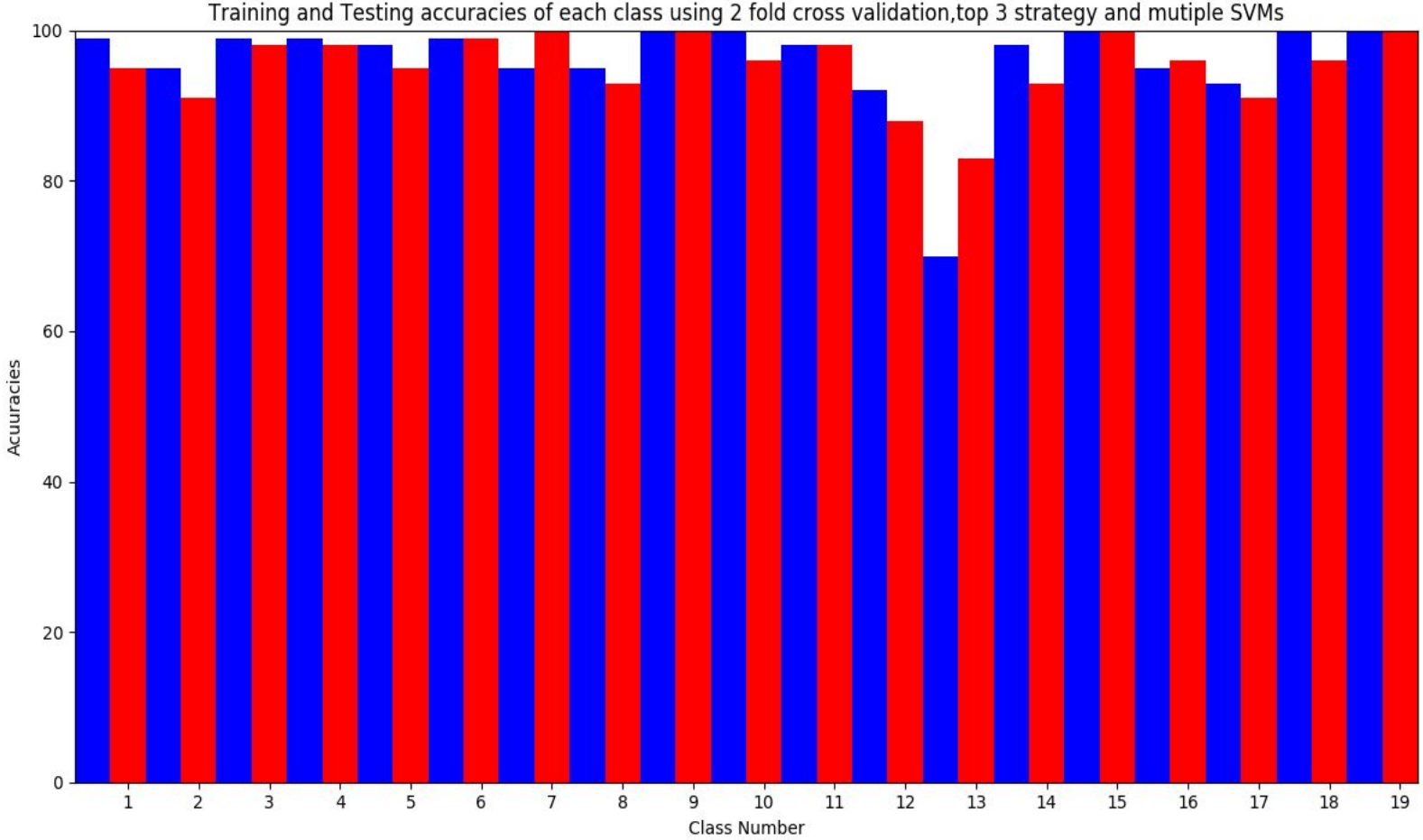

Training and Testing accuracies of each class using 2 fold cross validation and top 3 strategy
Acuuracies
Class Number
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

Approach 2

Here the data is fed into multiple SVMs. In the previous bar plot we can see that some classes have low accuracy. The reason can be due to low data with respect to others. So fed these classes into a new SVM i.e. all classes with low data. Even this did not show much improvement.

So finally divided the data into two parts such a way that both have approximately same data and equally balancing classes with high data points and one with low data points.

This has improved the testing accuracy from 93%(Single SVM) to 97%(Multiple SVM) using top 3 strategy.


Training and Testing accuracies of each class using 2 fold cross validation,top 3 strategy and mutiple SVMs
Acuuracies
Class Number
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

2 fold accuracy for single SVM

	Fold 1	Fold 2
Training accuracy	97.9	97.1
Testing accuracy	94.1	92.1

2 fold accuracy for Multiple SVMs

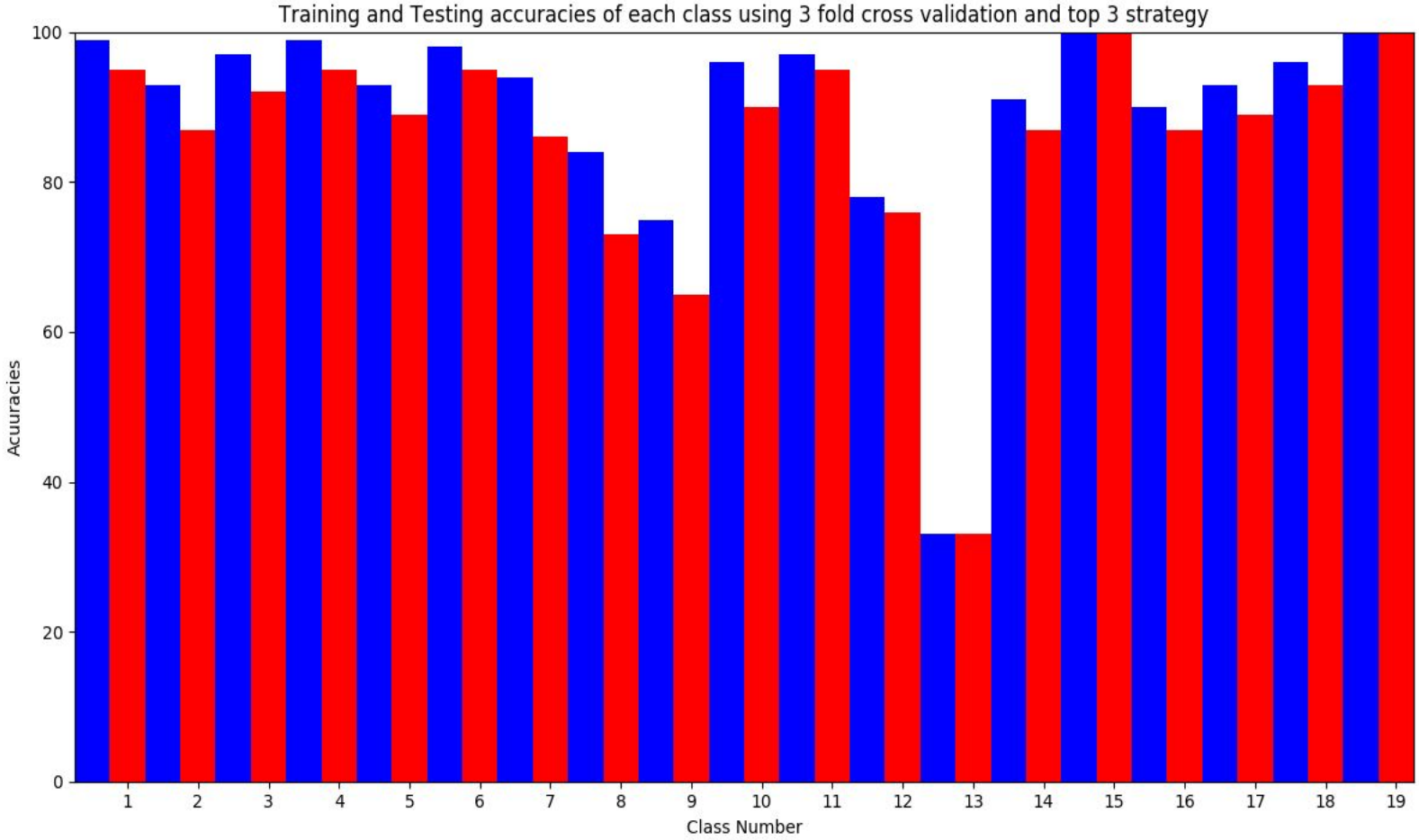
	Fold 1	Fold 2
Training accuracy	98.8	98.6
Testing accuracy	97.7	96.7

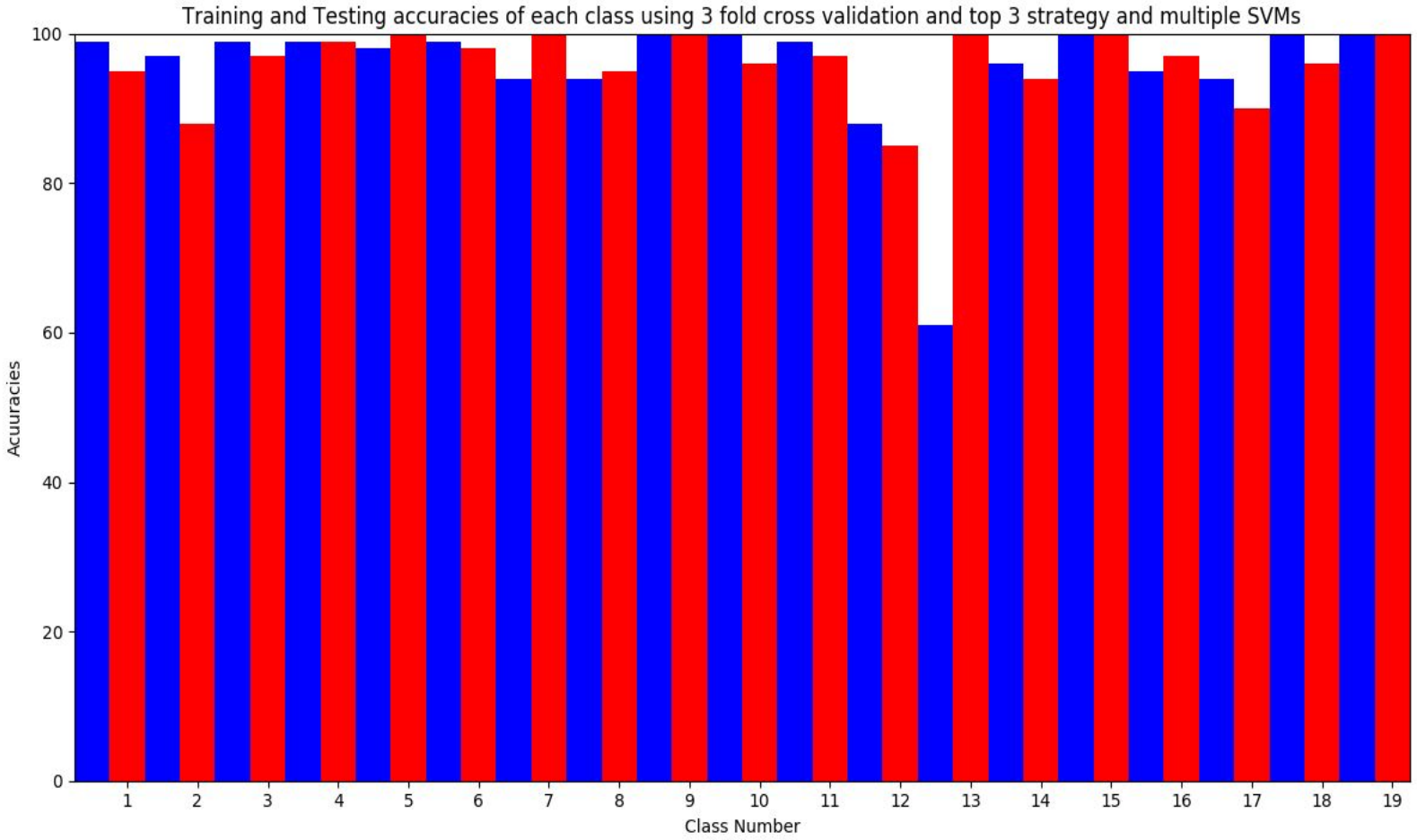
3 fold accuracy for single SVM

	Fold 1	Fold 2	Fold 3
Training accuracy	97.2	97.3	96.6
Testing accuracy	94.2	92	92.1

3 fold accuracy for Multi SVM

	Fold 1	Fold 2	Fold 3
Training accuracy	98.7	98.7	99.2
Testing accuracy	95.7	97.3	97.8


Training and Testing accuracies of each class using 3 fold cross validation and top 3 strategy
Acuuracies
Class Number
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19


Training and Testing accuracies of each class using 3 fold cross validation and top 3 strategy and multiple SVMs
Acuuracies
Class Number
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

Conclusion

- Input advice data is not uniform
- Weather is assumed to be same across all the farms
- Less number of input samples
- Results show that weather correlation is possible
- Model can be improved by using the actual weather information of farms.
- Will provide a solution to major problem in agricultural domain

Thankyou