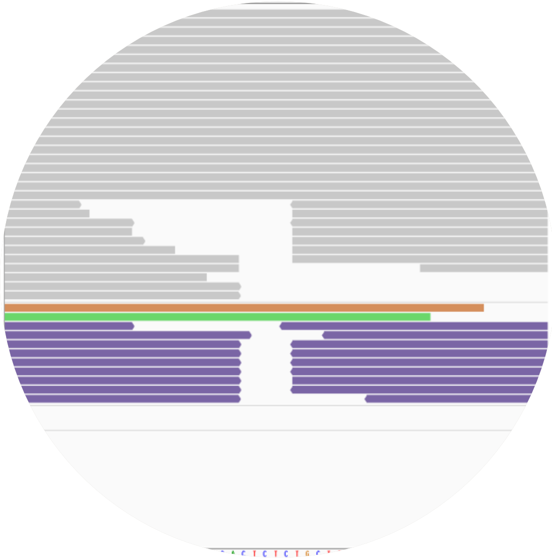
Deep learning approach to classify false and true positive chromosomal translocations

Ina Hulsegge, Aniek Bouwman, Roel Veerkamp and Claudia Kamphuis

Big Data Network lunch meeting 20-02-2020



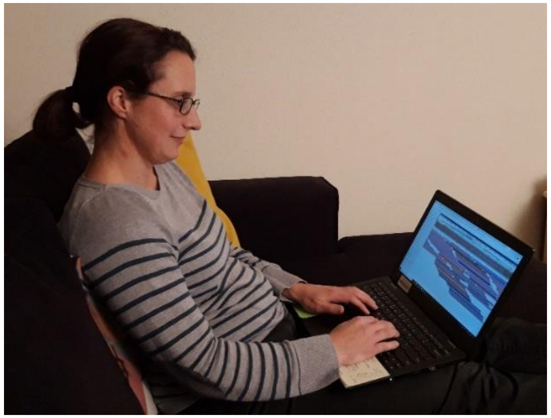
WAGENINGEN
UNIVERSITY & RESEARCH

100years
1918 — 2018

KB Artificial Intelligence

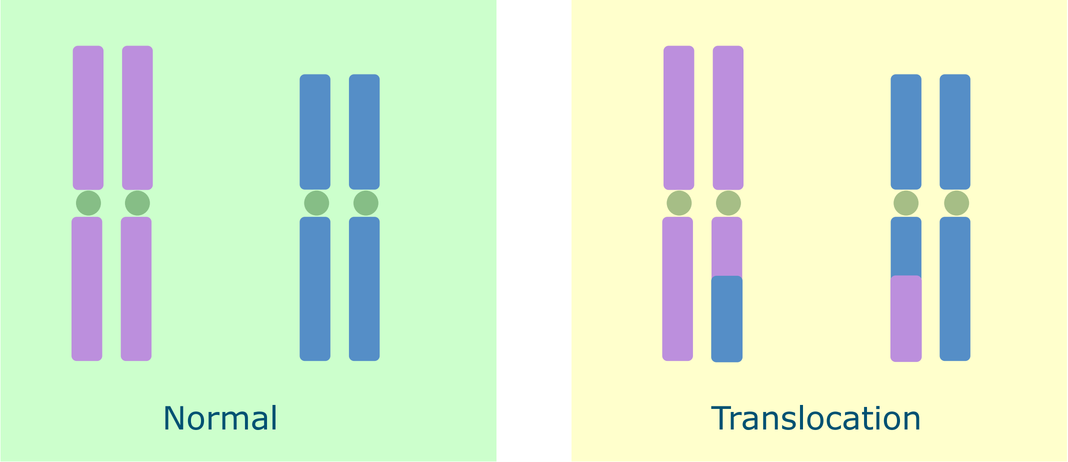
- Building knowledge on utilizing unstructured (images) data, by applying new AI techniques (deep learning).

- Case study



- Can machine learning reduce, or replace, the manual inspection of structural variants.

Translocations



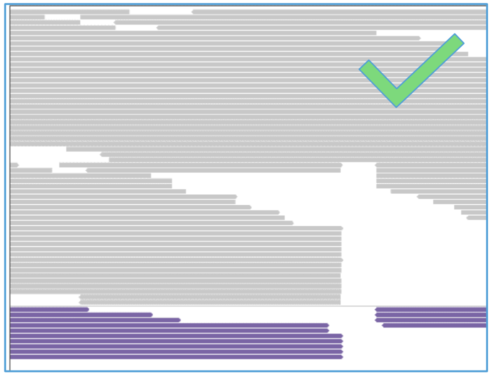
- change in location of a chromosomal segment

Motivation

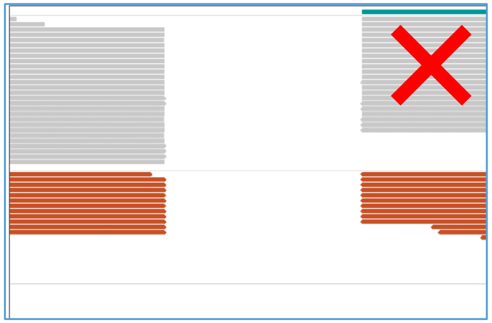
- Important to detect structural variants
- Bioinformatic tools to detect SVs
- False positives ➔ Manual inspection needed
- Manual inspection is time-consuming, costly and poorly standardized

Challenge

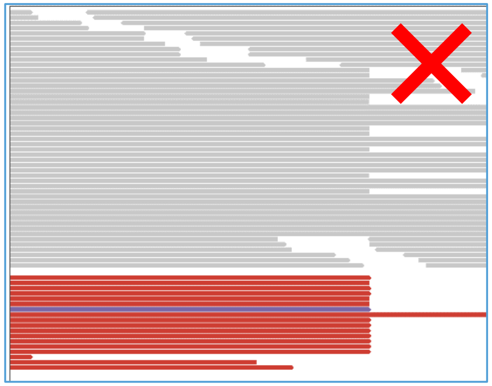
Good



No normal reads



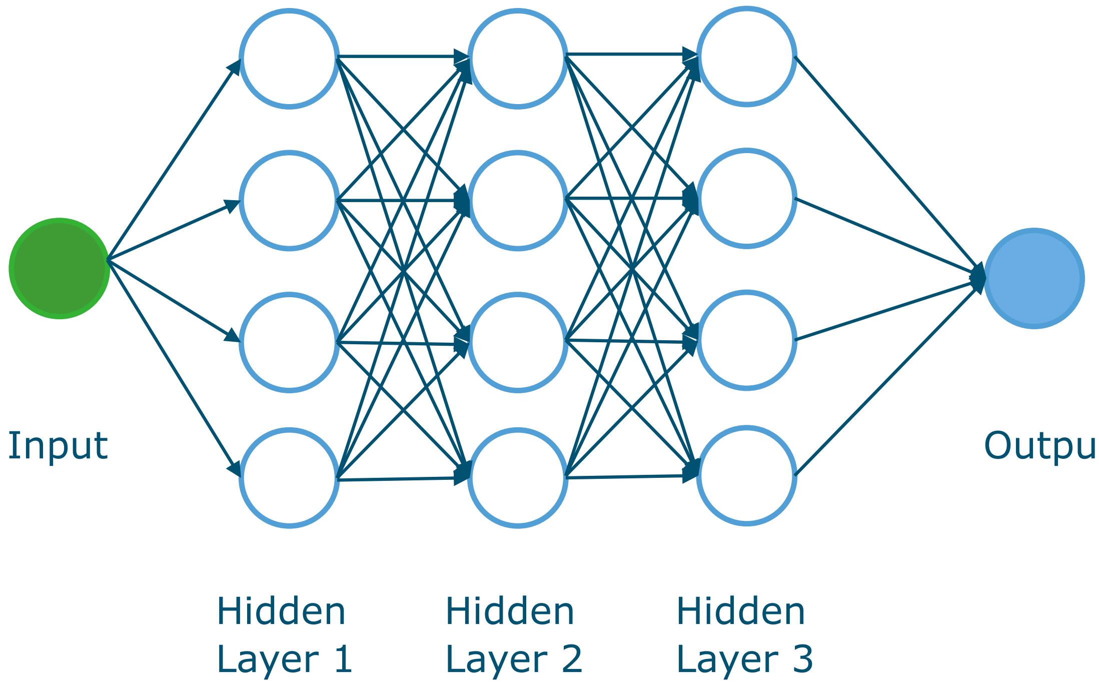
Overlap forward & reverse



- Develop a Deep Learning model to classify images of sequence reads into false and true translocations with the ultimate goal to replace manual inspection

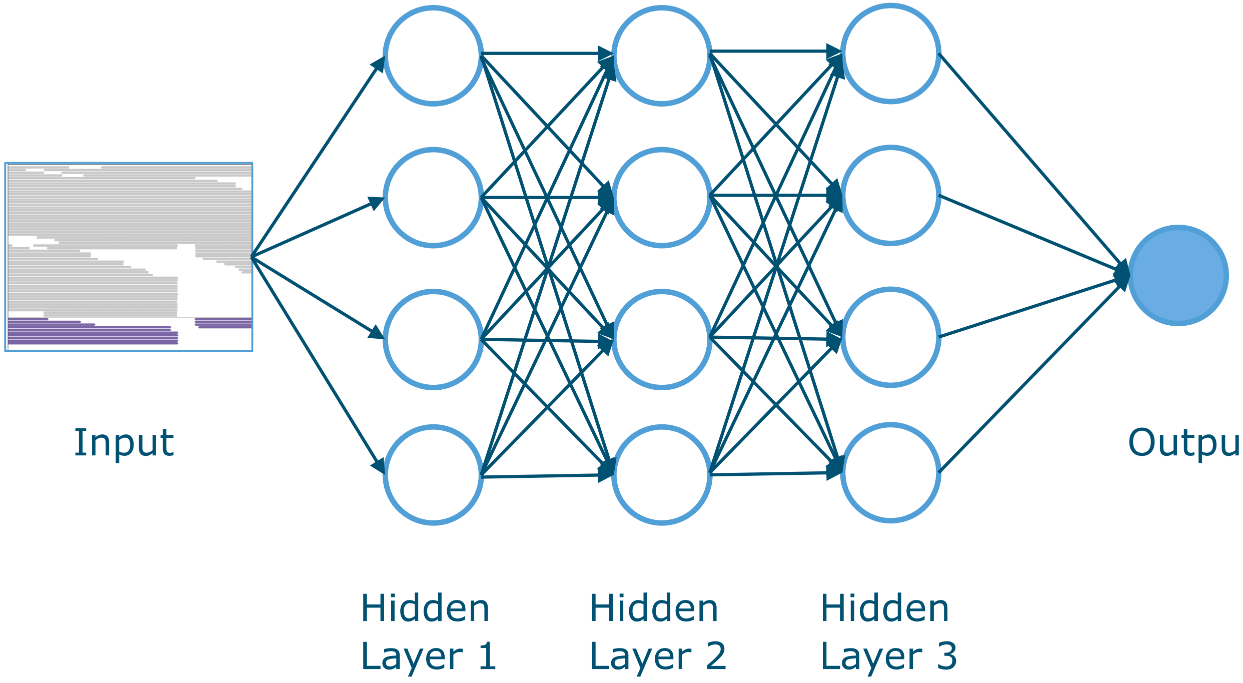
Deep learning

- Machine learning technique
- Continually analyze data with a logic structure
- Learn from unstructured data.
- Multiple layers

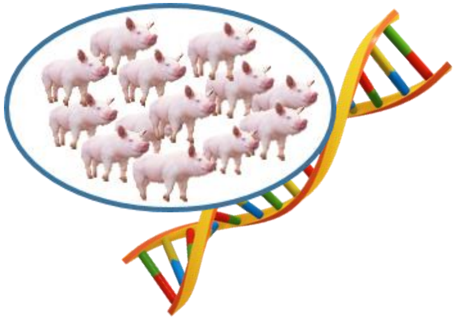


Convolutional neural networks

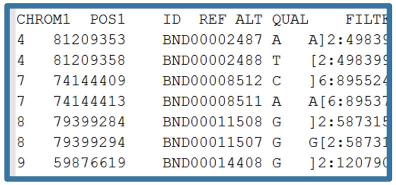
CNN is a **multi-layered** neural network with a unique architecture designed to **extract increasingly complex features** of the data **at each layer** to determine the output



Data



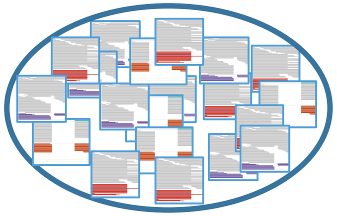
39 pigs



```
CHROM1   POS1      ID  REF ALT QUAL    FILTE
4   81209353    BND00002487 A   A]2:49839
4   81209358    BND00002488 T   [2:498399
7   74144409    BND00008512 C   ]6:895524
7   74144413    BND00008511 A   A[6:89537
8   79399284    BND00011508 G   ]2:587315
8   79399294    BND00011507 G   G[2:58731
9   59876619    BND00014408 G   ]2:120790
```

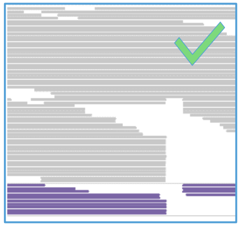
List of identified translocations (Delly output)

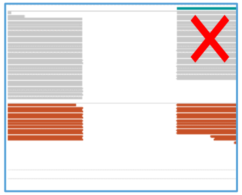
2992 detected translocations

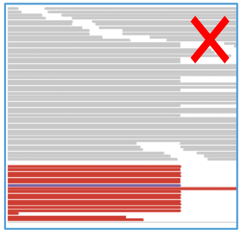


2992 images

Annotation







544

2448

Data

- Training and validation set: 34 animals
  - 80% for model training;
  - 20% for model validation.
- Performance evaluation: 5 animals

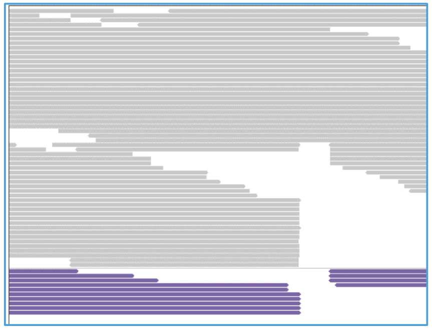
Data set	#animals	% data	#false	#true
Training	34	80	1723	395
validation	34	20	431	99
Testing	5	100	294	50

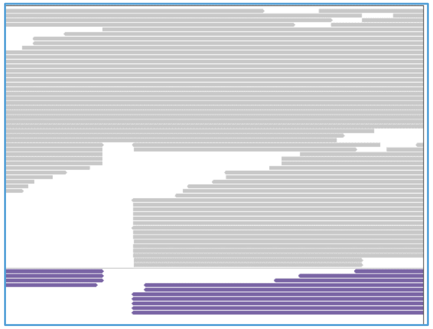
Automatic generation of images

- Integrative Genomics Viewer (igv)
- Automatic generation of the images based on chromosome number and position

Data augmentation

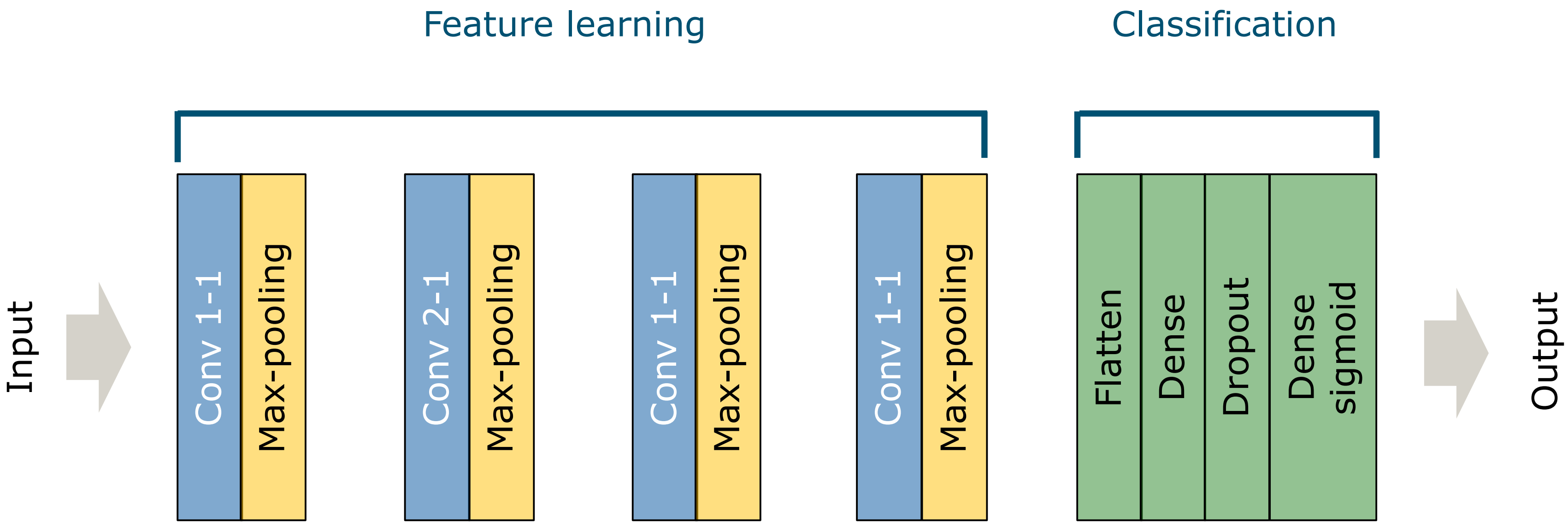
- Image data augmentation
  - artificially expand the size of dataset
  - creating modified versions of images
- shift, flip, brightness, zoom



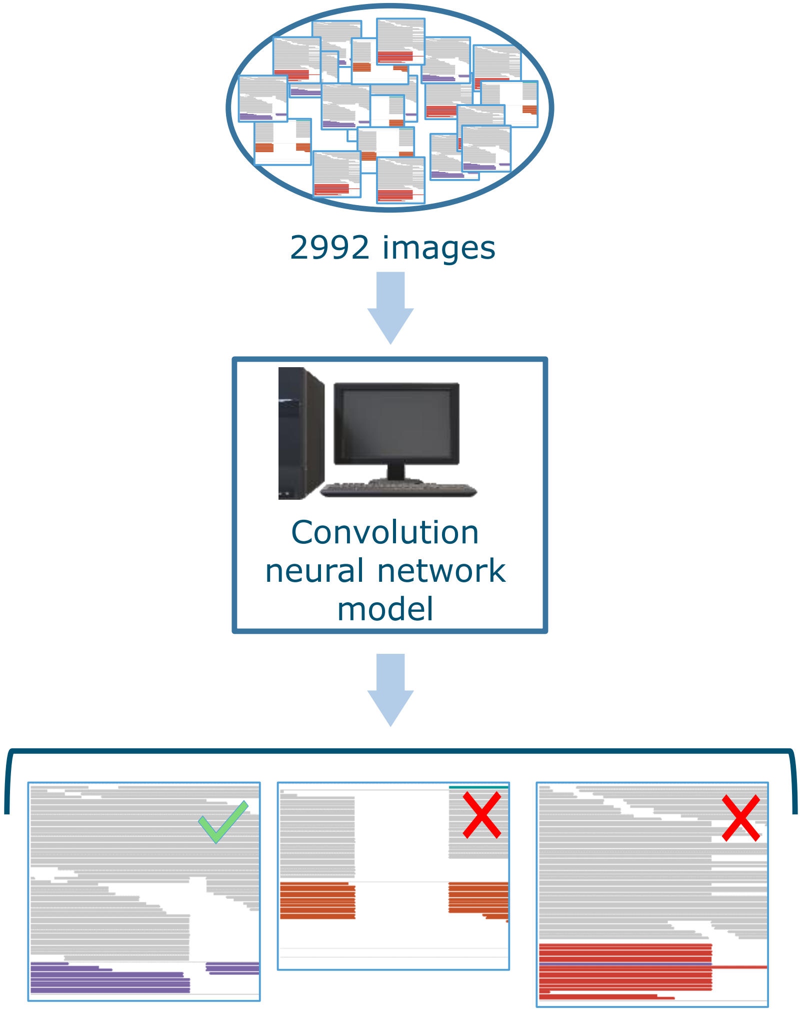


Horizontal flip

VGG-like convnet

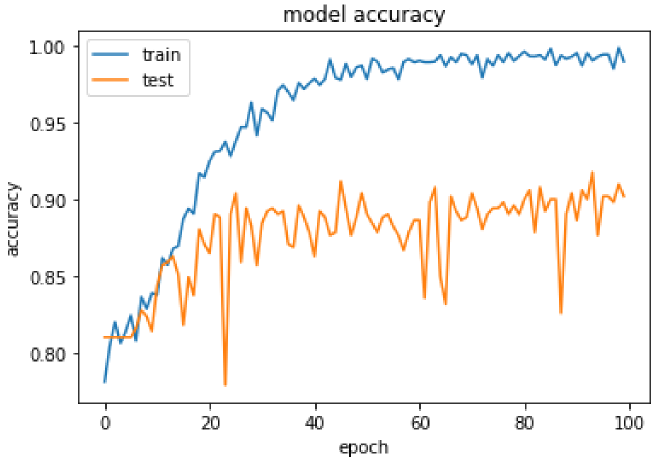


Results training and validation set



2992 images

Convolution neural network model



Model development
Accuracy training: 0.959
Accuracy validation: 0.906

Results on independent test set

- Performance of 5 animals 344 pictures

		CNN model	
		False	True
Gold standard Aniek	False	278	16
	True	11	39

● Accuracy	0.922	(278+39)/344
● Precision	0.709	39/(16+39)
● Recall	0.778	39/(11+39)
● Specificity	0.962	278/(278+11)
● F1_score	0.743	(2*39)/((2*39)+11+16

Performance evaluation

- Balancing false positive and true translocations
  - Added 3 copied of true SV in training and validation set
  - Trainings set :1723 and 395 -> 1723 and 1580
- Performance of 5 animals 345 pictures

	Without copies (Previous results)	With copies
Accuracy	0.922	0.942
Precision	0.709	0.812
Recall	0.778	0.780
Specificity	0.962	0.963
F1_score	0.743	0.796

Conclusions

- Much is already possible with small dataset
- The results looks promising
    - the sensitivity (recall), still need to be improved

Where to from here

- Finetuning model
- More data
- Images of bigger region of DNA 25 → 50
- Colouring chromosomes to 1 colour
- More balanced data

Suggestion are welcome

Thank you for your attention