Patterns of DNA variation between the autosomes, the X chromosome and the Y chromosome in *Bos taurus* genome

Bartosz Czech[1*], Bernt Guldbrandtsen[2,3], Joanna Szyda[1,4]

[1] Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland

[2] Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark

[3] Department of Animal Sciences, University of Bonn, Bonn, Germany

[4] Institute of Animal Breeding, Balice, Poland

[*] bartosz.czech@upwr.edu.pl http://theta.edu.pl

CONCLUSIONS

- fewer extreme variants are consistent with purging due to the homozygous state in males
- accumulation of nonsynonymous mutations on the BTY could be associated with loss of recombination
- variants in transcription regions on BTX have less severe consequences as compared to BTY and autosomes

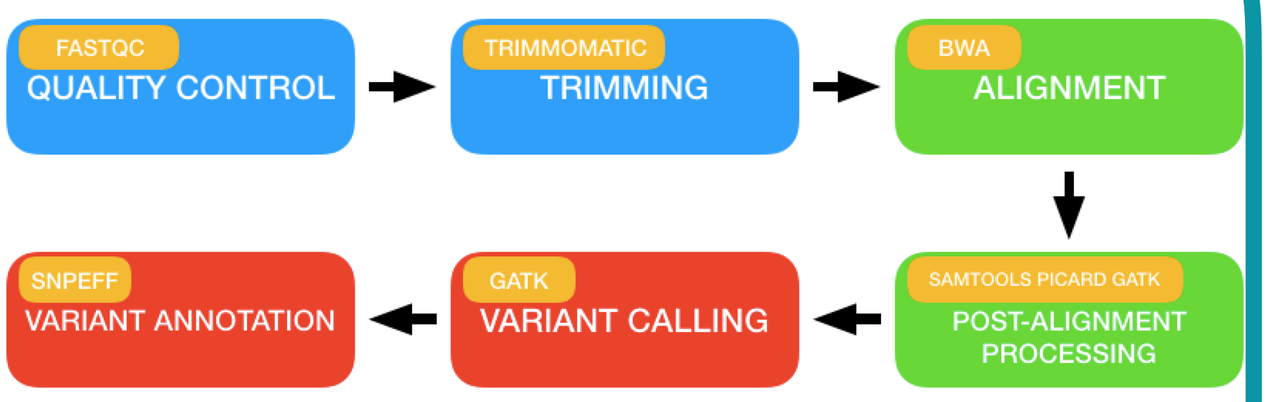
MATERIAL

- 217 individuals of 7 Danish cattle breeds
- WGS – Illumina HiSeq 2000
- assembly: ARS-UCD1.2_Btau5.0.1Y
- Btau_5.0.1 and ARS-UCD1.2 GFFs

RESULTS

- 23,655,295 SNPs / 3,758,781 InDels
- numbers of SNPs and InDels not uniformly distributed across 100kb non-overlapping windows ($P < 0.001$)
- Ka/Ks ratio: BTA = 0.79 BTX = 0.62 BTY = 2.00

METHODS



Statistical analysis:
- variant density on each chromosome
- InDel length • Ka/Ks ratio • nucleotide divergence
- Tajima's D • SIFT score

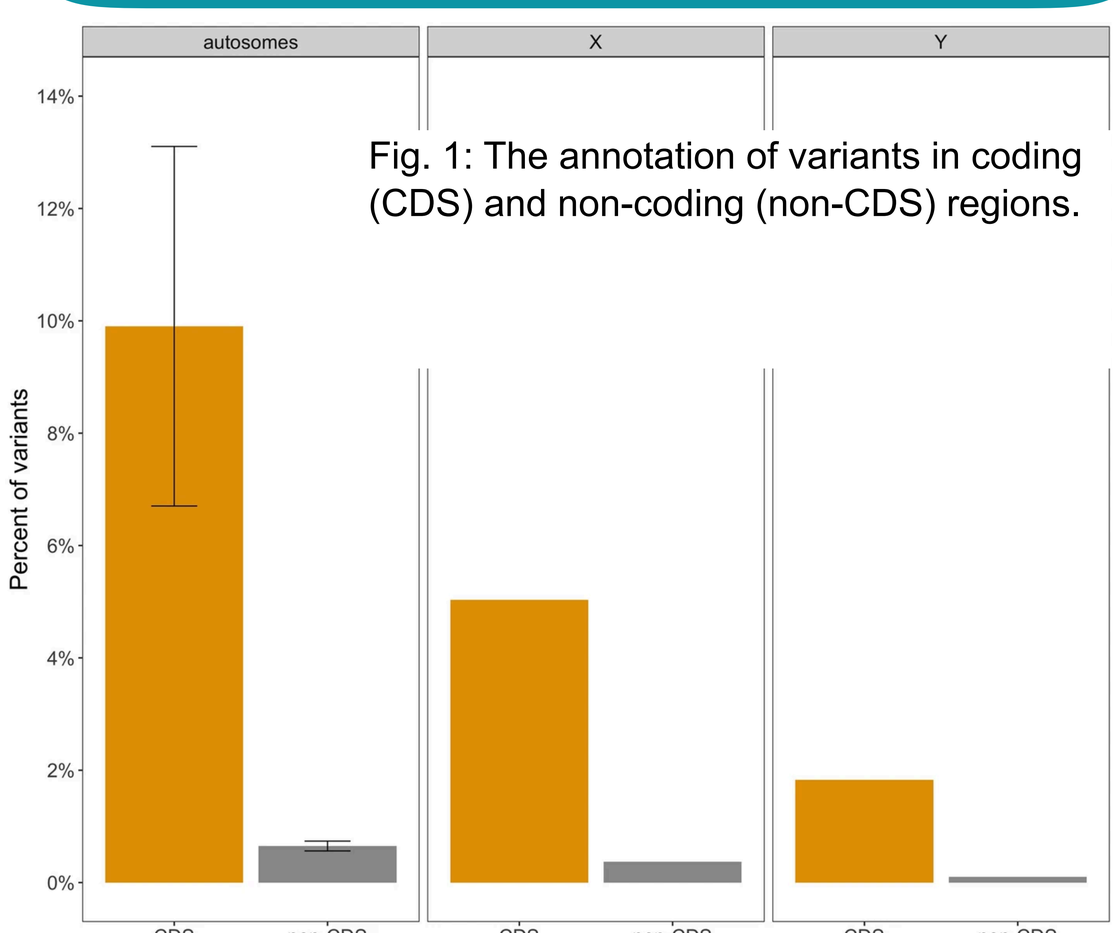


Fig. 1: The annotation of variants in coding (CDS) and non-coding (non-CDS) regions.

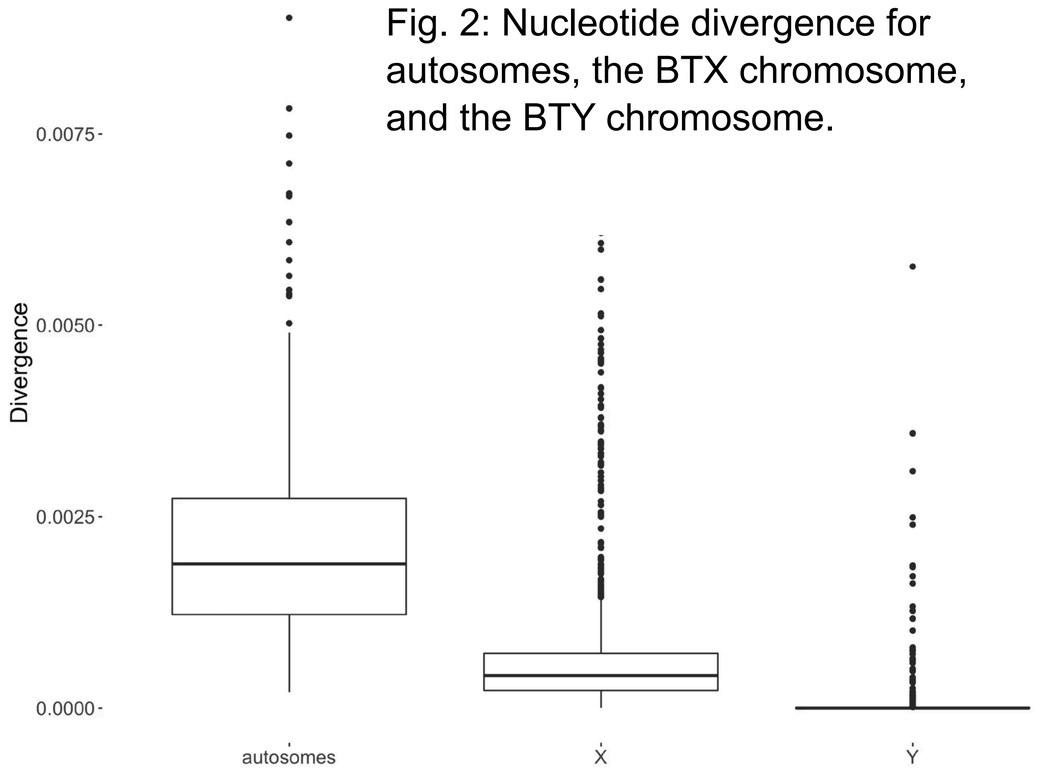


Fig. 2: Nucleotide divergence for autosomes, the BTX chromosome, and the BTY chromosome.

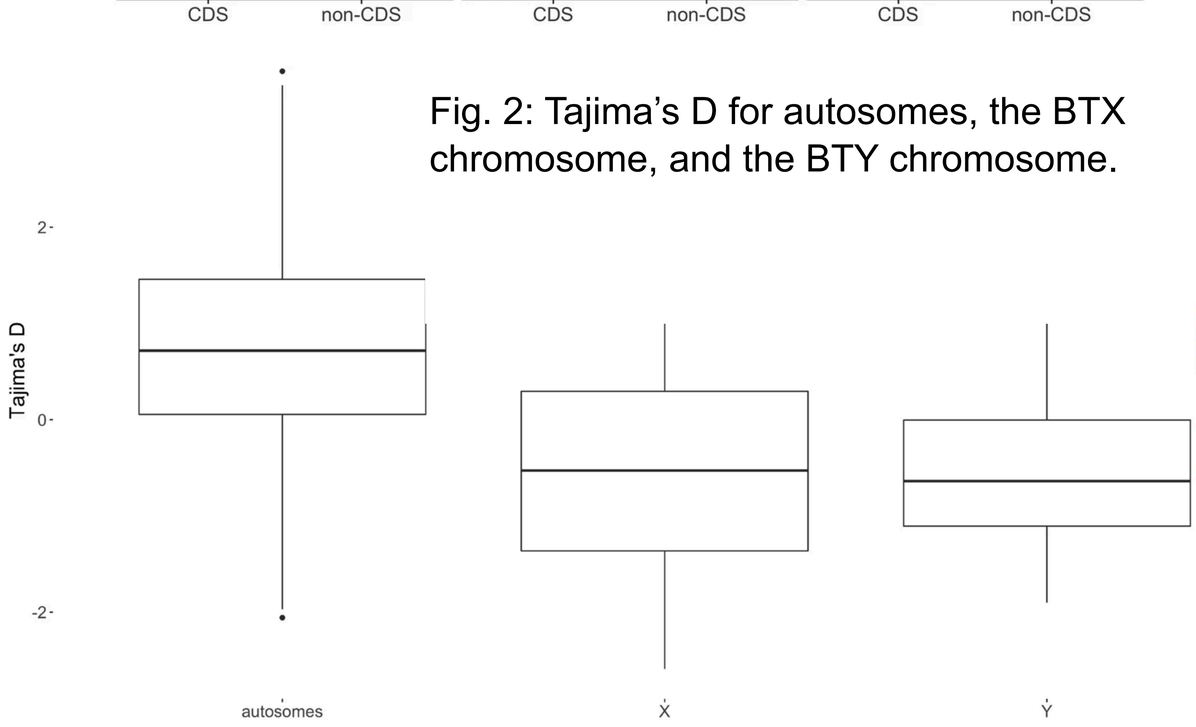


Fig. 2: Tajima's D for autosomes, the BTX chromosome, and the BTY chromosome.

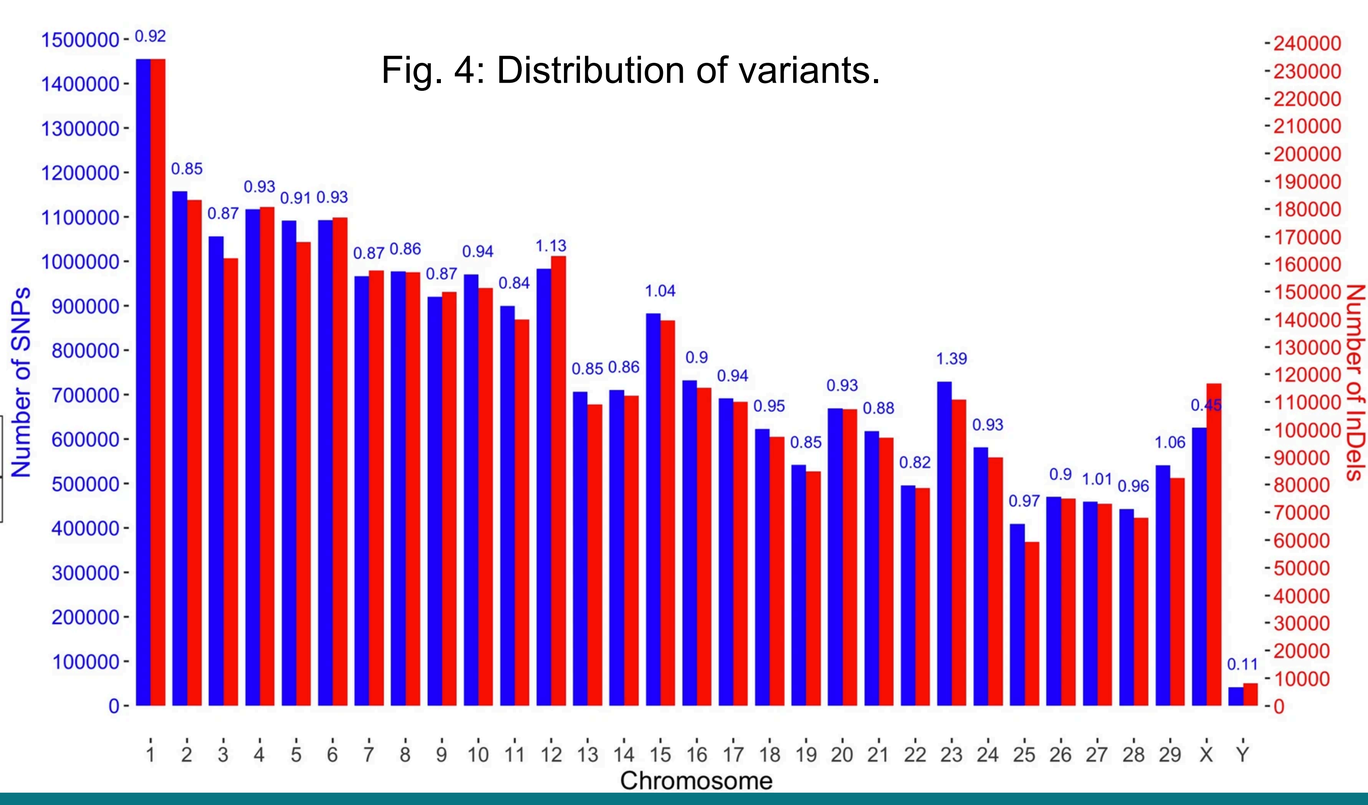


Fig. 4: Distribution of variants.

Mulitple sequence alignment analysis
master thesis

author: Paulina Dziadkiewicz (MiNI PW), advisor: dr hab. Norbert Dojer (MIMUW)
pedziadkiewicz@gmail.com, dojer@mimuw.edu.pl

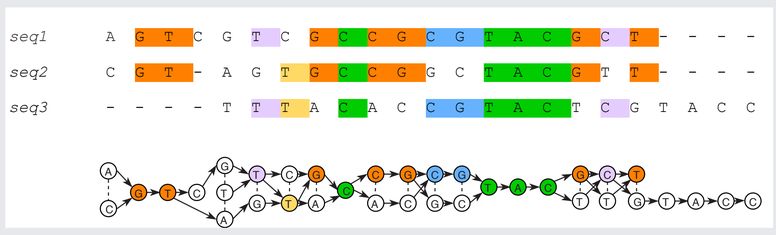
Introduction

Constant growth of genomic data leads to arising of a new research field called **pangenomics**. It is focused on delivering methods for joint multiple sequences processing. In this work a tool called *PangTree* is introduced.
The purpose of this tool is to extend currently used methods – **multiple sequence alignment**, **consensus search**, **multialignment graph representation** into new concept called **Affinity tree**. It is designed to be used as a taxonomic study or a reference genome for aligned sequences.

Multialignment as a graph

Graph representation of multiple alignment is based on partial order alignment graph.[1] The transformation is executed as follows:

1. Process multialignment column by column;
2. Merge identical nucleotides into single nodes;
3. Add directed edges between subsequent nodes and undirected for aligned nodes.



The representation is concise and intuitive. It is suitable to represent both short-length mutations and longer rearrangements, e.g. inversions or duplications.

Consensus idea

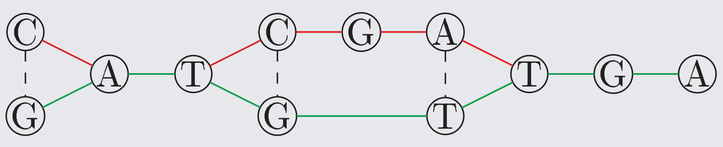
Typically, a consensus is determined by voting procedure on multialignment columns:

```
CATCGATGA
GATG-TTGA
CATG-TTG-

        ↓

CATG-TTGA
```

However, for multialignment given as a graph, Lee[1] proposed to find consensuses as minimum set of paths which describe all sequences.



Using Lee's approach we can build a graph model of multialignment and find a flat division of its component sequences into subgroups. Each of them has a consensus sequence assigned.

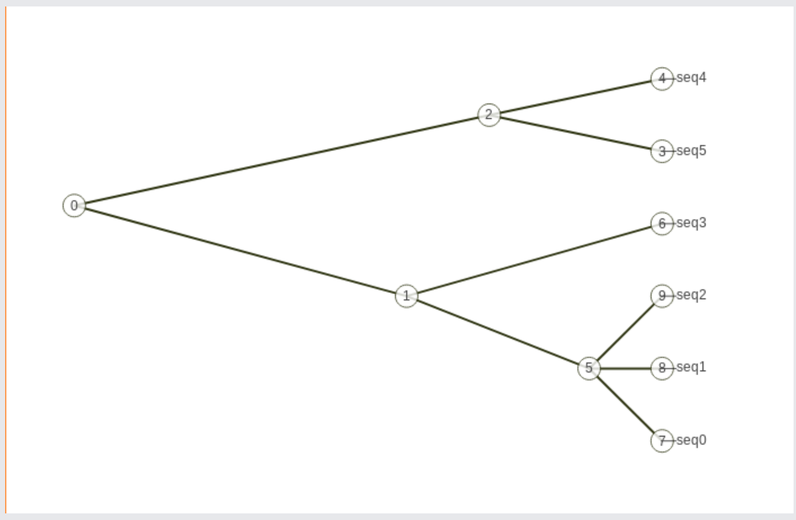
References

[1] Lee C. *Generating consensus sequences from partial order multiple sequence alignment graphs*, Bioinformatics. (2003) 22;19(8):999-1008.

[2] Dziadkiewicz, P., Dojer, *N. Getting insight into the pan-genome structure with PangTree.* BMC Genomics 21, 274 (2020).

Affinity tree

The introduced data structure is called Affinity tree. It serves as an extension of Lee's methods into hierarchical division of aligned sequences joint with consensus paths generation.
The root node has all input sequences assigned. Each non-leaf node has at least two children nodes that form a partition of the sequences assigned to their parent into more homogeneous subsets.



Each node has the following attributes assigned:

- a subset of input sequences,
- a linear consensus sequence being their common representation,
- a *minComp* (minimum compatibility) - value which reflects this node's homogeneity level.

Figure 1: An example of a Affinity tree

Affinity tree can be used as a reference genomes source, an evolution model or an assessment of heterogeneity for given dataset.

Simulated dataset

In order to evaluate the proposed solution a simulated multialignment was prepared using Evolver and evolverSimControl software. This alignment was based on a phylogenetic tree presented in Figure 2. It can be easily compared with the obtained Affinity tree which is shown in Figure 3.

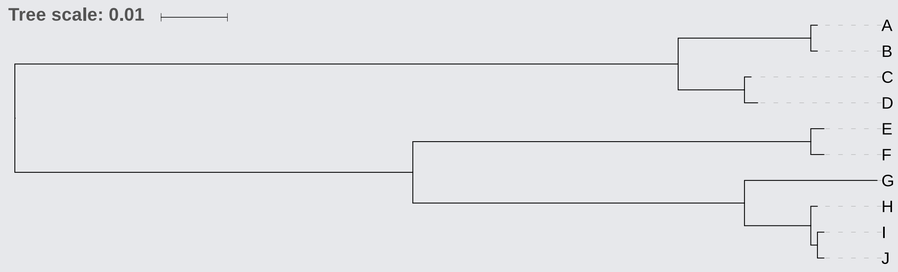


Figure 2: Phylogenetic tree for simulated data

The trees have similar forms which means, that the evolution pattern was correctly discovered by pangtree. However, the result includes not only the tree but also a consensus sequence assigned to each node. This is the main advantage of the Affinity tree over a phylogenetic tree.

For further simulations please follow the article[2].

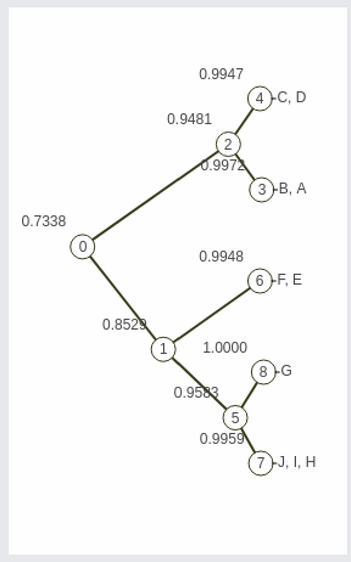


Figure 3: Affinity tree for simulated data

Ebola virus dataset

The proposed approach was also applied to Ebola virus alignment. The multialignment file was built using 160 genomes and is available in UCSC Ebola Portal together with associated studies.

The relationships between aligned sequences were correctly discovered.

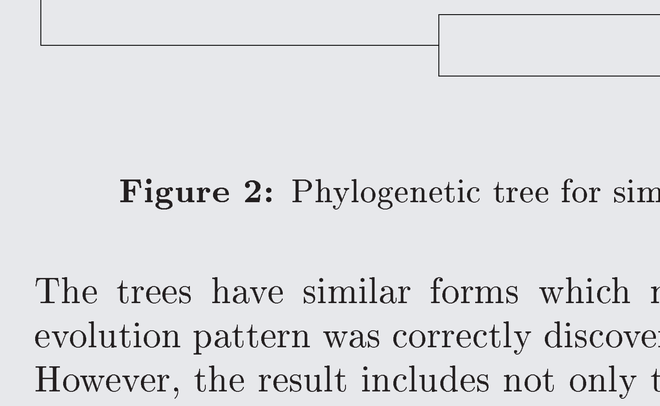


Figure 4: Local compatibilities between consensus sequences of ebolavirus species

Plots on Figure 4 show compatibilities with the consensus sequence of the species from the caption. Dark background indicates coding sequences, respective genes are listed below.

Figure 5: Ebola – Affinity tree

Cerebral Microbleeds detection on MR images with hybrid neural network

Aleksandra Suwalska[a], Yingzhe Wang[b], Ziyu Yuan[c], Yanfeng Jiang[c,d], Jinhua Chen[e], Mei Cui[b], Xingdong Chen[c,d], Chen Suo[c,f], Joanna Polanska[a]

[a] Silesian University of Technology, Department of Data Science and Engineering, 44-100 Gliwice, Poland
[b] Department of Neurology, Huashan Hospital, Fudan University, Shanghai, People's Republic of China
[c] Fudan University Taizhou Institute of Health Sciences, Taizhou, People's Republic of China
[d] State Key Laboratory of Genetic Engineering and Collaborative Innovation Centre for Genetic and Development, School of Life Sciences, Fudan University, Shanghai, People's Republic of China
[e] Taizhou People's Hospital, Taizhou, People's Republic of China
[f] Department of Epidemiology & Ministry of Education Key Laboratory of Public Health Safety, School of Public Health, Fudan University, Shanghai, People's Republic of China

Objective

The aim was to develop a novel tool for the automated detection of cerebral microbleeds (CMBs) based on magnetic resonance (MR) images. The system is expected to increase the sensitivity of CMB detection and to improve the accuracy of the diagnosis of the disease.

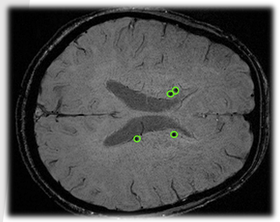


Figure 1: Example of brain image slice with CMB marked by an expert.

Introduction

Cerebral microbleeds are caused by structural abnormalities of the brain's small vessels. CMBs are linked with many neurological diseases; they can even lead to cognitive impairment, disability or death. They are visible on Susceptibility Weighted Imaging (SWI) sequences as round or elliptical areas with lower signal intensity and diameter up to 10 mm. Their manual detection is prone to errors and time-consuming.

Materials

In the study, MRI images from Taizhou People's Hospital were collected for a group of 304 patients and were used to train and test the system (Dataset 1). MR images from another 70 patients (Dataset 2) were used as an external independent validation. The process scheme is presented in Fig.2.

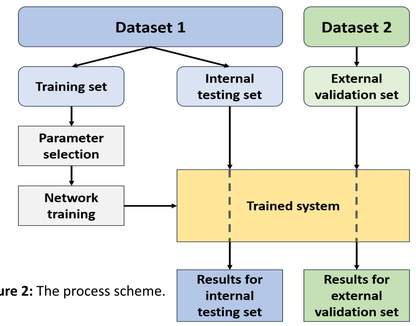


Figure 2: The process scheme.

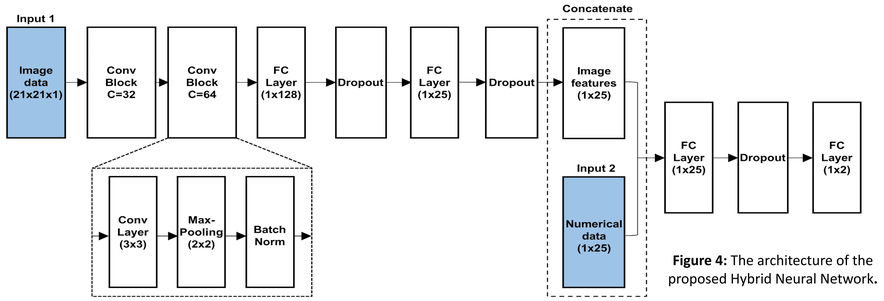


Figure 4: The architecture of the proposed Hybrid Neural Network.

Methods

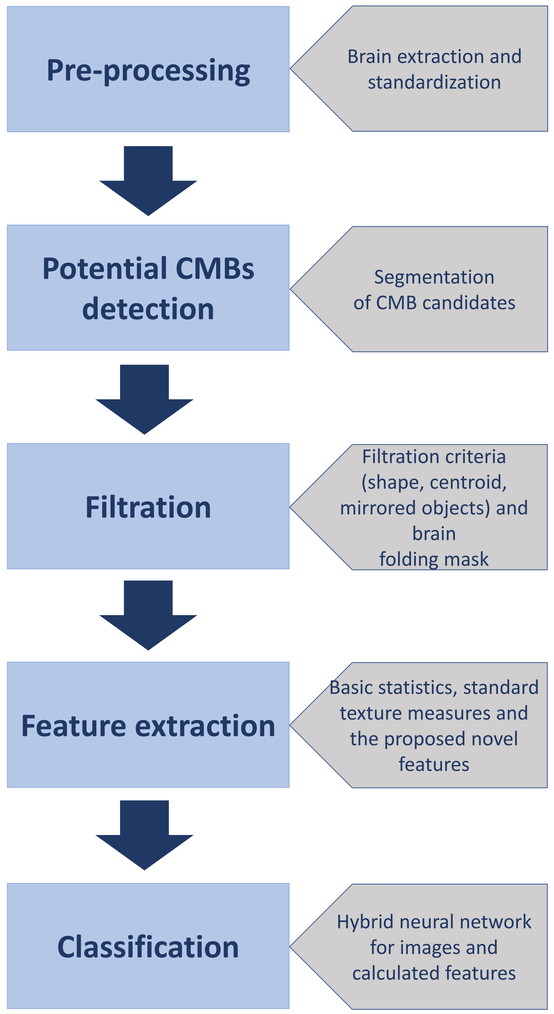


Figure 3: The pipeline of the CMB detection algorithm.

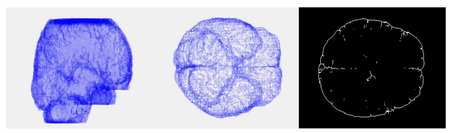


Figure 5: Brain folding mask in 3D and 2D.

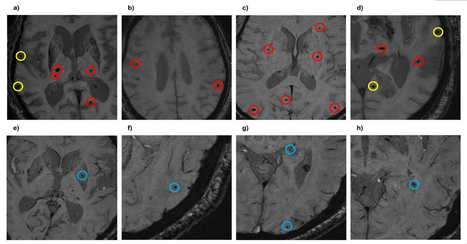


Figure 6: Exemplary positive samples. Red - CMBs identified correctly; yellow - CMBs lost by the system; blue - exemplary false positives.

Results

Dataset 1: The network reached a weighted accuracy of 94.48% with a sensitivity of 90.00% and specificity of 98.95%. The number of objects incorrectly classified as CMBs was 32 which gives an average of 0.54 false positives (FP) per patient.
Dataset 2: The system was able to detect 108 from 118 CMBs which resulted in the sensitivity of 91.5%. The number of false positives was 117 which gives 1.92 FPs per patient and the specificity of 95.2%.

		Training set			Test set					
Author	Modality	Patients without CMBs	Patients with CMBs	No. of CMBs	Patients without CMBs	Patients with CMBs	No. of CMBs	Sensitivity	Specificity	FPs/patient
Barnes et al. (2011)	SWI	-	6	120	-	6	6	81.70%	95.90%	107.50
Bian et al. (2013)	mIP SWI	-	5	116	-	10	304	86.50%	-	44.90
Chen, Yu et al. (2015)	SWI	-	15	62	-	5	55	89.13%	-	6.40
Van den Heuvel et al. (2016)	SWI+T1	18	23	491	-	10	136	89.00%	-	25.90
Dou, Chen et al. (2016)	SWI	-	270	270	-	50	117	93.16%	-	2.74
Ateeq et al. (2018)	SWI	-	14	104	-	6	63	93.70%	-	56.00
Chen et al. (2018)	SWI+echo scans	-	61	2458	-	12	377	94.70%	-	11.60
Liu et al. (2019)	SWI+phase	-	179	1473	10	31	168	95.80%	-	1.60
Suwalska, Wang et al. (2020) - Dataset 1	SWI	213	30	134	52	9	10	90.00%	98.95%	0.54
Suwalska, Wang et al. (2020) - Dataset 2	SWI	-	-	-	40	21	118	91.50%	95.20%	1.92

Table 1: Comparison with existing solutions (not all details were always available). Our results are marked red.

Conclusions

The use of both SWI images and numeric features allowed for the CMB's identification with high sensitivity and specificity without the need for additional imaging or complex models. On both test data, the developed system outperforms existing methods in terms of the number of false positives (FP) per patient. Our research confirms the usefulness of deep learning solutions to the problem of CMB detection based only on single MRI modality.

This work was partially financed by SUT BK-277/RAU4/2020. Calculations were carried out using infrastructure of GeCONiI (POIG.02.03.01-24-099/13).