978-1-7281-4452-8/20/$31.00 ©2020 IEEE

High Performance Computing for Detecting Complex Diseases using Deep Learning

Sahar I. Ghanem
Dept. of Computer and Systems Engineering,
Faculty of Engineering, Alexandria University,
Alexandria, Egypt
eng-sahar.ghanem@alexu.edu.eg

Ahmed A. Ghoneim
Dept. of Computer Engineering, Faculty of Engineering,
Arab Academy for Science, Technology and Maritime Transport,
Alexandria, Egypt
ghoneim@student.aast.edu

Nagia M. Ghanem
Dept. of Computer and Systems Engineering,
Faculty of Engineering, Alexandria University,
Alexandria, Egypt
nagia.ghanem@alexu.edu.eg

Mohamed A. Ismail
Dept. of Computer and Systems Engineering,
Faculty of Engineering, Alexandria University,
Alexandria, Egypt
maismail@alexu.edu.eg

***Abstract*—The study of the Genome-wide association study (GWAS) and the complex diseases is of high importance nowadays. The epistasis describes the analysis of the single nucleotide polymorphisms (SNPs) interactions and their effects on the complex diseases. However, enormous number of SNPs interactions should be tested against the disease that is highly computational expensive. In this paper, High Performance Computing (HPC) is being applied on a supercomputer to reduce the processing time. Parallel Deep Learning (PDL) is applied and tested using different datasets. Simulated datasets of 12 different models and the real WTCCC Rheumatoid arthritis (RA) dataset are being tested. Results show the high accuracy, specificity and true positive rate values. Moreover, they show low values of the false discovery rate and the robustness of power through the different simulated models. When tested on the real RA dataset, our model shows the ability to detect the 2-way interaction SNPs with their promising related genes with high accuracy due to the parallel deep learning architecture.**

***Index Terms*—GWAS, complex diseases, single nucleotide polymorphisms, high performance computing, deep learning**

I. Introduction

GWAS successfully reveals the analysis of the genotype-phenotype relationship to study different diseases with its binary or quantitative phenotype trait [1]. It explores the study of SNPs. SNPs are the genetic markers used to analysis their association with the diseases [2]. In fact, the study of the single linear SNP effect gives insufficient information in most cases. Hence, the SNPs pairwise interactions, higher order interactions (epistasis) and their associations with the disease, are of more concern recently.

Some approaches rely on choosing SNPs due to their allelic pair single marginal effect on the phenotypic trait, as a first stage. Then, they test the pairwise effect on the disease among these selected SNPs of high individual association effect, as a second stage.

Unfortunately, in many cases SNPs showing no significant single effect on the disease can have a remarkable interaction effect with other SNPs towards the disease trait, that evolves the genome-wide interaction studies (GWIS). It is more concerned with the SNPs associated effect on the phenotype trait than their single impact. Some of the well-known models expressing the combined effect are additive, dominant, recessive, multiplicative, threshold and XOR effects.

There are several multi-loci algorithms for the epistatic detection. BEAM [3] is a Stochastic search methods relies on Markov chain MonteCarlo model-based method. Stochastic algorithms depends on searching and usually reach an optimal solution that can easily stuck to a local minimum. EDCF [4], SNPRuler [5] and BOOST [6] are examples of of exhaustive search techniques that investigate all the possible SNPs loci combinations. BOOST depends on the binary representation to enhance the time complexity issue but it is limited to 2 loci interaction. The GPU parallel implementation presented in PBOOST [7] is one of the BOOST extensions. MDR [8] is another example of exhaustive search techniques that tends to reduce the high dimension space. MDR is non-parametric, non model based technique. Ant Colony optimization algorithm is presented in AntEpiSeeker [9] and its parallel implementation [10] using MapReduce Framework. Cuckoo search epistasis (CSE) [11] depends on dividing SNP sites into groups due to per-knowlwdge on dataset and the disase model is more reliable compared to BEAM and multifactor dimensionality reduction (MDR). In addition, Hi-Seeker [12] a parametric algorithm to detect higher order SNPs interactions, is not based on the highly significant SNPs pairs and insensitive to the marginal effect. But, it is affected by linkage disequilibrium and types of noise such as missing genotypic data. Heuristic search algorithms are another category. They are approximate algorithms used to reduce the space dimension. SNPHarvest [13] is one of its examples that reduces the SNPs due to statistical tests. Dynamic clustering and cloud computing (DCHE) [14] is constructed to speed-up the SNPs interaction of pairwise and higher order ones, using step-wise approach that is much faster than exhaustive algorithms. It shows better

978-1-7281-4452-8/20/$31.00 ©2020 IEEE

performance for complex diseases of both without or with low marginal effect. Both heuristic and Stochastic methods are better to be used for diseases with and without marginal effects. Lastly, machine learning approaches are regarded as classification problems that can select the most discriminative features affecting the output trait. DL [15] is an example of the machine learning algorithm that uses feed-forward DL algorithm to detect the SNP pairs ranked due to their variable of importance.

The detection of complex disease is regarded as high computational cost problem due to the excessive processing and the high dimensional data, especially when moving towards higher-order levels of interaction. Thus, moving towards accelerating approached as cloud computing in DCHE, using CUDA implementation for the GPU in GBOOST [16] and the hybrid FPGA-GPU architecture in [17] used for third-order interaction detection. A heterogeneous architecture with both GPUs and Xeon Phi coprocessors is introduced in [18] used for the epistatic detection using regression models. A parallel FaST-LMM [19] uses supercomputer with nodes offering a number of GPUs, implemented using MPI standards. Other studies are more concerned on parallel approaches for quantitative traits like epiSNP parallel version [20] and its extension implementation on TACCs Stampede Supercomputer using the Intel Xeon Phi coprocessors (MICs) [21] for serial optimization and load balance enhancement.

Although, DL has been applied on small scale of synthesis and real data and tested against different types of noise as in [22] showing high reliability, it is still not applied on large-scale real data that requires extensive processing. In this paper, parallel deep learning (PDL) approach applying Map/Reduce framework to enable multi-node processing is implemented. The PDL is being applied by the Bibliotheca Alexandrina [23] High Performance Computer BA-HPC C2 cluster architecture on different synthesis data models and being tested on RA WTCCC data.

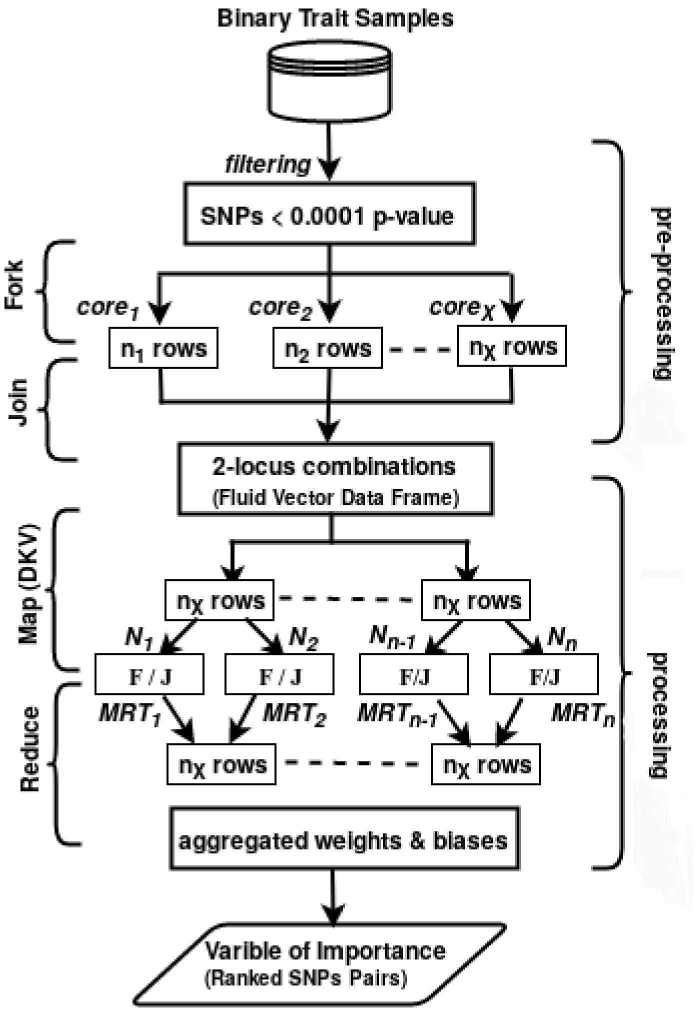

Fig. 1. The lower architecture of the pre-processing and the Parallel Deep Learning processing Flow Diagram. Pre-processing phase is composed of filtering stage and applying Fork/Join (F/J) multi-threading framework. The processing phase is applying Map/Reduce framework as a lower layer dividing tasks into map reduce tasks (MRT1, MRT2,..MRTn) to process the assigned data rows through nodes([N1,N2,..Nn). Each node applies F/J framework.

II. Methods

A. Parallel Deep Learning

In this paper 2 phases are presented as in fig. 1. Phase 1, the pre-processing stage, that filters the SNPs pairs due to p-value threshold value of 10^{-13} in a parallel Fork/Join (F/J) multi-threading framework. The output is the 2-locus combinations as an intermediate stage. F/J divides the data samples and shares the filtered SNPs pairs headers, in a way to get use of the maximum threads available to fasten the performance.

In Phase 2, the BA-HPC C2 is being used to apply the parallel deep learning stage (PDL). The parallel and distributed cluster computing environments is of multi-nodes with shared memory model is being handled by the H2O platform [24]. H2O is an open source, salable distributed machine learning with fine-grain parallelism. In our study, the H2o cluster server is being accessed by R interface that acts as a client front-end connecting to the H2o cluster server.

The data output from the intermediate stage is being input as Fluid Vector Data Frame that has the flexibility to be edited through adding, deleting or changing columns. Map/Reduce framework is applied in a way that each node is assigned a subset of the deep learning architecture parameters (group of weights and biases) as well as a block of data. Data storage is applied by a distributed Key/Value (DKV) used by the non-blocking HashMap. Each node N processes its assigned data using F/J framework. At the end, the nodes results (net weights and biases) are aggregated at the end to form the overall deep learning parameter values and the output results.

In this paper, parallel feed-forward deep learning is being applied. The model consists of an input layer, a number of hidden layers and an output one. Data is splitted in folds for the training and testing the model. In the training phase, backpropagation algorithm is used to calculate and update the network weights, until reaching steady state. Activation functions are being applied through the model layers and other parameters are set to avoid overfitting as dropout values, L1

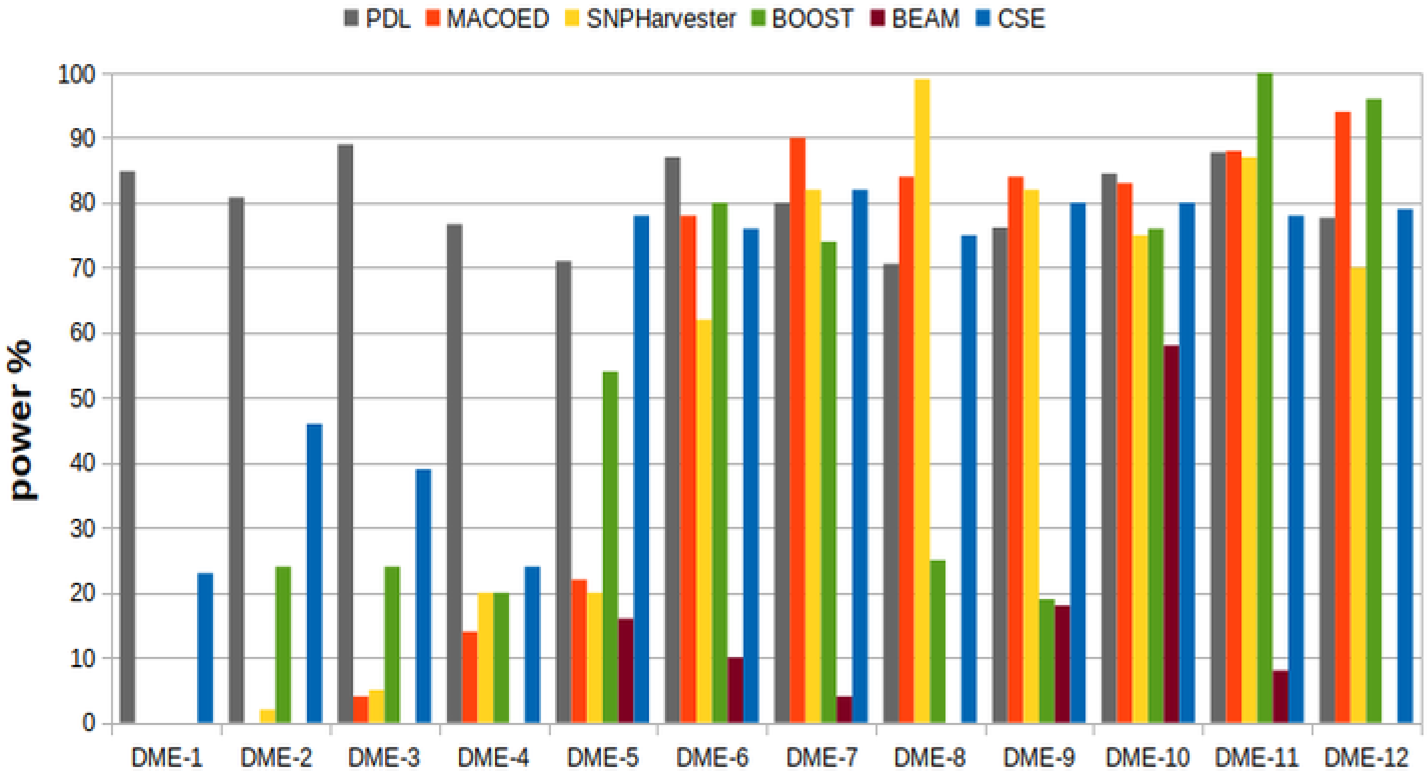

Fig. 2. Power percentage of the 12 DME models. Each model 100 datasets. Each dataset of 100 SNPs and 1600 samples (800 cases and 800 controls).

and L2 regularization.

B. Data collection

1) Simulated Data: Twelve 2-locus disease models with main effects (DME) are being investigated in our study [25]. The first four models expressing the multiplicative model that shows an increase of the 2-alleilic pair frequency multiplicatively with the disease presence under different minor allele frequencies (MAF) of 0.05, 0.1, 0.2 and 0.5. The other four models representing the threshold model under the same MAF values are (0.05, 0.1, 0.2 and 0.5). The last four models are concrete ones having low marginal effect and strong interaction effect. The genetic heritability is of 0.005 for the multiplicative models, but 0.02 for both threshold and concrete models. The simulated datasets generated is of 1600 samples (800 cases and 800 controls) and 100 SNPs with 2 functional SNPs representing the pair-wise interaction model. It is generated using the GAMETES tool [26], where random samples are automatically generated due to the defined penetrance functions and MAF for the 2 allelic pairs functional SNPs that determine the corresponding heritability. The 12 models are compared against other algorithms; MACOED [25], SNPHarvester [13], BOOST [6], BEAM [3] and CSE [11].

PDL is applied on each model dataset, by 5 hidden layers, each 100 neuron. PDL parameters are tuned to optimize the accuracy results [27]. Rho is 0.98, epochs is 10 with a value of 1.0e-4 for both L1 and L2 regularization and tanh activation function. A dropout effect is applied on all hidden layers by a value of 0.1 to avoid overfitting problem. Cross validation is applied of 10 folds, for training, testing and validating the data.

2) Real Data: In our experiment the real large scale dataset has been tested using the Rheumatoid arthritis (RA) disease from Wellcome Trust Case Control Consortium (WTCCC) [28] using the BA HPC supercomputer. RA disease is a chronic inflammatory complex disease that may lead to severe disability. RA dataset is composed of nearly 500K SNPs and 5,000 samples. Samples are divided into 2,000 cases and 3,000 controls (1,500 from the 1958 British Birth Cohort and 1,500 UK Blood Services). Chromosomes are to be scanned separately,each of average 30K SNPs.

PDL is applied in parallel approach on the HPC on RA dataset, by 5 hidden layers, each 200 neuron and 64 epochs. The rest of the parameters are of same setting for simulated data models.

III. Results and Discussion

A. Simulated Data

The experimental results show that the PDL is not highly affected by the genetic heritability among the three different models compared to other methods. PDL has the highest power concerning the multiplicative models (DME-1,DME-2, DME-3 and DME-4), shown in fig.2. However, MACOED,

TABLE I
ACCURACY(ACC), SPECIFICITY(SPC), TRUE POSITIVE RATE(TPR) AND FALSE DISCOVERY RATE(FDR) PERCENTAGES OF THE 12 DME MODELS OF THE SIMULATED DATA. EACH MODEL 100 DATASETS. EACH DATASET OF 100 SNPS AND 1600 SAMPLES (800 CASES AND 800 CONTROLS).

Model	**PDL**				**MACOED**				**CSE**				**SNPHarvester**				**BEAM**				**BOOST**			
	TPR	**Spc**	**Acc**	**FDR**	**TPR**	**Spc**	**Acc**	**FDR**	**TPR**	**Spc**	**Acc**	**FDR**	**TPR**	**Spc**	**Acc**	**FDR**	**TPR**	**Spc**	**Acc**	**FDR**	**TPR**	**Spc**	**Acc**	**FDR**
DME-1	100	100	94	11	15	99	70	8	54	54	54	46	0	100	50	NaN[a]	0	79	44	100	0	100	50	NaN
DME-2	100	99	95	7	68	99	87	3	46	46	46	54	0	100	50	NaN	3	68	41	94	2	100	51	0
DME-3	100	100	93	3	93	99	97	2	38	38	38	62	0	100	50	NaN	10	56	38	88	0	100	50	NaN
DME-4	100	92	85	5	98	79	87	24	25	25	25	75	1	99	50	50	46	26	33	77	0	100	50	NaN
DME-5	100	100	83	0	96	99	98	1	78	78	78	22	0	100	50	NaN	64	42	54	44	25	100	63	0
DME-6	100	99	97	2	98	87	92	13	74	74	74	26	34	97	66	8	79	28	57	41	61	99	80	2
DME-7	100	100	94	0	98	82	89	18	83	83	83	17	73	41	51	64	98	6	72	27	49	99	74	2
DME-8	100	100	94	0	100	64	78	36	75	75	75	25	100	2	12	90	97	15	83	15	4	100	52	0
DME-9	100	100	85	0	99	74	85	26	80	80	80	20	77	29	40	74	100	0	83	17	0	100	50	NaN
DME-10	100	100	94	0	100	83	91	17	81	81	81	19	94	47	63	53	100	0	83	17	67	100	84	0
DME-11	100	100	98	0	99	97	98	4	78	78	78	22	96	100	98	0	27	45	38	76	100	100	100	0
DME-12	100	100	91	1	97	86	91	14	79	79	79	21	98	55	71	44	98	2	58	41	98	99	99	1

[a] NaN can not be calculated when both True Positive (TP) and False Negative (FN) are equal to 0.

TABLE II
2-WAY SNPS INTERACTIONS OF RA DISEASE.

SNP-SNP Interaction	***Genes***	**Importance**	**Genomic Position**
rs17165379-rs41501546	ZDHHC14-ZDHHC14	1	chr6:(157506828-157464909)
rs3129249-rs3130233	LOC105375021-HLA-DPB2	0.911	chr6:(33141779-33127399)
rs35225149-rs41321645	NR-NR[a]	0.910	chr6:(115004900-86175699)
rs241403-rs241404	LOC100294145-LOC100294145	0.864	chr6:(32899215-32898220)
rs9275418-rs9469220	NR-NR	0.861	chr6:(32702467-32690533)
rs41505850-rs41391157	LOC105377964-HECA	0.855	chr6:(117261297-139140028)
rs35225149-rs1413342	NR-LOC340184	0.851	chr6:(115004900-11696274)
rs241404-rs2857136	LOC100294145-NR	0.841	chr6:(32898220-32807909)
rs916571-rs1610584	HCG17-NR	0.830	chr6:(30234178-29707838)
rs35225149-rs10498719	NR-NR	0.816	chr6:(115004900-23647806)
rs29233-rs17165379	NR-ZDHHC14	0.805	chr6:(29643452-157506828)
rs17165379-rs41327149	ZDHHC14-NR	0.789	chr6:(157506828-137568017)
rs41321645-rs41509944	NR-NR	0.785	chr6:(86175699-133166156)
rs9273363-rs9469220	HLA-DQB1-AS1-NR	0.780	chr6:(32658495-32690533)
rs9469220-rs9275418	NR-NR	0.763	chr6:(32690533-32702467)
rs1413342-rs41438346	LOC340184-NR	0.748	chr6:(11696274-122093340)
rs241404-rs9276815	LOC100294145-PSMB9	0.718	chr6:(32898220-32857730)
rs17165379-rs41345853	ZDHHC14-RPS6KA2	0.712	chr6:(157506828-166751831)
rs241404-rs1800454	LOC100294145-TAP2	0.712	chr6:(32898220-32832635)
rs17165379-rs41489047	ZDHHC14-ADGRB3	0.710	chr6:(157506828-68774553)

[a] NR refers to unknown gene.

SNPHarvester, CSE shows higher power values reaching to a maximum value of 99%, concerning models DME-7 and DME-8 related to the threshold model. Also, BOOST shows optimum power values of 100% for DME-11 and DME-12, concrete models. PDL, MACOED, SNPHarvester and CSE have nearly similar power values regarding the last 4 concrete models.

Table I shows the values of true positive rate(TPR), specificity(Spc), accuracy(Acc) and false discovery rate(FDR) metrics. PDL has stable higher accuracy [83-98]% for the 12 models compared to SNPHarvester, BEAM and CSE. Similarly, PDL outperforms BOOST in the first 10 models by a maximum difference of 44%. But, for DME-11 and DME-12 BOOST accuracy is slightly higher by only 8% maximum value. PDL shows approximate equal accuracy values compared to MACOED. Concerning the TPR the PDL shows optimum response of 100%. Similarly, the Spc outperforms other methods for all models, except for the DME-2 and DME-4 models. The BOOST have higer values of 100% for both DME-2 and DME-4 models compared to PDL. Lastly, The PDL FDR shows better results for both the threshold and concrete models. But, MACOED and BOOST have lower FDR regarding the multiplicative models.

$$t_{\mathrm{k}} = t_1/k \tag{1}$$

Equation (1) shows the linear scalability where t_1 is the serial time using a single core and t_{k} is the parallel time using k number of cores. Although, theoretically the number of cores is inversely proportional in a linear manner with time, practically that is not implied. The time taken is not only the time per each core, but also the time taken for their interactions. Fig. 3 and 4 show the effect of applying different number of cores on the time metric against the increase of the number of SNPs and samples respectively.

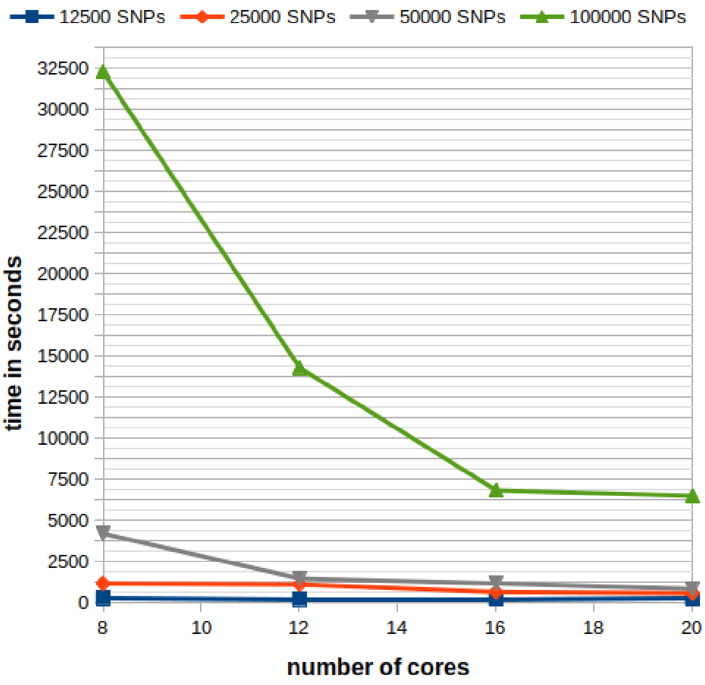

Fig. 3. Time taken of different number of SNPs and 5,000 samples across different number of cores.

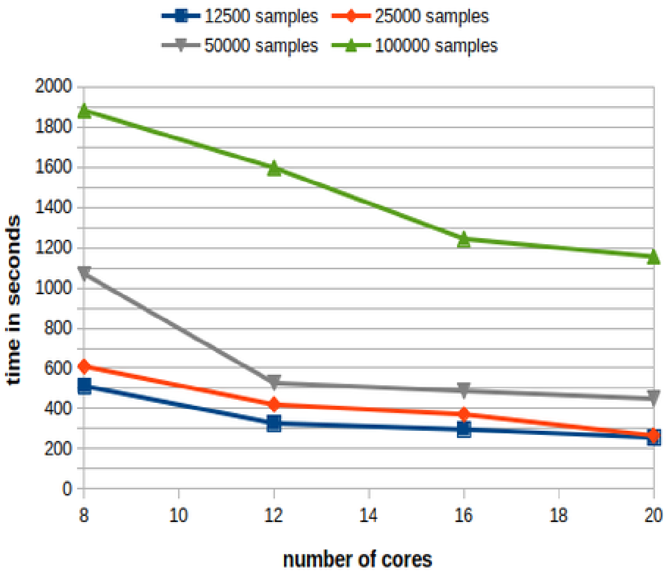

Fig. 4. Time taken of different number of samples and 10,000 SNPs across different number of cores.

B. Real Data

For each chromosome, SNPs combinations included in our investigation are nearly 25,000, after the pre-processing phase. The number of cores applied is 16 cores and 26.67 GB RAM in 859 seconds on the BA-HPC cluster. Table II shows the first 20 SNP-SNP interactions, with their variable of importance value, where NR reveals no specific related gene. By analyzing the top 100 2-ways interaction SNPs, it can be concluded that rs17165379 in (ZDHHC14) appears in 17 interactions, rs6457620 in (HLA-DRB1), rs6457617 in (MHC), that has been explored previously by AntEpiSeeker. SNPs rs2857136, rs2857129 were also shown in [10]study. rs2857154 in (HLA-DQB2) and other related genes (HLA-DPB2), (HLA-DRB1), (HLA-DQA1) and (HLA-DQA2) related to HLA class have been explored by many studies to be highly related to RA. Gene (TSBP1) , (C6orf10) and (BTNL2) are associated with RA and marked by DCHE. However, there are still other genes appearing have to be biologically interpreted.

IV. Conclusion

Parallel Deep Learning (PDL) on a supercomputer is being applied to investigate the SNP-SNP interactions. The results on the simulated data show the robustness of PDL among the 12 disease models datasets. But the average power is affected by the existence of model-based. PDL approach is also applied on WTCCC RA real dataset, showing remarkable interactions with some biological analysis of the contributed genes. Knowing the relations of other genes and RA disease is under investigation. Extensions for this work include studying higher-order epistasis interactions and applying more general diseases levels such as the quantitative traits.

Acknowledgment

The authors gratefully acknowledge the postive and constructive remarks given by anonymous referees which contributed positively to the final form of this paper. The authors would like to thank the access of Wellcome Trust Case Control Consortium data available at https://www.wtccc.org.uk/. We acknowledge with thanks the use of the High-Performance Computing (HPC) cluster,BA-HPC-C2” available at the Bibliotheca Alexandrina (BA). The help of Prof. Noha Yousri at Weill Cornell Medical Qatar is highly appreciated.

REFERENCES

[1] Upton, Alex, et al. "High-performance computing to detect epistasis in genome scale data sets." Briefings in bioinformatics 17.3 (2015): 368-379.

[2] MacArthur, Jacqueline, et al. "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)." Nucleic acids research 45.D1 (2016): D896-D901.

[3] Zhang, Yu, and Jun S. Liu. "Bayesian inference of epistatic interactions in case-control studies." Nature genetics 39.9 (2007): 1167.

[4] Xie, Minzhu, Jing Li, and Tao Jiang. "Detecting genome-wide epistases based on the clustering of relatively frequent items." Bioinformatics 28.1 (2011): 5-12.

[5] Wan, Xiang, et al. "Predictive rule inference for epistatic interaction detection in genome-wide association studies." Bioinformatics 26.1 (2009): 30-37.

[6] Wan, Xiang, et al. "BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies." The American Journal of Human Genetics 87.3 (2010): 325-340.

[7] Yang, G.; Jiang, W.; Yang, Q.; Yu, W. PBOOST: A GPU based tool for parallel permutation tests in genome-wide association studies. Bioinformatics 2015, 31, 14601462.

[8] Hahn, Lance W., Marylyn D. Ritchie, and Jason H. Moore. "Multifactor dimensionality reduction software for detecting genegene and geneenvironment interactions." Bioinformatics 19.3 (2003): 376-382.

[9] Wang, Yupeng, et al. "AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm." BMC research notes 3.1 (2010): 117.

[10] Zhou, Zhihui, Guixia Liu, and Lingtao Su. "A new approach to detect epistasis utilizing parallel implementation of ant colony optimization by MapReduce framework." International Journal of Computer Mathematics 93.3 (2016): 511-523.

[11] Aflakparast, M. et al. Cuckoo search epitasis: a new method for exploring significant genetic interactions. Heredity 112, 666674 (2014).

[12] Liu, Jie, et al. "Hiseeker: Detecting high-order snp interactions based on pairwise snp combinations." Genes 8.6 (2017): 153.

[13] Yang, C. et al. SNPHarvester: A Filtering-based Approach for Detecting Epistatic Interactions in Genome-wide Association Studies. Bioinformatics 25, 504511 (2009).

[14] Guo, Xuan, et al. "Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering." BMC bioinformatics 15.1 (2014): 102.

[15] Uppu, Suneetha, Aneesh Krishna, and Raj P. Gopalan. "A Deep Learning Approach to Detect SNP Interactions." JSW Vol. 11, No. 10, pp. 965-975, 2016.

[16] L.S.Yung, C. Yang, X. Wan et al., GBOOST: A GPU-Based Tool for Detecting Gene-Gene Interactions in Genome-Wide Case Control Studies, Bioinformatics, vol. 27, no. 9, pp. 13091310, 2011.

[17] Wienbrandt, Lars, et al. "Fast genome-wide third-order snp interaction tests with information gain on a low-cost heterogeneous parallel fpga-gpu computing architecture." Procedia computer science 108 (2017): 596-605.

[18] Gonzlez-Domnguez, Jorge, et al. "Parallel pairwise epistasis detection on heterogeneous computing architectures." IEEE Transactions on Parallel and Distributed Systems 27.8 (2015): 2329-2340.

[19] Martnez Prez, Hctor, et al. "FaST-LMM for Two-Way Epistasis Tests on High-Performance Clusters." (2018).

[20] Ma, Li, et al. "Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies." BMC bioinformatics 9.1 (2008): 315.

[21] Weeks, Nathan T., et al. "High-performance epistasis detection in quantitative trait GWAS." The International Journal of High Performance Computing Applications 32.3 (2018): 321-336.

[22] Ghanem, Sahar I., Nagia M. Ghanem, and Mohamed A. Ismail. "Noisy Epistasis Using Deep Learning." 2018 International Japan-Africa Conference on Electronics, Communications and Computations (JAC-ECC). IEEE, 2018.

[23] Bibliotheca alexandrina supercomputer project web page, ", available: http://www.bibalex.org/ISIS/Frontend/Projects/ProjectDetails.aspx?id=m8fC7jXMTFprEy98pIPBFw=="

[24] Aiello, Spencer, et al. "Machine Learning with R and H2O." (2018).

[25] Jing, Peng-Jie, and Hong-Bin Shen. "MACOED: a multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies." Bioinformatics 31.5 (2014): 634-641.

[26] Urbanowicz, Ryan J., et al. "GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures." BioData mining 5.1 (2012): 16.

[27] Uppu, Suneetha, and Aneesh Krishna. "Tuning hyperparameters for gene interaction models in genome-wide association studies." International Conference on Neural Information Processing. Springer, Cham, 2017.

[28] Wellcome Trust Case Control Consortium. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature 447.7145 (2007): 661.