Chapter 4

A Review on Machine Learning and Deep Learning Techniques Applied to Liquid Biopsy

Arets Paeglis, Boriss Strumfs, Dzeina Mezale and Ilze Fridrihsone

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.79404

Abstract

For more than a decade, machine learning (ML) and deep learning (DL) techniques have been a mainstay in the toolset for the analysis of large amounts of weakly correlated or high-dimensional data. As new technologies for detecting and measuring biochemical markers from bodily fluid samples (e.g., microfluidics and labs-on-a-chip) revolutionise the industry of diagnostics and precision medicine, the heterogeneity and complexity of the acquired data present a growing challenge to their interpretation and usage. In this chapter, we attempt to review the state of ML and DL fields as applied to the analysis of liquid biopsy data and summarise the available corpus of techniques and methodologies.

Keywords: machine learning, deep learning, data analysis, biomarker detection, automated discovery, literature review

1. Introduction

Biological and medical sciences are becoming increasingly data-rich and information-intensive. This tendency, along with the growing availability of such data, provides a better understanding of important questions regarding functions of organisms, causes of diseases, etc. However, both the inherently massive complexity of biological systems and the high dimensionality and noisiness of data thus acquired can make it remarkably difficult to correctly infer such mechanisms. Machine learning (ML) and deep learning (DL) techniques are quickly becoming highly useful tools for solving difficult problems in biology and medicine by providing mathematical apparatus for analysing vast amounts of information that would

otherwise be difficult to process and interpret. Additionally, these fields themselves provide new challenges for machine learning that can ultimately advance existing ML techniques and give rise to new ones.

The mutual history of machine learning and biological and medical disciplines is both long and complex. An early ML technique, the perceptron, was made in attempt to model the behaviour of biological neurons [1] and was used early on to define the start sites of translation initiation sequences in *E. coli* [2], and can be considered the starting point of the entire field of machine learning. In the last few decades, the power, flexibility, and accessibility of ML and DL techniques have grown considerably, and it can be expected that they will provide significant assistance in the discovery and understanding of the mounting volume of biological and medical data.

In this chapter, we first provide an overview of the commonly used ML and DL techniques and strategies and outline their broad areas of applicability with regard to processing and analysis of biological and medical data. Next, we attempt to summarise the available corpus of research and development concerning the application of ML and DL techniques to the process of analysis and interpretation of biomedical data, focusing on liquid biopsy analysis, outline several of the main avenues of such research, and predict the potential improvements and changes in this highly dynamic and quickly developing field. Expertise in ML is not a prerequisite for this chapter, although we assume basic overall familiarity with the most well-known ML and DL models, techniques, and methodologies.

2. Machine learning strategies

This overview is limited to classical software-based tools and techniques for brevity's sake. Several hardware-based approaches are mentioned in the Future Prospects section.

The ML ecosystem is both extensive and complex [3–5], with many possible ways to subdivide or classify its members. One frequently used classification scheme outlines two broad groups of ML algorithms: supervised learning, where the model is presented with both a set of labelled example inputs and desired outputs (called the training dataset), with the goal to learn a mapping from inputs to outputs, and unsupervised learning, where no labels are given to the model, leaving it to learn the input-output mapping in unstructured data. A notable specific case of supervised learning is reinforcement learning (RL), where training data consists only of positive ("reward") and negative ("punishment") feedback, given according to the model's performance in the training environment.

Another informative approach to classifying ML algorithms is based on the desired type of output of the given model, such as classification (division of the input data into two (binary classification) or more (multi-label classification) predetermined groups), clustering (similar to classification but with the groups not known beforehand), dimensionality reduction (simplification of high-dimensional input data by mapping them into a lower-dimensional space), search, etc. Of these, clustering is particularly notable due to its broad and general applicability

http://dx.doi.org/10.5772/intechopen.79404

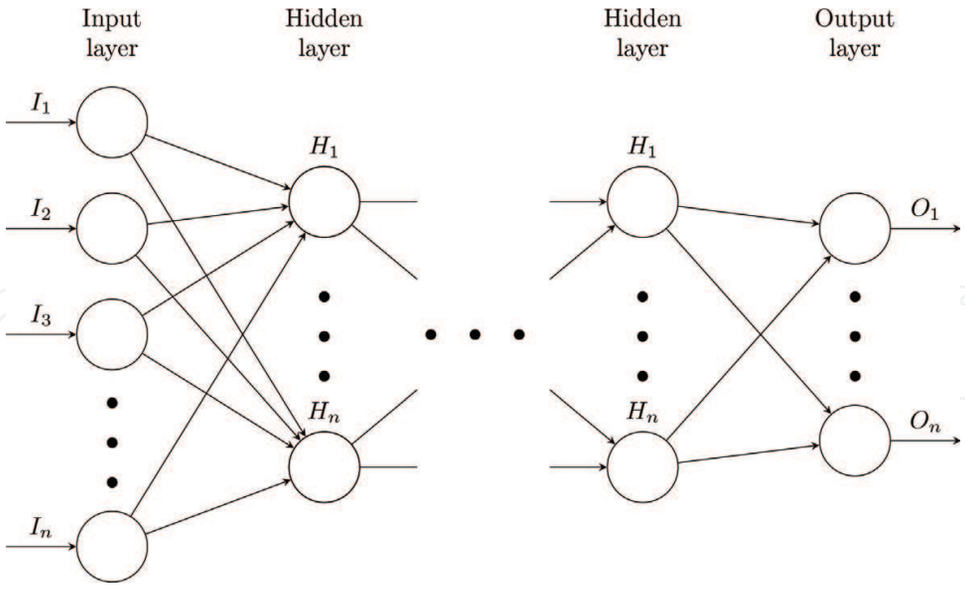

Figure 1. The structure of a typical feed-forward deep neural network, with a fully connected input layer *I*, an unspecified number of hidden processing layers *H*, and an output layer *O*.

and the wide range of models, methods, and algorithms [6–8] that can be employed to carry out cluster analysis.

The notion of "cluster" is often not precisely defined and tends to serve as an umbrella term for various types of data objects, typically groups of data points with small distances (appropriately defined) between group members, higher-density areas of some parameter space, particular statistical distributions, etc. The desired clustering algorithm, therefore, depends on both the given data set and the intended application of the returned results. Due to these complications, clustering, like many other data analysis methods, is typically not fully automated even within the domain of machine learning but instead tends to partially rely on preprocessing and initial parameter selection, based on the specifications of the task at hand.

Deep learning is a subclass of machine learning problems, the distinction being based on the training data representations instead of specific algorithms. Similar to ML in general, deep learning can be both supervised and unsupervised [9]. Deep learning models tend to be vaguely similar to information processing patterns in biological brains (and are therefore often called artificial neural networks), in that they use multiple layers [10] of non-linear processing units (frequently called "neurons", even though their similarity to biological neurons is usually limited) for pattern recognition and transformation, with each successive layer using as inputs the output from a previous layer, forming a hierarchy of representations and levels of abstraction. The number of hidden layers of an artificial neural network broadly determines the "computational power" of the network [11] (**Figure 1**).

Machine learning models have been applied to a wide variety of fields and problem classes, including computer vision, natural language processing, machine translation, bioinformatics and biochemistry [12], with results often similar or superior [13] to those produced by human domain experts.

3. Using machine learning techniques in blood tests

3.1. Classifying blood cells with deep convolutional neural networks

An important part of the data acquired by blood tests is the number of white blood cells (WBCs) or leukocytes, usually differentiated into total and differential WBC count, where the latter describes the absolute and relative numbers of WBC subtypes (neutrophils, lymphocytes, basophils, eosinophils, and monocytes) in the sample. The amount of WBCs in the sample provides information on the state of the patient's innate and adaptive immune system, e.g., a significant changes in the WBC count relative to the patient's baseline is evidence that their body is being affected by an antigen, whereas variations in the specific WBC subtypes can correlate with specific types of antigens or different pathways of immune and inflammatory reaction. Therefore, detailed measurement and understanding of the WBC counts is an important part of the quantitative picture of health and the organism's general condition.

Traditional methods of estimating the WBC count generally fall into one of two categories—manual and automated. The historical manual inspection of the blood sample involved counting the number of cells in a blood sample under a microscope and extrapolating under the assumption of uniform cell distribution across the entire bloodstream. Automated methods involve specialised equipment such as Coulter counters [14] or laser flow cytometers [15] which can provide accurate results and good performance [16] but are generally expensive and require specialised training to operate.

In this light, the ML-based approach provides a potential improvement over the aforementioned techniques due to several reasons. First, it requires far less expensive equipment due to being built around simple imaging solutions. Furthermore, unlike earlier methods, it is able to provide almost instantaneous results after the initial training stage. Finally, its performance can be expected to improve over time, in proportion to growing dataset sizes and, being mostly software-based, it can be expanded and advanced continually and "over the air", without requiring extensive changes in the underlying infrastructure.

We illustrate this approach using an example problem provided by *Athelas* team [17], namely, binary classification of a stained image of a WBC as either polymorphonuclear or mononuclear.[1] The training dataset consisted of hand-labelled images of stained WBCs of all given types in various proportions. Before the dataset could be used, several preprocessing steps were taken, including removing images with multiple cells and using transformations such as flips and rotations in order to increase the size and variability of the training dataset. By using transformed versions of the images, the training dataset size was increased from approximately 350 to 10^4.

For the ML model, *Athelas* team used the LeNet-5 [18] convolutional neural network (CNN) [3, 4, 9] due to its simplicity and availability. The model was tested against a test dataset of 71 images (20% of the original training set and 0.7% of the training set after transformations), with

[1] Eosinophils, basophils, and neutrophils are polymorphonuclear, while lymphocytes and monocytes are mononuclear.

the high accuracy of 98.6%. While the presently used model performs less well (accuracy of 86%) when classifying WBCs into multiple individual type categories as opposed to binary classification, given the high performance and simplicity of this purely software-based approach, *Athelas* team plans to extend it to more complex problems, including datasets containing other cell types, which could enable faster improvement cycles, increased accessibility, and better patient outcomes, compared to previously used methods of cell count analysis.

3.2. Using deep neural networks for detection of ageing-related biomarkers

During the last decade, human ageing research has received an increasing amount of mainstream interdisciplinary attention [19, 20], with an emerging tendency to approach various aspects of the natural ageing process as potentially treatable conditions.

Insilico team developed a DL system designed to predict human chronological age from biochemical data obtained from a basic blood test [21], narrowing an extensive set of potential ageing-related biomarkers to a limited subset of the most salient ones. A dataset of $> 6 \times 10^4$ records was used, with each record consisting of a patient's age, sex, and 46 blood biochemical markers. The dataset was preprocessed, normalising all blood marker values to 0–1 range, and then split into training and test datasets with ratio 90:10.

An ensemble of 21 feed-forward deep neural networks (DNNs) was created as the ML model, with a range of values assigned to DNN parameters such as the number of hidden layers, the number of processing units per layer, activation function, and optimization and regularisation methods. The permutation feature importance method [22] was used to evaluate the relative importance of the various biochemical markers with regard to ensemble accuracy. Batch normalisation [23] was used to reduce the effects of overfitting and increase the stability of convergence of the models.

The best results were obtained from a DNN with five hidden layers, using regularised mean squared error (MSE) function as the loss function, parametric rectified linear unit (PReLU) [24] activation function in each layer, and AdaGrad [25] optimiser of the loss function. The highest-scoring DNN performed with 82% ε-prediction accuracy at $\varepsilon = 10$ (i.e., considering the sample as correctly recognised if the predicted age is ± 10 years of the true age), out-performing several classes of competing ML models. Multiple models for combining individual DNNs into an ensemble (stacking) were evaluated, with the best being the elastic net model [26]. The most important blood markers were discovered to be albumin, glucose, alkaline phosphatase, urea, and erythrocyte count.

Insilico team created an online service (http://www.aging.ai) to make the DNN ensemble available to the general public, allowing patients to use their blood test data to evaluate the age prediction system and serving as a proof of concept for estimating ageing-related variables using readily available biochemical data. Additional data sources, including transcriptomic and metabolomic markers from liquid and individual organ biopsies, as well as imaging data, are being considered. *Insilico* team suggests that similar systems could also be developed for model organisms in order to perform cross-species analysis of individual biological markers and their importance in predicting both chronological and biological age.

3.3. Machine learning-based approach to Alzheimer disease biomarker discovery

In their study, *Smalheiser* team has developed [27] a ML-based model for predicting Alzheimer disease (AD) status of individual samples with high accuracy, using miRNAs and other small RNAs extracted from circulating exosomes obtained from liquid biopsy (blood plasma) samples.

A sample set of $N = 70$ was used to construct the training dataset consisting of normalised miRNA expression data across 465 loci. Cross-validation was used in feature selection to evaluate the impact of values from specific loci as features. The samples were randomly divided into 7 partitions of 5 positive and 5 negative samples each and cross-validation was performed on these partitions, using 6 partitions for training and 1 for evaluation. The random partitioning was repeated 10 times in order to acquire 70 estimate points of the performance measures of interest, one for each sample in the set. These values were averaged and their relative performance was assessed using area under the curve of the receiver operating curve (ROC), Matthews correlation coefficient (MCC) [28], and F1 score.

Smalheiser team evaluated three different ML classifier algorithms—C4.5 decision trees [29] (using the J48 implementation), support vector machines (SVMs) [30], and adaptive boosting (AdaBoost) [31]. After selecting 50 most significant features, as per Mann-Whitney *U* test [32], the C4.5 classifier produced the best results, based on which it was selected as the feature selection method. The feature significance was measured by the number of times the given miRNA locus was used as a node in the decision tree over the 70 runs. The 18 highest-scoring features were selected to move on to the next step. AdaBoost algorithm was used for the final feature selection from the set of 18 features, producing an optimised set of 7 features which were then used with all 70 data samples to produce the final dataset.

The best model used by *Smalheiser* team was able to correctly classify, on average, 29 out of 35 samples from the AD group and 31 out of 35 samples from the control group, yielding accuracy in the range of 83–89%. *Smalheiser* team concluded that ML-based classifiers are able to produce highly accurate predictions of AD occurrence, using a dataset of only 7 miRNAs and that integrating exosome miRNA data with other data is likely to further increase performance of these models.

3.4. Detection and classification of circulating tumour cells using machine learning methods

The presence of circulating tumour cells (CTCs) in blood samples indicates the tumour response to chemotherapeutic drugs and contributes to the mechanism for subsequent growth of derived tumours (metastasisation) in distant tissues. Evaluation of CTCs can yield the diagnosis or help to follow the tumour response to chemotherapeutic drugs.

Mao team designed a deep (six layers) CNN for image-based circulating tumour cell detection with automatically learned network parameters [33]. They used a dataset of 45 phase contrast microscopy [34, 35] images, of which 35 randomly selected images were used for training and the remaining 10 for testing the network. The experiment was repeated 5 times in order to minimise network bias.

The CNN received normalised 40 × 40 pixel images as input. They were passed to a layer of 6 convolutional filters with the size of 5 × 5, followed by a max-pooling layer in order to extract the local signal in every 2 × 2 pixel region, defined by the max-pooling function,

$$z^i_{p,q} = \max_{0 \le m,\, n \le 2} \left\{ y^i_{2\times p+m,\, 2\times q+n} \right\}, \tag{1}$$

where (p, q)—pixel coordinates, y—input map, z—output map. This layer was followed by another convolutional filter layer, consisting of 12 filters, and, subsequently, by another max-pooling layer. The last layer was fully connected to the output layer by way of dot product between the weight and input vectors, passed to the sigmoid function which maps the values to the $[-1, 1]$ range. The filter parameters, network bias terms, and weight matrices were automatically adjusted by backpropagation with learning rate set to 0.1.

Mao team compared their CNN-based classifier to a simpler, SVM-based method that depended on hand-crafted feature sets. Using the F-score (harmonic mean of precision and recall scores) as the comparison metric, they found that, after two rounds of five iterations, the F-score of the CNN-based classifier was 0.97, by 18.6 points exceeding the F-score (0.784) of the SVM-based classifier and hand-crafted feature set. They concluded that the CNN-based classifier presents a promising development towards automated CTC detection in images taken from blood samples, and that the technique could be adapted for use with microfluidics-based liquid biopsy platforms for early diagnosis and monitoring.

4. Cancer detection and monitoring using neural network-based methods

4.1. Using artificial neural networks for lung cancer detection and diagnosis

Goryński team describes [36] an artificial neural network (ANN)-based model class used for early detection and diagnosis of lung cancer. In their study, a dataset consisting of a wide range of biochemical parameters obtained from blood samples, as well as results from medical interviews (48 values in total) from 193 patients of mixed age and sex was used to train a family of 10 multilayer perceptron network (MLP) [3, 4] architectures, using a range of activation functions (linear, logistic, and tanh) for both hidden and output layers, as well as varying number of processing units ("neurons") in the hidden layer and different training algorithms (gradient descent, Broyden-Fletcher-Goldfarb-Shanno (BFGS) [37], and scaled conjugate gradient (SCG)) [38].

Goryński team found that two of the trained models, named MLP 48–9-2[2] (trained using BFGS algorithm and using linear and tanh activation functions for hidden and output layers, respectively) and MLP 48–15-2 (SCG algorithm, logistic and tanh activation functions) gave highly

[2]The naming scheme represents the number of "neurons" in the input, hidden, and output layers of the MLP model, respectively.

accurate results in terms of inferring the presence or absence of lung cancer from the given set of variables, with ROC value reaching 99.83%.

Goryński team concluded that these, relatively simple, ANN solutions, while not viable as a full substitute of expert opinion, are nonetheless efficient in early diagnosis and risk prognosis of lung cancer and therefore are promising as potential improvements over and additions to the existing inventory of diagnostic and prognostic methods.

4.2. Mutation prediction and early lung cancer detection in liquid biopsy using convolutional neural networks

The proliferation of cancer cells is driven by specific somatic mutations in the cancer genome [39]. To fulfil the high expectations associated with liquid biopsy, such as comprehensive characteristics of the whole tumour in contrast to limited sampling in the traditional tissue biopsy, or dynamic assessment during treatment, the somatic mutations must be detected with high sensitivity and accuracy; limited coverage depth is not sufficient. *Kothen-Hill* team has demonstrated a CNN-based classifier system named "Kittyhawk" [40] that enables the detection of cancer-related mutations even in extremely low variant allele frequencies (VAFs), more than 2 orders of magnitude lower than is possible with the currently available methods.

For training dataset, whole genome sequencing (WGS) data from 4 non-small cell lung cancer (NSCLC) patients and 3 melanoma patients were used, with $> 1.2 \times 10^7$ reads in total. To ensure adequate genetic context regardless of variants appearing at the end of the read, additional bases were added to both ends of the read. Additional bases were also added to ensure equal read length in cases where a read is shorter than 150 bp.

Kothen-Hill team chose an 8-layer CNN with a single fully connected output layer, similar to the VGG[3] architecture [41], with a perceptive field of size 3 used to convolve the features, based on results of [42] who showed that the tri-nucleotide context contains distinct mutagenesis-related signatures. After 2 successive convolutional layers, downsampling by max-pooling with a receptive field of 2 and a stride of 2 was applied, forcing the model to retain only the highest-importance features, as per [43]. The output of the last convolutional layer was directly connected to a fully connected sigmoid output layer for final classification. A logistic regression layer was used to retain the features associated with the position of the read.

The model was trained using minibatch stochastic gradient decent (SGD) with batch size of 256, initial learning rate of 0.1, and momentum of 0.9, with batch normalisation [23] and a rectified linear unit (RLU) [44] applied after each convolutional layer.

Kothen-Hill team presents the Kittyhawk architecture as a first of its specific kind, being able to avoid the information loss associated with similar earlier architectures. To evaluate the performance of the model, a test dataset consisting of $> 2 \times 10^5$ reads that were split off the training set of reads from the 4 NSCLC patients was used. *Kothen-Hill* team found that the model achieves F1 accuracy of 0.961 when using this test dataset, and 0.92 when using data from an

[3] A CNN architecture developed by the Visual Geometry Group at University of Oxford.

additional independent NSCLC case. When further tested against data from a melanoma case, F1 accuracy of 0.71 was achieved, indicating that the model had learned specific mutation patterns associated with NSCLC, as well as a more general pattern associated with both NSCLC and melanoma.

Kothen-Hill team presents the Kittyhawk CNN model as the first ML architecture designed specifically for detecting cancer-related mutations in a low allele frequency environment, such as liquid biopsy and might serve as the foundation for novel early stage cancer detection techniques that could be used for both screening and prognosis.

4.3. Machine learning and nanofluidics in pancreatic cancer diagnosis

Issadore team has developed a ML-based platform [45] for isolating exosomes from liquid biopsy samples and, using the RNA inside these exosomes to diagnose pancreatic cancer in human and murine cohorts.

Using the Exosome Track-Etched Magnetic Nanopore (ExoTENPO) nanofluidics chip developed as part of the study, *Issadore* team successfully isolated exosomes from cell cultures, as well as human and mouse liquid biopsy (blood plasma) samples. Exosomal mRNA was subsequently extracted and used to develop a predictive panel for pancreatic cancer biomarkers.

Training datasets of 15 mouse and 10 patient profiles, respectively, were created. Linear discriminant analysis (LDA) [46] was used to identify combinations of mRNA profile that discriminated between healthy and tumour-bearing samples. The prediction algorithm was generated by running LDA on the training set, which produced a vector that was used to calculate a weighted sum such that it maximally separates the control group from the sample group with tumours. Two independent blinded test sets, mouse ($N = 18$) and patient ($N = 34$), respectively, were used to evaluate the performance of the LDA classifier. Fisher's exact test was used to quantify the predictive value of the classifier, yielding $P < 0.001$.

Although in their study *Issadore* team focused primarily on the development and evaluation of the ExoTENPO nanofluidics platform, they conclude that even very simple ML algorithms such as LDA can produce good quality predictive models for classifying biochemical and genetic markers and note that more advanced ML solutions could be used in future research in order to further improve performance.

4.4. Machine learning-based RNA sequencing for multi-class cancer diagnostics

Wurdinger team demonstrated a ML-based approach to sequencing and analysis of mRNAs obtained from tumour-educated platelets (TEPs) [47] as a tool for accurate tumour diagnosis, both within a single class and across six different tumour classes.

The initial dataset consisted of blood platelet samples from healthy donors ($N = 55$) and both treated and untreated patients with six different tumour types (NSCLC, colorectal cancer, glioblastoma, pancreatic cancer, hepatobiliary cancer, and breast cancer) in various stages of advancement and metastasis ($N = 228$). After the mRNA extraction, amplification, and sequencing, a set of approximately 5000 different mRNAs was selected for further analysis.

The accuracy of TEP-based multi-class cancer classification in the training dataset ($N = 175$) was estimated, using an SVM algorithm. To cross-validate the SVM for the entire sample set, leave-one-out cross-validation (LOOCV) method was applied. The percentage of correct predictions was reported as the accuracy score. The algorithm was performed 175 times, in order to classify and cross-validate the entire dataset. To determine specific input gene lists for the algorithm, *Wurdinger* team performed ANOVA testing. They selected a set of 1072 mRNAs to use with the training dataset, yielding final accuracy of 96% and ROC value of 0.986. From the patient cohort, all 39 patients with localised tumours and 33 of the 39 patients with primary tumours in the CNS were classified as cancer patients.

Wurdinger team concluded that using the SVM classifier with TEP-based data produces high-accuracy, high-specificity models for liquid biopsy-based diagnostics for several common cancer types. They expect that using more advanced ML algorithms capable of self-learning could further improve the performance of these diagnostic models. They also suggest evaluating systemic factors such as inflammatory diseases and other non-cancerous diseases as potential factors that can influence the mRNA profile.

5. Using machine learning to accelerate DNA sequencing and biomarker development

5.1. A supervised machine learning-based approach to DNA sequence analysis

DNA sequencing and sequence analysis is an important task in many scientific and medical fields that is well-known for being both data-rich and computationally intensive. *Memeti* & *Pllana* describe a ML-based solution for optimised DNA sequence analysis [48, 49]. Their algorithm leverages the increased performance and parallelisation capabilities of heterogeneous (a host central processor (CPU) in combination with a 61-core Intel Xeon Phi co-processor) multi-core computing platform.

Memeti & *Pllana* used the widely known Aho-Corasick (AC) algorithm [50] as the basis for their work, since DNA analysis is a specific case of a string matching problem, where the input text is the given DNA sequence and the alphabet consists of characters corresponding to the four nucleotide bases. AC uses finite automata (FA), a simple type of formal machine in the form of a prefix tree with additional links between internal nodes. These links allow for fast failure transitions (also known as ε-transitions) between branches of the tree that share a common prefix, thus avoiding backtracking. A known drawback of the AC algorithm is its being non-deterministic. *Memeti* & *Pllana* solved the non-determinism issue by modifying the AC finite automaton so that it computes the correct transition for each state, thus eliminating failure transitions and guaranteeing that every character always has the same number of operations associated with it.

A boosted decision tree regression-based predictor [51] was used to estimate the execution time of DNA sequence analysis for both the host CPU and the Intel Xeon Phi co-processor. The predictor's output was used to partition the DNA sequence based on the S-factor,

$$S = \frac{T_{host}}{T_{device}}, \tag{2}$$

where T_{host} and T_{device} are execution times for the host CPU and the co-processor, respectively, and using the partitioning scheme

$$I_{host} = I - I_{device} \tag{3}$$

$$I_{device} = \frac{I}{S+1}, \tag{4}$$

where I is the original DNA sequence, I_{host} is the part of I analysed by the host CPU, and I_{device} is the part of I analysed by the co-processor.

Memeti & *Pllana* used the "single instruction, multiple data" (SIMD) parallelism [52] of both the host CPU and the Xeon Phi co-processor to achieve teraFLOP (10^{12} floating point operations per second) performance. For experimental evaluation of their deterministic finite automata (DFA) algorithm, *Memeti* & *Pllana* used reference genomes of human and 11 different animals from the GenBank sequence database of the National Center for Biological Information, with the average dataset size of 2043 MB. In total, data from approximately 4000 experiments was used to train the performance predictor and to evaluate the DFA performance. The DFA performance was evaluated using different thread affinity modes (*compact*, *balanced*, and *scatter*) and numbers of threads for each of the DNA sequences. The *balanced* thread affinity mode evenly distributes the threads among the computing cores, *compact* mode completely fills a single core with threads before assigning the remaining threads to the next core, while the *scatter* mode distributes threads among the cores in a round-robin sequence.

Memeti & *Pllana* discovered that the balanced thread affinity mode is overall fastest for all of the tested DNA sequences, with second best being the scatter mode. The evaluation of DFA with regard to varying thread counts showed that the algorithm scales well up to approximately 120 threads, whereas in the 180–240 thread range the performance improvement becomes modest due to overhead from thread management operations. Performance-wise, *Memeti* & *Pllana* found that the parallel version of DFA running on a heterogeneous platform has a speed-up from 35.6× up to 206.6×, compared to a sequential (single-thread) version running on the host CPU, with the exact speed-up degree depending on the given host CPU. *Memeti* & *Pllana*. intend to use this work to study and develop highly parallel DNA analysis solutions on more powerful hardware in the future.

6. Future prospects

While the ML models currently used in liquid biopsy analysis in particular and biological and medical research in general (typically different classes of neural networks and linear classifiers) appear to both produce accurate results and show generally high performance, they represent only a narrow subset of machine learning and artificial intelligence solutions [5]. For instance,

a potentially valuable research direction might be in the form of highly advanced probabilistic graphical models [53] augmented with functionality such as one-shot learning [54] and probabilistic program synthesis [55], which could potentially allow researchers to reduce the size of the commonly massive training datasets required for creating ANN- or DL-based models.

Furthermore, with a single exception, all of the studies reviewed here have been focused on the performance and accuracy of software ML models, which is currently the predominant class of machine learning solutions. However, recent advances in general purpose computation using both graphics processing units (GPUs) and specialised application-specific integrated circuits (ASICs) tailor-made for machine learning [56] provide a strong case for the exploration and exploitation of hardware or hybrid ML solutions, as evidenced by, e.g., the results from the AlphaGo experiments and public performance [57].

7. Conclusions

Liquid biopsy-based approaches open many so far little explored and promising opportunities for studying and measuring biological and biochemical markers with broad applications for the monitoring, diagnosis, and prognosis of a large class of diseases and processes. Machine learning, with its advanced pattern recognition capabilities, will likely play an increasingly important role in these fields, as the amount and complexity of data produced by scientific and medical sources already by far exceeds the capacity of unaided human experts and is rapidly increasing with no foreseeable slowdown.

In addition, machine learning tools form a natural synergy with distributed, highly parallel, or cloud-based computation solutions, thus easily yielding to collaboration among researchers and medical professionals from distant locations and involving amounts of data storage and processing power previously available only on dedicated high performance computing (HPC) platforms and supercomputers. It is likely that in the near future the importance of decentralised collaboration will continue to grow, increasing the demand for powerful and easy to use toolset for analysis and processing of biological data.

Based on these trends, we expect that the next generation of liquid biopsy technologies will include many types of machine learning as an integral part of their operation and that this trend could have a significant positive impact on both diagnosis and treatment of patients.

Acknowledgements

The present work was carried out within the frame of scientific project № 1.1.1.2 VIAA 1 16 242.

Conflict of interest

The authors declare that the chapter was written in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author details

Arets Paeglis[1]*, Boriss Strumfs[2], Dzeina Mezale[1] and Ilze Fridrihsone[1]

*Address all correspondence to: arets.paeglis@protonmail.com

1 Department of Pathology, Riga Stradins University, Riga, Latvia

2 Latvian Institute of Organic Synthesis, Riga, Latvia

References

[1] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review. 1958;**65**(6):386-408. DOI: 10.1037/h0042519

[2] Stormo GD, Schneider TD, Gold L, et al. Use of the 'perceptron' algorithm to distinguish translational initiation sites in *E. coli*. Nucleic Acids Research. 1982;**10.9**:2997-3011. DOI: 10.1093/nar/10.9.2997

[3] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. May 2015;**521**(7553):436-444. DOI: 10.1038/nature14539

[4] Goodfellow I, Bengio Y, Courville A. Deep Learning. USA: MIT Press; 2016. URL: http://www.deeplearningbook.org

[5] Russell S, Norvig P. Artificial Intelligence: A Modern Approach. 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall Press; 2009. ISBN: 0136042597; 9780136042594

[6] Estivill-Castro V. Why so many clustering algorithms. ACM SIGKDD Explorations Newsletter. 2002;**4**(1):65-75. DOI: 10.1145/568574. 568575

[7] MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. Berkeley, California: University of California Press; 1967. pp. 281-297

[8] Kriegel HP, Kröger P, Sander J, et al. Density-based clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011;**1**(3):231-240. DOI: 10.1002/widm.30

[9] Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks. 2015; **61**:90-93. DOI: 10.1016/j.neunet.2014.09.003

[10] Deng L. Deep learning: Methods and applications. Foundations and Trends® in Signal Processing. 2014;**7**(3–4):197-387. DOI: 10.1561/2000000039

[11] Hornik K. Approximation capabilities of multilayer feedforward networks. Neural Networks. 1991;**4**(2):251-257. DOI: 10.1016/0893-6080(91) 90009-t

[12] Ghasemi F, Mehridehnavi A, Fassihi A, et al. Deep neural network in QSAR studies using deep belief network. Applied Soft Computing. 2018;**62**:251-258. DOI: 10.1016/j.asoc.2017.09.040

[13] Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; June 2012. DOI: 10.1109/cvpr.2012.6248110

[14] DeBlois RW, Bean CP. Counting and sizing of submicron particles by the resistive pulse technique. Review of Scientific Instruments. 1970;**41**(7):909-916. DOI: 10.1063/1.1684724

[15] Telford WG. Lasers in flow cytometry. In: Methods in Cell Biology. USA: Elsevier; 2011. pp. 373-409. DOI: 10.1016/b978-0-12-374912-3.00015-8

[16] Fleisher TA, de Oliveira JB. Flow cytometry. In: Clinical Immunology. USA: Elsevier; 2008. pp. 1435-1446. DOI: 10.1016/b978-0-323-04404-2.10097-1

[17] Parthasarathy D. Classifying White Blood Cells. Blog. 2017. URL: https://github.com/dhruvp/wbc-classification/blob/master/notebooks/binary%5C_training.ipynb

[18] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998;**86**(11):2278-2324. DOI: 10.1109/5.726791

[19] de Grey A, Rae M. Ending Aging: The Rejuvenation Breakthroughs that could Reverse Human Aging in our Lifetime. USA: St. Martin's Griffin; 2008. ISBN: 0312367074

[20] Cortese F. Longevity Industry Landscape Overview 2017. Technical Report Johns Hopkins University, Biogerontology Research Foundation; January 2017. URL: http://lir.website/pdfdata/pdf2017/

[21] Putin E, Mamoshina P, Aliper A, et al. Deep biomarkers of human aging: Application of deep neural networks to biomarker development. Aging. 2017;**8**(5):1021-1033. DOI: 10.18632/aging.100968

[22] Altmann A, Toloşi L, Sander O, et al. Permutation importance: A corrected feature importance measure. Bioinformatics. 2010;**26**(10):1340-1347. DOI: 10.1093/bioinformatics/btq134

[23] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Bach F, Blei D, editors. Proceedings of the 32nd International Conference on Machine Learning. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR; July 2015. pp. 448-456

[24] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: 2015 IEEE International Conference on Computer Vision (ICCV). IEEE; December 2015. DOI: 10.1109/iccv.2015.123

[25] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research. 2011;**12**:2121-2159. ISSN: 1532-4435

[26] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B. 2005;**67**:301-320

[27] Lugli G, Cohen AM, Bennett DA, et al. Plasma exosomal miRNAs in persons with and without Alzheimer disease: Altered expression and prospects for biomarkers. PLoS One. 2015;**10**(10):e0139233. DOI: 10.1371/journal.pone.0139233. Zhang B, editor

[28] Matthews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) – Protein Structure. 1975;**405**(2):442-451. DOI: 10.1016/0005-2795(75)90109-9

[29] Quinlan JR. C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning). USA: Morgan Kaufmann; 1992. ISBN: 1558602380

[30] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;**20**(3):273-297. DOI: 10.1007/bf00994018

[31] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences. 1997;**55**(1):119-139. DOI: 10.1006/jcss.1997. 1504

[32] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics. 1947;**18**(1):50-60. DOI: 10.12 14/aoms/1177730491

[33] Mao Y, Yin Z, Schober J. A deep convolutional neural network trained on representative samples for circulating tumor cell detection. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE; March 2016. DOI: 10.1109/wacv.2016.7477603

[34] Zernike F. Phase contrast, a new method for the microscopic observation of transparent objects. Physica. 1942;**9**(7):686-698. DOI: 10.1016/s0031-8914(42)80035-x

[35] Zernike F. Phase contrast, a new method for the microscopic observation of transparent objects part II. Physica. 1942;**9**(10):974-986. DOI: 10.1016/s0031-8914(42)80079-8

[36] Goryński K, Safian I, Grądzki W, et al. Artificial neural networks approach to early lung cancer detection. Open Medicine. 2014;**9**(5). DOI: 10.2478/s11536-013-0327-6

[37] Nocedal J, Wright S. Numerical Optimization (Springer Series in Operations Research and Financial Engineering). USA: Springer; 2006. ISBN: 0387303030

[38] Møller MF. A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks. 1993;**6**(4):525-533. DOI: 10.1016/s0893-6080(05)80056-5

[39] Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;**505**(7484):495-501. DOI: 10.1038/nature12912

[40] Kothen-Hill ST, Zviran A, Schulman RC et al. Deep Learning Mutation Prediction Enables Early Stage Lung Cancer Detection in Liquid Biopsy. ICLR 2018 Conference Vancouver, Canada. 2018

[41] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: CoRR abs/1409.1556; 2014. arXiv: 1409.1556

[42] Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature. 2013;**500**(7463):415-421. DOI: 10.1038/nature12477

[43] Boureau YL, Ponce J, LeCun Y. A theoretical analysis of feature pooling in visual recognition. In: 27th International Conference on Machine Learning, Haifa, Israel; 2010

[44] Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10. Haifa, Israel: Omnipress; 2010. pp. 807-814. ISBN: 978-1-60558-907-7

[45] Ko J, Bhagwat N, Yee SS, et al. Combining machine learning and nanofluidic technology to diagnose pancreatic cancer using exosomes. ACS Nano. 2017;**11**(11):11182-11193. DOI: 10.1021/acsnano.7b05503

[46] Hilbe JM. Logistic Regression Models (Chapman & Hall/CRC Texts in Statistical Science). United Kingdom: Chapman and Hall/CRC; 2009. ISBN: 1420075756

[47] Best M, Sol N, Kooi I, et al. RNA-Seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. Cancer Cell. 2015; **28**(5):666-676. DOI: 10.1016/j.ccell.2015.09.018

[48] Memeti S, Pllana S. Accelerating DNA sequence analysis using Intel Xeon Phi. In: CoRR abs/1506.08612; 2015. arXiv: 1506.08612

[49] Memeti S, Pllana S. A machine learning approach for accelerating DNA sequence analysis. The International Journal of High Performance Computing Applications. 2018;**32**(3):363-379. DOI: 10.1177/1094342016654214

[50] Aho AV, Corasick MJ. Efficient string matching: An aid to bibliographic search. Communications of the ACM. 1975;**18**(6):333-340. DOI: 10.1145/360825.360855

[51] Rohrer B. How to Choose Algorithms for Microsoft Azure Machine Learning. Blog. 2017. https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-choice

[52] Patterson DA, Hennessy JL. Computer Organization and Design: The Hardware/Software Interface. USA: Morgan Kaufmann; 1997. ISBN: 1558604286

[53] Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques – Adaptive Computation and Machine Learning. USA: The MIT Press; 2009. ISBN: 0262013193, 9780262013192

[54] Fei-Fei L, Fergus R, Perona P. One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006;**28**(4):594-611. DOI: 10.1109/tpami.2006.79

[55] Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. Science. 2015;**350**(6266):1332-1338. DOI: 10.1126/science.aab3050

[56] Jouppi NP, Young C, Patil N, et al. In-datacenter performance analysis of a tensor processing unit. In: CoRR abs/1704.04760; 2017. arXiv: 1704.04760

[57] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. Nature. 2017;**550**(7676):354-359. DOI: 10.1038/nature24270