Directory of Machine Learning Tools, Platforms and Packages

Compiled by

H3ABioNet 2018

https://h3abionet.org

Work Package: Big Data Analytics and Machine Learning Tools for Application To Biomedical Data

Project lead: Amel Ghouila

Handbook development lead: Zahra Mungloo- Dilmohamud

Table of Contents

Contributors:

H3ABioNet
Amel Ghouila *

KRISP, Kzn
San Emmanuel James

Malawi-Liverpool-Wellcome Trust
Anmol Kiran

University of Cape Town, South Africa
Gerrit Botha
Nicki Tiffin

University of Illinois Urbana-Champaign, USA
Liudmila Mainzer
Prakruthi Burra
Christopher Fields

University of Mauritius, Mauritius
Shakuntala Baichoo
Zahra Mungloo-Dilmohamud
Dassen Sathan
Anisah Ghoorah

*** Handbook development lead: Zahra Mungloo- Dilmohamud**
Email: z.mungloo@uom.ac.mu

Introduction

Machine learning consists of programming computers to optimize a performance criterion by using example data or past experience. There are many tools, platforms and packages that are available in this field. A platform provides all you need to run a project, whereas a library only provides discrete capabilities or parts of what you need to complete a project. Below is a list of some of the most common ones (ordered alphabetically).

1. Accord.Net

Type	.NET machine learning framework
Open Source	Open Source
Availability	http://accord-framework.net/
Description	Accord.NET is a framework for scientific computing in .NET. It is combined with audio and image processing libraries which encompass a range of scientific computing applications such as machine learning, statistical data processing and pattern recognition
Operating Systems	Microsoft Windows, Xamarin, Unity3D, Windows Store applications, Linux and Mobile.
Input and Output - formats/ types of data etc	The Accord.IO library provides data readers for formats like Excel, comma-separated values, Matlab matrix files, LibSVM and LibLinear's data formats, and others.
Limitations	Not a very popular framework. Slow compared to TensorFlow.
Bioinformatics Use	Mostly used for audio and image processing Not specific to Bioinformatics but can be used for Classification (Kernel SVMs, Naive Bayes and Decision Trees), Regression (Kernel SVM), Clustering (K-Means and MeanShift, Gaussian Mixture Models), and for Feature Selection (L1-regularized Logistic SVMs)

2. Alibaba Machine Learning Platform

Type	Cloud service
Open Source	Fee Paying
Availability	https://www.alibabacloud.com
Description	The cloud computing business offers a machine learning service to help enterprise customers streamline analytics software development. It provides end-to-end machine learning services, including data processing, feature engineering, model training, model prediction, and model evaluation. It has more than 100 sets of algorithms covering data preprocessing, neural network, regression, classification, predictions, evaluations,

	statistical analysis, feature engineering and deep-learning architecture.
Operating Systems	Cloud-based
Input and Output - formats/ types of data etc	Text and image data
Limitations	Sparse documentation. Geared toward companies.
Bioinformatics Use	None

3. Amazon AWS Services

Type	Cloud Service
Open Source	Fee paying for heavy use
Availability	https://aws.amazon.com/
Description	AWS provides a number of stable machine learning APIs to be consumed off the shelf. ● **Lex**, which is the underlying technology for its Alexa AI voice assistant; ● **Polly** for text-to-voice services, and ● **Rekognition** for adding image analysis and facial recognition to apps. ● **Transcribe** for converting speech to text; ● **Amazon Translate** for translating text between languages; ● **Amazon Comprehend** for understanding natural language; ● **Amazon Rekognition Video**, a computer vision service for analysing videos in batches and in real-time.
Operating Systems	Cloud-based
Input and Output - formats/ types of data etc	Text, Audio, Video
Limitations	Limits in input and output E.g. Input: Speech:15 s, Text:1024 chars E.g. Output:25 kb, Speech output:10 mins
Bioinformatics Use	None

4. Amazon AWS Machine Learning Service

Type	Cloud Service
Open Source	Fee paying for heavy use
Availability	https://aws.amazon.com/machine-learning/
Description	Amazon Machine Learning offers a managed service for developers and data scientists building machine learning models and generating predictions. Amazon Machine Learning combines powerful machine learning algorithms with interactive visual tools to guide you towards easily creating, evaluating, and deploying machine learning models.

Operating Systems	Cloud-based
Input and Output - formats/ types of data etc	Input: text, audio, video data Output: prediction model as manifest file Results as csv gzipped compressed file
Limitations	Training data: max 100GB Prediction input: max 1 TB Variable schema:1,000 max Longest run time job: 7 days
Bioinformatics Use	Genomics Analysis Classification

5. Amazon SageMaker

Type	Cloud Service
Open Source	Fee paying for heavy use
Availability	https://aws.amazon.com/sagemaker/
Description	SageMaker is a fully managed machine learning platform which intends to take away some of the heavy lifting previously involved with running models on AWS. SageMaker is a platform for authoring, training and deploying machine learning algorithms to business applications without provisioning infrastructure and managing and tuning training models.
Operating Systems	Cloud-based
Input and Output - formats/ types of data etc	Input: text (CSV, JSON, …) , audio, video data Output: prediction model as manifest file Results as csv gzipped compressed file
Limitations	10000 request per second Daily Data storage limit: 10GB Query timeout: 30 mins
Bioinformatics Use	Genomics Analysis Classification

6. Amazon's Deep Scalable Sparse Tensor Network Engine (DSSTNE)

Type	Library
Open Source	Open Source
Availability	https://github.com/amzn/amazon-dsstne
Description	The open source deep learning library, pronounced 'destiny', allows data scientists to train and deploy deep neural networks using GPUs.
Operating Systems	Cloud-based (Amazon AWS), Ubuntu (as a docker container)

Input and Output - formats/ types of data etc	Input: text, audio data Output: CSV
Limitations	-
Bioinformatics Use	Genome Classification

7. Apache Mahout

Type	Framework / Library
Open Source	Open Source
Availability	https://mahout.apache.org/ https://github.com/apache/mahout
Description	Apache Mahout(TM) is a distributed linear algebra framework and mathematically expressive Scala DSL designed to let mathematicians, statisticians, and data scientists quickly implement their own algorithms.
Operating Systems	Unix-based environment
Input and Output - formats/ types of data etc	Uses the Hadoop MapReduce framework. The MapReduce paradigm bases itself on two operations, **map** and **reduce**. The **map** operation takes a series of key/value pair and performs an operation on these. It emits zero or more key/value pairs. These pairs are sorted by key and fed to the reduce operation which iterates through the values of a certain key and performs some operation on them, again emitting zero or more new pairs.
Limitations	In order to perform any MapReduce processing of the bioinformatics data analyses, the analysis tasks must be modelled in a way that they can be implemented using the MapReduce framework and launched as individual tasks on a number of nodes in a cluster.
Bioinformatics Use	There is no specific bioinformatics analysis that has used Apache Mahout. But it can be easily used for bioinformatics data processing as it is an advanced distributed data processing library for machine learning and data mining and can process very large datasets.

8. Apache PredictionIO

Type	Server
Open Source	Open Source
Availability	https://predictionio.apache.org/
Description	It is built atop Spark and Hadoop, and serves Spark-powered predictions from data using customizable templates for common tasks. Apps send data to PredictionIO's event server to train a model, then query the engine for predictions based on the model. Spark, MLlib, HBase, Spray, and and Elasticsearch all come bundled with PredictionIO, and Apache offers supported SDKs for working in Java, PHP, Python, and Ruby.

Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	Recommender, Classification, Regression, NLP, Clustering, others Input: Json Output: e.g. a ranked list for recommender systems
Limitations	-
Bioinformatics Use	None. Recently developed tool (since 2015)

9. Apache Spark MLib

Type	Library
Open Source	Open Source
Availability	https://spark.apache.org/mllib/
Description	Apache Spark MLlib is an in-memory data processing framework. Spark offers a large and growing library of useful algorithms and utilities incorporating classification, regression, clustering, collaborative filtering and more (for in-memory data processing).
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	Inputs and outputs as per the specific analyses mentioned.
Limitations	Does not seem to have any limitation. It has been empirically validated that the libraries in Spark outperform a number of other libraries for most the analyses.
Bioinformatics Use	Can be used for: ● Alignment and mapping ● Assembly ● Sequence Analysis ● Phylogeny ● Drug Discovery ● Single-cell RNA sequencing ● Variant association and population genetics studies

10. Azure Machine Learning Studio

Type	Browser-based application
Open Source	Fee Paying
Availability	https://azure.microsoft.com/en-us/services/machine-learning-studio/
Description	A cloud service that enables you to build, deploy, and share predictive analytics solutions. Supports modelling in Python, Scala and PySpark.
Operating Systems	Cloud-based

Input and Output - formats/ types of data etc	Input: Delimited files such as CSV, TSV, etc; Fixed width files; Plain text files; Excel (.xls/xlsx) ; JSON files; Parquet files Output: .xls
Limitations	Rows and columns are each limited to the .NET limitation of Max Int: 2,147,483,647 Fewer algorithms (e.g. no XGBoost) and other transformations (e.g. NLP) built-in. Price Limited storage in the free version.
Bioinformatics Use	Tracking medical genetic literature through machine learning. (pubmed/27268407) A statistical approach to identify, monitor, and manage incomplete curated data sets (pubmed/29609549) A supervised machine learning classification model using the Azure Machine Learning (ML) Platform for analyzed medical literature (no ref)

11. Azure Machine Learning Service

Type	Cloud Service
Open Source	Fee Paying
Availability	https://azure.microsoft.com/en-us/services/machine-learning-service/
Description	Allows developers to manage and deploy machine learning workflows and models with the following modelling capabilities: ● Model versioning ● Model checking ● Deploying models to production ● Creating Docker containers with the models and testing them locally ● Automated model retraining ● Capturing model telemetry for actionable insights
Operating Systems	Cloud-based
Input and Output - formats/ types of data etc	Input: SQL,Blob storage, ARFF,SVMlight,TSV,CSV Output: .sql, blob, weburls
Limitations	Data upto 10GB maximum 1 hour continuous experimentation
Bioinformatics Use	See 10. Azure Machine Learning Studio

12. BigML

Type	Cloud Service
Open Source	Limited access when free / Fee-paying
Availability	https://bigml.com/releases
Description	BigML offers a wide variety of basic Machine Learning resources that can be composed together to solve complex Machine Learning tasks.

	BigML covers not only classification, regression and time series forecasting but also unsupervised learning tasks such as cluster analysis, anomaly detection, topic modeling, and association discovery BigML combines multiple ML approaches to arrive at better answer (“fusions”)
Operating Systems	Cloud-based
Input and Output - formats/ types of data etc	Documentation is under “Support” on here: https://support.bigml.com/hc/en-us
Limitations	-
Bioinformatics Use	-

13. Caffe2

Type	Deep learning framework
Open Source	Open source, under BSD-2 license
Availability	https://caffe2.ai/
Description	Caffe2 allows user to experiment with deep learning. Users can create applications to scale using the power of GPUs in the cloud or to the masses on mobile with Caffe2's cross-platform libraries. Caffe2 comes with Python & C++ APIs
Operating Systems	Linux, Mac OS, Microsoft Windows AWS Cloud Service and Docker
Input and Output - formats/ types of data etc	Input: Image data
Limitations	Not optimised for recurrent architectures, models need to be implemented in C++ (not always easy) (https://onlinelibrary.wiley.com/doi/full/10.15252/msb.20156651)
Bioinformatics Use	-

14. Cloudera Oryx2

Type	A framework for building real-time machine learning applications
Open Source	Open Source
Availability	https://github.com/OryxProject/oryx http://oryx.io/
Description	Oryx 2 is a realization of the lambda architecture built on Apache Spark and Apache Kafka, but with specialization for real-time large scale machine learning. It consists of three tiers, each of which builds on the one below: ● A generic lambda architecture tier, providing batch/speed/serving layers, which is not specific to machine learning ● A specialization on top providing ML abstractions for hyperparameter selection,

	etc. • An end-to-end implementation of the same standard ML algorithms as an application (ALS, random decision forests, k-means) on top.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	Input: text data
Limitations	-
Bioinformatics Use	Same as Apache Spark as it is built on top of Apache Spark. No specific bioinformatics analysis task using Cloudera Oryx.

15. ConvNetJS

Type	Library
Open Source	Open Source
Availability	https://cs.stanford.edu/people/karpathy/convnetjs/
Description	ConvNetJS is a Javascript library for training Deep Learning models (Neural Networks). It supports common Neural Network modules (fully connected layers, non-linearities), Classification (SVM/Softmax) and Regression (L2) cost functions. It provides the ability to specify and train Convolutional Networks that process images. It has an experimental Reinforcement Learning module, based on Deep Q Learning.
Operating Systems	Browser-based
Input and Output - formats/ types of data etc	Input: digital images Output: Json
Limitations	Slower processing times
Bioinformatics Use	Convolutional NN has been used in: diabetic retinopathy detection and heart disease detection

16. Deeplearning4j

Type	Deep learning library for Java
Open Source	Open Source
Availability	https://deeplearning4j.org/
Description	Deeplearning4j is a computing framework with wide support for deep learning algorithms. It includes implementations of the restricted Boltzmann machine, deep belief net, deep autoencoder, stacked denoising autoencoder and recursive neural tensor network, word2vec, doc2vec, and GloVe.
Operating Systems	Android, Linux, Mac OS, Microsoft Windows

Input and Output - formats/ types of data etc	Typically images
Limitations	Perception is that other (Python-based and more) libraries are used more frequently in the research community
Bioinformatics Use	Model prediction (for instance antibiotic resistance in bacteria) etc.

17. GoLearn

Type	Machine Learning Library for the Go language (by Google)
Open Source	Open Source
Availability	https://github.com/sjwhitworth/golearn
Description	It is still being developed. It can be used to perform classification using KNN | Trees | liblinear. It can be used to perform regression and filtering as well
Operating Systems	Linux, Mac OS X
Input and Output - formats/ types of data etc	Inputs and outputs will depend on specific algorithms being used.
Limitations	It is still under development and is a machine learning library for Google language Go (often referred to as Golang) .
Bioinformatics Use	No specific bioinformatics analysis task using GoLearn, but it can be used as it contains a number of machine learning algorithms.

18. GoML

Type	Library
Open Source	Open Source
Availability	https://github.com/cdipaolo/goml
Description	Generalized Machine Learning Libraries in Go Language Allows the average developer to include machine learning into their applications Models that have been implemented are Generalized Linear Models, perceptron, clustering and text classification
Operating Systems	FreeBSD (10.3 or later), Linux (2.6.23 or later with glibc), Mac OS (10.10 or later), Microsoft Windows (7, Server 2008R2 or later)
Input and Output - formats/ types of data etc	E.g. Clustering: Input: CSV Output: CSV, Json
Limitations	Rather recent library. Not well known yet among bioinformaticians, and limitations not reported

Bioinformatics Use	Rather recent library. Not well known yet among bioinformaticians, and limitations not reported

19. Google TensorFlow

Type	JavaScript Library
Open Source	Open Source
Availability	https://js.tensorflow.org/
Description	The machine learning library is developed for a multitude of tasks such as image search and improving its speech recognition algorithms. TensorFlow can produce C++ or Python graphs that can be processed on CPUs or GPUs. These flow graphs depict the movement of data running through a system. It targets environments with WebGL 1.0 or WebGL 2.0
Operating Systems	Browser-based
Input and Output - formats/ types of data etc	Input: supports various text format because data can be read in data structures like array, graph, etc, and image data
Limitations	Not suitable for data with small sample sizes such as identification genes or methylation probes associated with a given disease
Bioinformatics Use	Protein structure prediction; Medical imaging applications, such as segmenting brain images to help diagnose Alzheimer's disease; Determining correspondence between images; Predicting regulatory effects directly from DNA sequences. https://arxiv.org/ftp/arxiv/papers/1603/1603.06430.pdf https://doi.org/10.1016/j.cels.2016.01.009

20. H2O

Type	Software that can be called from the statistical package R, Python, and other environments
Open Source	Open Source
Availability	https://www.h2o.ai/
Description	It is used for exploring and analyzing datasets held in cloud computing systems and in the Apache Hadoop Distributed File System as well as in the conventional operating-systems Linux, macOS, and Microsoft Windows. It is written in Java, Python, and R. It has a graphical-user interface that is compatible with four browsers: Chrome, Safari, Firefox, and Internet Explorer.
Operating Systems	Linux (Ubuntu 12.04 ; RHEL/CentOS 6 or late), Mac OS X (10.9 or later), Microsoft Windows (7 or later)

Input and Output - formats/ types of data etc	Can be used to import, manipulate, and export data. It contains key machine-learning concepts, such as cross-validation and validation of data sets. It can also run on big-data systems, particularly Apache HDFS, several popular version: Cloudera (5.1 or later) and Hortonworks. It also operates on cloud computing environments, e.g. Amazon EC2, Google Compute Engine & Microsoft Azure.
Limitations	In order to run on the big-data platforms the bioinformatics analysis must be modelled in an appropriate way in order to leverage on the parallel and concurrency advantages of these platforms.
Bioinformatics Use	Contains the implementation of most of the ML algorithms which can be run on a cluster or cloud for optimization. There is no specific bioinformatics analysis that has used H2O. But it can be easily used for bioinformatics data processing as it supports distributed data processing for machine learning and can process very large datasets using cluster-based processing.

21. Hector

Type	Library
Open Source	Open source
Availability	https://github.com/xlvector/hector
Description	Hector is a Golang machine learning library. Currently, it can be used to solve binary classification problems.
Operating Systems	Run on command line, should work on all operating systems
Input and Output - formats/ types of data etc	Libsvm like data format
Limitations	Only supports algorithms which can solve binary classification problems https://godoc.org/github.com/xlvector/hector
Bioinformatics Use	Not explored much

22. IBM Watson Analytics

Type	Cloud Service
Open Source	Fee paying for heavy use
Availability	https://www.ibm.com/watson-analytics
Description	IBM Watson Analytics is a smart data analysis and visualization service on the cloud that helps users quickly discover patterns and meanings in their data. IBM Watson for Genomics provides access to expertly validated, up-to-date, and comprehensive content based on peer-reviewed publications, genomic databases, professional guidelines, clinical trials and approved therapies.
Operating Systems	Cloud-based

Input and Output - formats/ types of data etc	Data must be in list format (excel or csv format)
Limitations	Cannot process structured data directly
Bioinformatics Use	Has been used to: ● Provide enhanced insight into cancer kinases ● Drug Repurposing (IBM Watson:How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research - Chen et al) ● Uncover patterns that cause diseases by correlating data from genome sequencing to reams of medical journals, new studies and clinical records

23. Java-ML

Type	Library
Open Source	Freely available library
Availability	http://java-ml.sourceforge.net/
Description	Java API with a collection of machine learning algorithms (data manipulation, clustering, feature selection, classification, databases) implemented in Java.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	Text Data
Limitations	
Bioinformatics Use	Can be used for: • Gene clustering for gene family construction • Gene expression profile clustering • Core promoter clustering

24. JuliaML

Type	Library
Open Source	Open Source
Availability	https://github.com/JuliaML
Description	Julia ML is a set of various libraries to perform data mining at high speed using machine learning algorithm in Julia language.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	Depends on the library being used
Limitations	Julia dictionaries are hashed differently than Python dictionaries, which can make them slower in many cases. The data visualization packages are not as powerful as some other packages/libraries

Bioinformatics Use	A number of different libraries exist for bioinformatics use as follows: **EEG.jl** to process EEG files in Julia. **Rosalind.jl** provides a bioinformatics library for solving problems from rosalind.info. **taxize.jl** provides a taxonomic toolbelt for Julia. **COSMIC.jl** is a data analysis engine for COSMIC written in Julia. **Ensemble.jl** provide Ensemble Samplers for Julia. **FastaIO.jl** provides utilities to read/write FASTA format files in Julia. **GenomicTiles.jl gtf-parse-off** is used for experiments with parsing gene transfer format (GTF). **HyperNEAT.jl** provides a generative encoding for evolving ANN based on the NeuroEvolution of Augmented Topologies (NEAT) algorithm for evolutionary computation. **LCS.jl** is a package for finding longest common and longest contiguous subsequences. **OBC.jl** is used for the Optimal Bayesian classification for RNA-Seq data. **Pagel.jl** is used to detect correlated evolution on phylogenies. **Pathogen.jl** provides utilities to simulate and perform inference of disease dynamics. **ProgressiveAligner.jl** provides progressive alignment scripts for protein sequences. **PseudoGenomes.jl** is used to read alleles without a VCF parser. **SeqUtils.jl** provides seqencing analysis Utilities for Julia. **StatGenData.jl** is used for the statistical analysis of genomic data. **YARS.jl** provides YARS communication for RNA/proteins.

25. Keras

Type	Library
Open Source	Open Source
Availability	https://keras.io/
Description	Keras is a neural network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or MXNet. It is designed to enable fast experimentation with deep neural networks, and focuses on being user-friendly, modular, and extensible.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	Model - core data structure
Limitations	It is not always easy to customise Keras library for particular need, there is a need to understand what is going under the hood. It has limited functionality.
Bioinformatics Use	Has been used for predicting protein function from sequence and interactions https://academic.oup.com/bioinformatics/article/34/4/660/4265461

26. Massive Online Analysis (Java)

Type	Framework
Open Source	Open Source
Availability	moa.cms.waikato.ac.nz
Description	MOA is a framework for data stream mining. It includes a collection of machine learning algorithms (classification, regression, clustering, outlier detection, concept drift detection and recommender systems) and tools for evaluation.

Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	
Limitations	
Bioinformatics Use	Can be used for classification, regression, clustering, outlier detection, concept drift detection and recommender systems No specific bioinformatics analysis task using Massive Online Analysis

27. Matlab Statistics and Machine Learning Toolbox

Type	Application
Open Source	Fee Paying
Availability	https://www.mathworks.com/downloads/
Description	**Matlab** (*matrix laboratory*) is a multi-paradigm numerical computing environment. It can be used to compare approaches such as logistic regression, classification trees, support vector machines, ensemble methods, and deep learning. It uses model refinement and reduction techniques to create an accurate model that best captures the predictive power of your data. It integrates machine learning models into enterprise systems, clusters, and clouds, and target models to real-time embedded hardware.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	
Limitations	Code generation in Statistics and Machine Learning Toolbox does not support sparse matrices. Code generation does not support categorical arrays and tables. Statistics and ML Toolbox is still very new and requires a lot of improvement in diverse areas. The performance of the existing algorithms can be improved. More modern techniques should be added.
Bioinformatics Use	No specific bioinformatics analysis task using the Statistics and ML Toolbox

28. Microsoft Cognitive Toolkit

Type	Framework
Open Source	Open Source
Availability	https://www.microsoft.com/en-us/cognitive-toolkit/
Description	Microsoft Cognitive Toolkit, was previously known as Microsoft Computational Network Toolkit -0 (CNTK). It enables users to create neural networks depicted in directed graphs. Although primarily created for speech recognition technology, since April 2015 it has become a more general machine learning toolkit supporting image, text and RNN training (recurrent neural network - a type of neural network).

Operating Systems	Linux, Docker on Mac OS, Microsoft Windows and Cloud-based
Input and Output - formats/ types of data etc	
Limitations	Deeper models needs much data and memory Difficult to build - requires programming skills and deeper knowledge of ML Limited community support.
Bioinformatics Use	Supports Reinforcement learning, generative adversarial networks, supervised and unsupervised learning No specific bioinformatics analysis task using MIcrosoft Cognitive Toolkit - used mostly for image processing, speech and language processing

29. Microsoft Distributed Machine Learning Toolkit (DMTK)

Type	Framework
Open Source	Open Source
Availability	http://www.dmtk.io/document.html
Description	DMTK aims to ease crowded machine learning clusters, making it easier to run multiple (and differing) machine learning applications at the same time. It contains two distributed machine learning algorithms, which can be used to train the fastest and largest topic model and the largest word-embedding model in the world.
Operating Systems	Linux, Microsoft Windows
Input and Output - formats/ types of data etc	
Limitations	
Bioinformatics Use	No specific bioinformatics analysis task using DMTK

30. Mocha

Type	Library
Open Source	Open Source
Availability	https://github.com/pluskid/Mocha.jl
Description	It is inspired by the C++ framework Caffe. It is modular, has a high-level Interface, is portable and is known for its speed.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	Mocha uses the widely adopted HDF5 format to store both datasets and model snapshots, making it easy to inter-operate with Matlab, Python (numpy) and other existing computational tools.
Limitations	

Bioinformatics Use	Used for image classification - did not find any specific example for bioinformatics

31. MXNet

Type	Library
Open Source	Open source
Availability	https://mxnet.apache.org/
Description	MXNet is a powerful open-source deep learning framework. It is used to train, and deploy deep neural networks.
Operating Systems	Linux, Mac OS, Microsoft Windows, Cloud-based
Input and Output - formats/ types of data etc	Text and image data
Limitations	It has a much smaller community behind it compared with Tensorflow. It's not so popular among the research community.
Bioinformatics Use	Image classification tasks for: 1. Diabetic retinopathy detection 2. Diagnosis of heart disease

32. Octave

Type	Software/Programming Language
Open Source	Open Source
Availability	http://octave.sourceforge.net/
Description	GNU Octave is software featuring a high-level programming language, primarily intended for numerical computations. The LIBSVM, LIBLINEAR packages can be used for support vector machine/machine learning classification, regression, and distribution estimation problems.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	A large number of input formats Different output formats are produced
Limitations	Lacks an explicitly identified associative array, dictionary, map, mapping, or hash table data type although it has data structures like in C
Bioinformatics Use	Octave-bioinfo: **(Unmaintained package)** - This package contains bioinformatics manipulation for Octave. Toolbox available for optimisation and predictive analysis

33. Orange3

Type	Library + platform

Open Source	Open Source
Availability	http://blog.biolab.si/tag/orange3/
Description	Orange3 features a visual programming front-end for exploratory data analysis and interactive data visualization. It can also be used as a Python library
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	Feature , sample matrix (.xls, tab, space separated files etc)
Limitations	Install is big, limited list of algorithms (Orange - 2009) https://pdfs.semanticscholar.org/dd3c/89280a078131cd2bc1be6c3c6db2bc38c58f.pdf
Bioinformatics Use	Statistical patterns of epistatis in genetic studies https://www.sciencedirect.com/science/article/pii/S0022519305005217

34. PyBrain (Python)

Type	Library
Open Source	Open Source
Availability	http://pybrain.org/
Description	PyBrain, (short for Python-Based Reinforcement Learning, Artificial Intelligence and Neural Network Library), contains algorithms for neural networks, for reinforcement learning (and the combination of the two), for unsupervised learning, and evolution. Since most of the current problems deal with continuous state and action spaces, function approximators (like neural networks) must be used to cope with the large dimensionality. Our library is built around neural networks in the kernel and all of the training methods accept a neural network as the to-be-trained instance.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	Input data: CSV Pybrain provides specific data structure for each of the following categories of machine learning (supervised, sequential, classification)
Limitations	Slow performance
Bioinformatics Use	Pybrain has been used to predict protein-ligand binding sites. Used in a program called Silent Variant Analyzer (SilVA), a machine-learning approach to identify synonymous (silent) exonic mutations in a number of disorders

35. PyTorch

Type	Python Module
Open Source	Open Source
Availability	http://pytorch.org/
Description	PyTorch is a python package that provides two high-level features: Tensor computation (like numpy) with strong GPU acceleration and Deep Neural Networks built on a

	tape-based autodiff system.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	Multiple data types including lists, longTensor and many more
Limitations	It lacks interfaces for monitoring and visualization such as Tensorboard – though you can connect externally to Tensorboard. It lacks model-serving
Bioinformatics Use	Modelling systems https://openreview.net/forum?id=rJex9Mc0Y7

36. R

Type	Software/ Programming Language
Open Source	Open Source
Availability	https://cran.r-project.org/bin/windows/base/ https://www.rstudio.com/products/rstudio/download/
Description	R is a free software environment for statistical computing and graphics and has a large number of libraries available for ML.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	A large number of input formats Different output formats are produced
Limitations	When switching between different models one has to learn a new package written by a different author
Bioinformatics Use	Analysis of high-throughput genomic data Analysis of transcriptomics data

37. Rapid Miner (Java)

Type	Application
Open Source	Open Source
Availability	https://rapidminer.com/
Description	Data Science Platform providing a GUI and a Java API for developing your own applications. It provides data handling, visualization and modeling with machine learning algorithms. Currently used in the automotive industry, banking, insurance, healthcare, life sciences, telecommunications, manufacturing industry and many more
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	Supports more than 40 file types including SAS, ARFF, Stata, and via URLs. Has wizards for Microsoft Excel & Access, CSV, and database connections Can be used to access NoSQL databases like MongoDB and Cassandra

Limitations	Memory intensive and slows down your system. Not easy to use for new users
Bioinformatics Use	Has been used to estimate the midline shift and Intracranial pressure estimation based on Brain CT Images Used together with Taverna for assay clustering of microarray data and generating heatmaps

38. Rust bio

Type	Rust Bioinformatics Library
Open Source	Open Source
Availability	https://github.com/rust-bio/rust-bio
Description	Rust Bio provides implementations of many algorithms and data structures that are useful for bioinformatics. It has most major pattern matching algorithms and a convenient alphabet implementation.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	
Limitations	
Bioinformatics Use	Can be used for: pairwise alignment, suffix arrays, BWT and FM-Index, FMD-Index for finding supermaximal exact matches, a q-gram index, a rank/select data structure, FASTQ and FASTA and BED readers and writers, helper functions for combinatorics and dealing with log probabilities.

39. Rusty-Machine

Type	Library
Open Source	Open Source
Availability	https://crates.io/crates/rusty-machine/
Description	Rusty-machine is a general purpose machine learning library implemented entirely in Rust. It aims to combine speed and ease of use - without requiring a huge number of external dependencies. The goal of this library is to combine safety and speed without sacrificing simplicity.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	

Limitations	
Bioinformatics Use	No relevant Bioinformatics applications yet

40. Scikit-Learn

Type	Library
Open Source	Open Source
Availability	http://scikit-learn.org/stable/
Description	Sci-kit is a simple and efficient tools for R data mining and data analysis. It is accessible to everybody, and reusable in various contexts. It is built on NumPy, SciPy, and matplotlib. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, *k*-means and DBSCAN.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	any numeric data stored as numpy arrays or scipy sparse matrices and other types that are convertible to numeric arrays such as pandas DataFrame
Limitations	Can be challenging to start working with Python and sci-kit learn Not the best for building models. Not very efficient with GPU.
Bioinformatics Use	Pattern recognition tasks such as predicting and prioritizing putative functional interactions between proteins, learning regulatory modules and linking complex genotypes to phenotypes

41. Shogun

Type	Library
Open Source	Open Source
Availability	http://www.shogun-toolbox.org/
Description	Shogun is an open-source machine learning library that offers a wide range of efficient and unified machine learning methods. It supports many languages (Python, Octave, R, Java/Scala, Lua, C#, Ruby, etc) and integrates with their scientific computing environments. It can be used in the cloud from a browser.
Operating Systems	Linux/Unix, Mac OS, Microsoft Windows and Cloud-based
Input and Output - formats/ types of data etc	Shogun provides the capability to load datasets of different formats.
Limitations	
Bioinformatics Use	Has been used for: splice site recognition, gene prediction,genome annotation

42. Theano

Type	Deep Learning Python Module
Open Source	Open Source
Availability	http://deeplearning.net/software/theano/
Description	Theano allows users to define, optimize, and evaluate mathematical expressions involving multi-dimensional arrays efficiently.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	Typically, a graph that the user has to build
Limitations	Low-level framework, users have to be able to program the details of model Frequently long compile time when using large models, harder to learn https://www.sciencedirect.com/science/article/pii/S2405471216000107 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4965871/
Bioinformatics Use	

43. Veles

Type	Platform
Open Source	Open Source
Availability	https://velesnet.ml/
Description	Veles is Samsung's distributed deep learning platform, which is written in C++ and uses Python for coordination between nodes. Veles offers an API enabling immediate use of trained models and can be used for data analysis.
Operating Systems	Linux
Input and Output - formats/ types of data etc	Use their Loaders to preprocess data for use in veles
Limitations	Hard to say without running it, but seems very powerful and advanced. Could be worth trying out.
Bioinformatics Use	Useful for protein folding detection

44. Weka

Type	Software
Open Source	Open Source
Availability	https://svn.cms.waikato.ac.nz/svn/weka/

Description	Weka has a graphical interface as well as a Command line interface. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection.
Operating Systems	Linux, Mac OS, Microsoft Windows
Input and Output - formats/ types of data etc	Input: CSV, ARFF, Relational Table, Output: CSV
Limitations	Memory issues for very large data sets in the GUI version Cannot easily feed the results of one algorithm as input to another algorithm
Bioinformatics Use	Has been used for: Automated protein annotation Probe selection for gene-expression arrays Experiments with automatic cancer diagnosis Developing a computational model for frame-shifting sites Plant genotype discrimination Classifying gene expression profiles and extracting rules from them