Prospects and Challenges of Machine learning based Research in Bioinformatics: The Case for Africa

Ogechukwu N. Iloanusi*, Ijeoma J. F. Ezika, Kosisochukwu J. Madukwe, Ijeoma O. Okeke, Charles C. Osuagwu
Department of Electronic Engineering, University of Nigeria, Nsukka 410001, Enugu State. Nigeria.
{ogechukwu.iloanusi; ijeoma.ezika; kosisochukwu.madukwe; ijeoma.okeke; charles.osuagwu}@unn.edu.ng

Abstract

Bioinformatics, a field of nanobiotechnology, employs mathematical and statistical algorithms in studying biological data. Bioinformatics is also at the service of nanomedicine. Methods adopted in bioinformatics are typically software based and currently involve machine learning approaches which, when successful, are far more accurate in biological data analysis or prediction than traditional software-based methods. Considering a wide scope of diseases that could be combatted through machine learning-based bioinformatics research, the first goal of this paper is to highlight the prospects for research on resisting diseases peculiar to Nigeria and Africa at large, such as, improving the genetic resistance to malaria; eradicating diabetes disease, sickle cell anemia; and improving genetic resistance to HIV/AIDS. The second goal of this paper shows how machine learning can be applied in key areas of bioinformatics. Many researchers newly introduced to the field of bioinformatics often find the bioinformatics tools very confusing. As such, the second goal of this paper would help such researchers understand why, where and how to apply machine learning in bioinformatics research.

1 Introduction

Bioinformatics employs mathematical and statistical algorithms in studying biological data. Fields that benefit from bioinformatics include but are not limited to genetic engineering, genetics and genomes, gene ontologies and structural biology. Bioinformatics also involves image processing, feature extraction and representation of biological data. Bioinformatics encompasses all that have to do with identifying abnormal cells, such as cancerous cells amongst normal cells; transcribing and translating the genes of complex diseases; predicting protein structure from expressed genes, etc. Methods adopted in bioinformatics are typically software based and currently involve machine learning approaches which, when successful, are far more accurate in biological data analysis or prediction than traditional software-based methods.

Machine learning (ML) is an important field of artificial intelligence where it is believed that given a lot of data input, a machine can learn from its interactions with the data and eventually develop a model, which could classify or identify a new input, or predict an output for a given new input. The data used for "teaching" the system, technically referred to as "training" the system is called the training data and the result of the training is referred to as a trained model. A

model is developed by adequately training it with sufficient data till the model learns to categorize a new input correctly.

Typical ML algorithms such as regression, clustering and artificial neural networks (ANN) have been applied to the major fields of bioinformatics such as in genetic testing of microscopic spots stored in DNA microarrays; genomics – the analysis and mapping of genomes; proteomics – the analysis of proteins from gene expressions from some tissue, organism or cell. Medical or biological fields that benefit from the application of machine learning to bioinformatics include: medical engineering, epidemiology and the study of genetic diseases and disorders, such as, Alzheimer's disease, diabetes, cancer, arthritis, high blood pressure, hemochromatosis, cystic fibrosis, Huntington's disease, sickle cell anemia and Marfan syndrome [1], [2]. With respect to a wide scope of diseases that could be combatted through machine learning-based bioinformatics research, the first goal of this paper is to highlight the prospects for research on diseases common to Africans in Nigeria and Africa at large, such as, improving the genetic resistance to malaria; eradicating diabetes disease, sickle cell anemia; minimizing uterine fibroids in women and improving genetic resistance to HIV/AIDS. Uterine fibroids have been found to be more prevalent in black than in white women [3]. The second goal of this paper shows how machine and deep learning can be applied in key areas of bioinformatics. The reason for the second goal is to help many researchers newly introduced into the field of bioinformatics, who often find the bioinformatics tools very confusing, to understand the role of machine learning in bioinformatics research.

One of the challenges of machine learning based bioinformatics in Africa, is the limited availability of diverse and high volume biomedical data for accurate analysis [4]. Data is the crux of any analysis or learning method and its importance cannot be overemphasized. Currently, the Human Heredity & Health in Africa (H3Africa: www.h3africa.org) consortium is championing the research in bioinformatics in Africa, with the goal of generating and publicly publishing large genomics dataset of Africans [5]. On the other hand, due to the inadequate computing infrastructure and unreliable internet connection for cloud computation, there is the possibility of data export to the west for analysis [4], [6]. This eventually deprives researchers and students working in this area the hands-on experience and impedes the advancement of bioinformatics research in Africa.

2 Literature Review

ML has been used in different parts of the world for a variety of healthcare applications, including the identification of treatment sub-groups [7]; detection of anomalies in medical records [8]; predicting drug resistance [9]; improving patient understanding and profiling [10] and informing health policies [9]. Specifically, many research works have been carried out on the analysis and knowledge discovery of African datasets and African-related issues. In the area of disease diagnosis and prevention, malaria, sickle cell and fibroid have received considerable

attention and we highlight a few of these findings. In [11], the Random Forest ML algorithm was used to predict combination treatment for malaria that circumvents existing resistance. Because there is a large number of possible compound combinations, using a traditional data analysis method will require more time and cost, thus an ML technique, capable of effectively handling a large amount of data is used to predict new coactive compound pairs [11]. With images of unstained red blood cells as input dataset, Park et. al. [12] automated the analysis method for detection and staging of these quantitative phase images infected by the malaria parasite Plasmodium falciparum at trophozoite or schizont stage. They employed various machine learning techniques, including linear discriminant classification (LDC), logistic regression (LR), and k-nearest neighbor classification (NNC); to formulate algorithms that combine all the calculated physical parameters to distinguish cells more effectively. Still on malaria detection and diagnosis, Z. Liang et. al. [13] applied the infamous convolutional neural network (CNN) method to differentiate between infected and uninfected cells in thin blood smears. Other approaches have also been applied to both thin and thick blood smears [14]. Finally, Sharma et.al. [15] employed the Support Vector Machine (SVM) and Artificial Neural Network (ANN) techniques to develop a model for malaria outbreak prediction.

Another unique disease plaguing the African race in Sickle Cell Anemia. Beyond selection of reproductive sexual partners and other medical procedures for the selection of non-infected embryo, bioinformatics approaches have been investigated for the management and possible elimination of this disorder. The authors in [16] developed a model to automate the classification of red blood cell shapes for a better prognosis of the sickle cell disease. The convolutional neural network (CNN) technique was used in their classification task. The developed CNN model was able to spot the slight differences in texture alteration in the oxygenated and deoxygenated red blood cells. The achieved processing speed using the developed model is also impressive. Khalaf et al [17] investigated the accuracy and performance of several machine learning techniques to classify the dose of medication needed to treat patients with Sickle Cell disease. Their results show that the Random Forest Classifier had the highest overall performance.

With respect to malaria and sickle cell diseases, approaches existing in the literature so far have focused on detection, treatment and prognosis of these diseases already contracted. The goal of this paper is to incite the prevention of malaria disease, sickle cell disorder and other diseases common to Africans starting at the genetic level by controlling the expressions of the genes causing these diseases

3 Some Applications of Machine Learning in Bioinformatics

In this section, we will summarize the key classes of machine learning algorithms and application of results in select fields of bioinformatics. Machine learning algorithms are broadly classified into supervised and unsupervised learning. These classes of algorithms include: supervised learning, unsupervised learning, reinforcement learning, graphical models, optimization techniques and deep learning. We will simply summarize the two major classes.

3.1 Supervised learning

For a given set of data inputs, the goal is often to train a model that can classify a new dataset into a group of classes, c, based on a set of attributes or properties characterizing each data input. This classification task can also be extended to the case of predicting real-valued outputs, such that the number of classes, c, is infinite. For example, given a set of DNA sequences, each defined by the nucleotide at various positions, the goal may be to identify donor or acceptor splice sites [18]. To build a supervised learning classifier, we would require a training dataset of DNA sequences, each labelled as true donor site or false donor site[1]. Hence, given that the class (labels) of the training datasets are known prior to training, this method of building an ML classifier is referred to as supervised. Intuitively, this set of algorithms can only be employed in problems with known labels, such that the training set forms a set of input-output pairs or attributes-class pairs. The resulting classifier can then be used to classify new sequences (with unknown labels) as true or false donor site [19]. Supervised learning algorithms are either classification algorithms or regression algorithms. The former group refer to tasks with the number of classes are finite, while the latter refers to an infinite class task. Therefore, for regression algorithms, the associated tasks are referred to as prediction tasks rather than classification tasks.

Existing supervised learning algorithms include: logistic regression, Bayesian classifier, nearest-neighbor rule, neural networks, support vector machines, etc. These algorithms often serve as a black-box, receiving your training data as inputs and producing a classifier/predictor or classification/prediction model. However, parameter tuning/selection is important for producing an optimal model. It is also important to note that there is no best algorithm for the different problem sets in bioinformatics.

Typical examples of supervised learning tasks include: classification of erythrocytes infected with the malaria parasite [20], decision-tree based cancer cell classification from gene expression data [21], Bayesian network-based prediction of illness susceptibility [22], among others. Recently, a very important type of supervised learning method known as deep learning has attracted widespread interest and has generated insightful and novel results [13], [16], [23].

3.2 Unsupervised learning

This is applied in situations where training data labels do not exist. Basically, we have datasets with attributes, and the aim is to group these data based on similar attributes. For example, given the unlabeled data of Hepatitis B virus (HBV) DNA sequence from the GenBank database of the

[1] An example of publicly available Splice site dataset is the HS^{3}D dataset found at http://www.sci.unisannio.it/docenti/rampone/

National Center for Biotechnology Information (NCBI)[2], Bustamam et. al. [24] used the K-means clustering algorithm to group them into two clusters. The clusters differentiate more harmful HBV, containing increased protein structure to the less harmful HB viruses. A detailed review of the application of unsupervised learning to bioinformatics research can be found in [25].

4 Typical Machine Learning Workflow

In order to employ the described algorithms in the analysis of a given dataset, we outline a typical workflow, from data collection up to the prediction or classification of new unseen data. The workflow is also illustrated in Fig. 1 for easy reference.

4.1 Data Collection

Here the data to be analyzed are collected via reliable methods. Examples include: blood film images using a high resolution digital camera connected to a microscope [20], Colon adenocarcinoma specimens from patients in a hospital [26] and multiple health indices from children suffering from Sickle Cell Anemia [27].

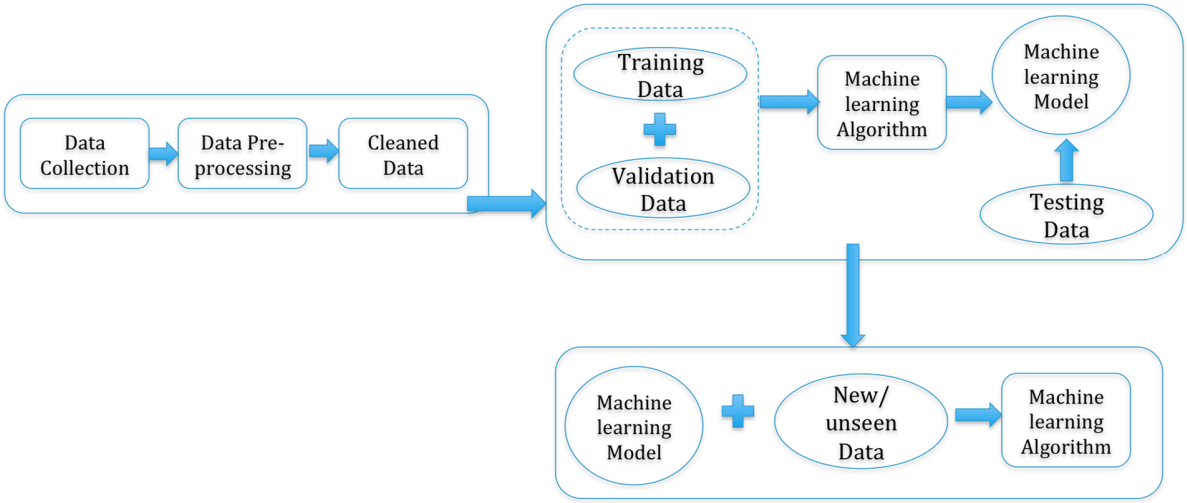

Figure 1: Typical Machine learning workflow

4.2 Data Pre-processing

Often times, the input data for computer analysis is characterized by additional noise emanating from human error, environment or acquisition device. It is therefore important to perform some preprocessing operations. In addition, some data come in non-numerical form and need to be digitized as computations performed by machine learning algorithms are on numerical data. A typical example can be found in [18], where splice sites in genes are converted from string (alphabetical) data to numerical data. Other pre-processing steps include: normalization [13],

[2] This database is located at www.ncbi.nlm.nih.gov

[24], multiple arithmetic scaling operations [26], discretization and binning [17], etc. In [20], image luminance and color correction using adaptive low pass filtering was performed on the input blood film images for improved performance. With respect to images, there is also need for image segmentation, i.e. extraction of useful information or a section of the image for the required analysis [20][28].

Furthermore, one preprocessing step that is important in most fields of machine learning is known as dimensionality reduction. Given that most data elements are defined by several attributes, it is important to perform some mathematical processing in other to reduce correlation between attributes and further reduce the size of the resulting dataset [25]. Because dimensionality reduction is a non-reversible operation that can affect accuracy results, selection of a subset of attributes using some statistical ranking algorithms in a process known as feature selection is also common practice [25]. It might also be necessary at this stage to visualize the data using different graph plots to get a better understanding of the data.

4.3 Algorithm Selection

One of the toughest decisions to make in any ML problem is the choice of algorithm for best performance. However, the good news is that there is no one-size-fits-all algorithm for a particular problem. This is even referred to as the "No Free Lunch" theorem in machine learning. Therefore, depending on the nature of your problem, the size of your dataset, the set of input features and choice of parameters for a given algorithm, different performance results can be obtained. An in-depth analysis of the effect of the choice of algorithm and set of hyper-parameters on bioinformatics data can be found in [29]. Most algorithms, as listed in Section 3, are available as library functions in software packages like Python, R, Matlab, etc and other ML softwares like Weka, Azure Machine Learning Studio, Google Cloud ML Engine, etc.

4.4 Model Training, Evaluation and Hyper-parameter tuning

This is the part where we plug our clean and pre-processed data into a chosen ML algorithm to produce a model for predicting/classifying new datasets. The first step in this stage is to split your data in three sets: training set, validation set and test set. This is important for performance analysis as training with entire dataset can result in "over-learning" your data and most likely performing poorly when faced with a new dataset. This is referred to as overfitting. The training set is used to train or build the model. That is, the algorithm "learns" from the properties of the training data. The validation set is used to test the trained model while tuning the various parameters of the algorithm in order to avoid overfitting. Both training and validation datasets are sometimes collectively referred to as the training set, as there are functions that can automatically validate during training[3]. Finally, the test set is used for model evaluation after training and producing performance reports of the trained model. The ratio of training set to test

[3] A typical example is the k-fold cross validation technique.

set is left to the discretion of the user. However, typical values of 80:20 are very common [17]. It is also important to note that some ML algorithms do not divide data into training and testing set, as their aim is simply to discover latent characteristics for dataset transformation [30]. This is true for some unsupervised learning algorithms like Principal Component Analysis (PCA).

4.5 Performance analysis

For supervised learning based ML methods, the classification or prediction accuracy is often the measurement criteria for choosing the best predictor model [20]. Depending on the nature of the problem, other performance measures can be employed. They include: sensitivity, specificity, precision, F1 score, receiver operating characteristics (ROC), area under ROC curve (AUC) etc [15], [17], [20].

4.6 Predict

The end result of all these is simply to predict or classify or assign to a given cluster some new unseen dataset. It is expected that if training and testing are done appropriately, the new dataset can be predicted with approximately the same accuracy/sensitivity achieved with the test set.

5 Key Problems of ML-based Research in Africa:

The introduction of bioinformatics in Africa has been met with several challenges, resulting in a slow growth rate. These problems include inadequate infrastructure and training opportunities, research funding, availability of bio-repositories, human resources, etc. Some of these problems are discussed below:

5.1 Lack of Expertise

For a successful application of machine learning to bioinformatics, one requires an in-depth understanding of both biology and mathematical analysis. It could be in form of machine learning experts learning the rudiments of bioinformatics in order to apply their knowledge or bioinformatics experts leveraging the advantages of machine learning to solve problems native to them. The lack of trained bioinformatics experts in Africa with a working knowledge of machine learning techniques and vice versa directly affects the availability of expertise in research centers. This has a negative effect on the initiation and completion of projects in African Institutions. The inadequate supply of people with the technical know-how is also as a result of the lack of appropriate national policies and strategies that foster bioinformatics education and research. Additionally, the unavailability of adequate training and capacity-building opportunities adversely affects the supply of relevant experts in the continent [31]–[34]. For a field like bioinformatics that is constantly evolving, open access to publications is pertinent for extensive research work. Some developing countries in the continent may not have access to

journals with high subscription fees [35] . Sadly, local survey have also identified the lack of collaboration by the few experts in the field [36].

5.2 Lack of Infrastructure

Reliable Internet access is a recurring problem in some African countries. For machine learning to be applied to bioinformatics, there has to be steady internet connectivity as some of the available tools can only be accessed online. In addition, some databases and learning materials are available online and can only be retrieved when connected. An effort called The eBioKits initiative [37] has been established to help solve the problem of unstable inter connectivity by housing biological databases and bioinformatics applications on a low-cost server. Some of the databases available on this platform are Ensembl, NCBI-BLAST, EMBOSS, PLINK [32]. Additionally, bioinformatics data is very sensitive, therefore a secure and stable internet connectivity is required for its transfer.

Constant supply of electricity is also another pertinent issue. Training machine learning algorithms take some time and thus requires the necessary hardware to be constantly powered through the duration of the training [38]. Machine learning applied to any field including bioinformatics requires powerful dedicated computing and storage hardware [32], [39]. There is a lack of laboratories housing said computers for training and testing data on developed algorithms.

5.3 Lack of Funding

Adequate funding is required from both public and private sectors as international funding might not be enough. The percentage of the gross domestic product (GDP) allocated to research and development is quite low, with Ghana reporting a 0.3% and South Africa a 0.87% [38]. Funding specific for bioinformatics research is necessary as there are other research endeavors that might compete for generic allocation. Some countries like South Africa has made great effort to establish funding for their bioinformatics research through their South African National Bioinformatics Institute (SANBI) [32].

5.4 Paucity of Research Data and Stigma Associated with Data Collection

Machine learning techniques require a large amount of data for the training and testing set. When applied to bioinformatics, this data could be clinical, demographic or genomic [40]. It is vital that this data is complete and represents a good number of the population, for the research results to be representative. The genome sequencer is not widely available thus restricting the illustrative nature of the outcome.

Phenotype data is a type of clinical data that needs to be collected from patients during consultations. This may prove difficult in some areas in the African continent that have not

embraced Western medicine or have superstitious/religious beliefs that keep them from engaging in such activities. There is also the issue of the protocol involved in getting ethical clearance from the appropriate body before collecting genomic data. Every country has her own process that must be followed before commencing research work [41]. This process might take longer than necessary if not considered a priority by these bodies, thus slowing down research activities and discouraging the researchers. Furthermore, some of the available data are characterized as low quality, due to the over reliance on traditional collection means such as questionnaires which need to be manually completed rather than digital means such as using electronic surveys or electronic apps which are paired with biosensors on smart devices. Moreover, these traditional methods are more prone to error and can sometimes be difficult to analyze due to incorrect completions.

6 PROSPECTS OF MACHINE LEARNING FOR BIOINFORMATICS IN AFRICA

The potentials of ML for improving various aspects of living and providing a plausible means of data-driven decision making makes it an attractive tool worth exploring. There is a growing concern by humans who feel that they may be replaced by machines and are therefore unwilling to embrace the technology; however, these machines have been shown to exceed human ability in very critical application areas and can be used as commentary tools to enhance research and healthcare delivery in Africa. We discuss a few of the prospects of this area of research in Africa.

6.1 Possible Cure of Some Diseases

Machine learning speeds up processes to which it is applied. It can, when applied correctly, reduce the time to achieve research results. This can also increase the rate of discovery for new drugs and therefore lead to the cure or eradication of some diseases.

6.2 Increase in Career Opportunities

The increase in the generation of biological data calls for professionals that can analyze and extract valuable insight from it. This creates new job opportunities in the field of bioinformatics. With the correct systems in place, educational institutions and research centers will start providing adequately trained professionals to fill those positions.

6.3 Growth of Native Databases

Since genes and other biological information vary for people from different races, it is important for researchers in Africa to have access to data native to the African people. In time, as Machine Learning is applied to bioinformatics, there will be a need to have databases containing such data. Also, diseases like malaria and sickle cell will receive more tailored attention from indigenous researchers. The growth of native databases would also help to reduce any bias that

could be associated with machine learning models developed using only data from a particular kind of race.

7 Conclusion

Different research areas have bought into the ML wave due to its promising future and solutions. It is therefore pertinent that even in the face of the continental challenges facing Africans, there is need for current researchers in the field to build stronger networks to be able to attract international grants and provide possible solutions for the problems plaguing us. Africa has a genetically diverse population, amounting to huge potentials for genetic health research. The development of ML models would on the long run, reduce experimental/research cost and help overcome the limitations of advancing research in Africa due to lack of physical laboratory equipment. Models can be easily trained to replace equipment-based diagnosis and save time by avoiding repeated experiments. So basically, we transition from analyzing data using traditional means to exploring available data and using it to create models that could be used to complement or improve health service delivery in Africa. Finally, the proliferation of nano-sensors and embedded sensors in smart devices are enablers for fast digital data harvesting.

References

[1] S. T. Chou *et al.*, "Whole-exome sequencing for RH genotyping and alloimmunization risk in children with sickle cell anemia," *Blood Adv.*, vol. 1, no. 18, pp. 1414–1422, 2017.

[2] A. J. Nevado-Holgado and S. Lovestone, "Determining the Molecular Pathways Underlying the Protective Effect of Non-Steroidal Anti-Inflammatory Drugs for Alzheimer's Disease: A Bioinformatics Approach," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 1–7, 2017.

[3] S. K. Laughlin, J. C. Schroeder, and D. D. Baird, "New directions in the epidemiology of uterine fibroids," *Semin. Reprod. Med.*, vol. 28, no. 3, pp. 204–217, 2010.

[4] N. Mulder *et al.*, "Genomic Research Data Generation, Analysis and Sharing – Challenges in the African Setting," *Data Sci. J.*, vol. 16, 2017.

[5] C. Rotimi *et al.*, "Research capacity. Enabling the genomic revolution in Africa," *Science*, vol. 344, no. 6190, pp. 1346–1348, 2014.

[6] L. Nordling, "African scientists call for more control of their continent's genomic data," *April, 2018*. [Online]. Available: https://www.nature.com/articles/d41586-018-04685-1. [Accessed: 12-Jul-2018].

[7] T. Chen *et al.*, "Using model-based recursive partitioning for treatment-subgroup interactions detection in real-world data: A myocardial infarction case study," *Stud. Health Technol. Inform.*, vol. 247, pp. 576–580, 2018.

[8] N. Sun, M. Xu, M. Cai, X. Ma, and Y. Qin, "Clinical Similarity B ased Framework for Hospital Medical Supplies Utilization Anomaly Detection : A Case Study," vol. 0.

[9] J. Wiens and E. S. Shenoy, "Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology," *Clin. Infect. Dis.*, vol. 66, no. 1, pp. 149–153, 2018.

[10] P.-Y. Hsueh and S. Das, "Interpretable Clustering for Prototypical Patient Understanding: A Case Study of Hypertension and Depression Subgroup Behavioral Profiling in National Health and Nutrition Examination Survey Data," in *American Medical Informatics Association Symposium*, 2017.

[11] Y. Kalantarmotamedi, R. T. Eastman, R. Guha, and A. Bender, "A systematic and prospectively validated approach for identifying synergistic drug combinations against malaria," *Malar. J.*, vol. 17, no. 1, 2018.

[12] H. S. Park, M. T. Rinehart, K. A. Walzer, J. T. Ashley Chi, and A. Wax, "Automated Detection of P. falciparum using machine learning algorithms with quantitative phase images of unstained cells," *PLoS One*, vol. 11, no. 9, 2016.

[13] Z. Liang *et al.*, "CNN-based image analysis for malaria diagnosis," *2016 IEEE Int. Conf. Bioinforma. Biomed.*, pp. 493–496, 2016.

[14] M. Poostchi, K. Silamut, R. J. Maude, and S. Jaeger, "Image analysis and machine learning for detecting malaria," *Transl. Res.*, pp. 1–20, 2016.

[15] V. Sharma, A. Kumar, L. Panat, G. Karajkhede, and A. Lele, "Malaria Outbreak Prediction Model Using Machine Learning," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 4, no. 12, pp. 4415–4419, 2015.

[16] M. Xu, D. P. Papageorgiou, S. Z. Abidi, M. Dao, H. Zhao, and G. E. Karniadakis, "A deep convolutional neural network for classification of red blood cells in sickle cell anemia," *PLoS Comput. Biol.*, vol. 13, no. 10, pp. 1–16, 2017.

[17] M. Khalaf *et al.*, "Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models," *Neurocomputing*, no. September, pp. 0–1, 2016.

[18] P. K. Meher, T. K. Sahu, A. R. Rao, and S. D. Wahi, "Identification of donor splice sites using support vector machine: A computational approach based on positional, compositional and dependency features," *Algorithms Mol. Biol.*, vol. 11, no. 1, 2016.

[19] P. Larranaga *et al.*, "Machine learning in bioinformatics," *Briefings in Informatics*, vol. 7, no. 1, pp. 86–112, 2005.

[20] G. Díaz, F. A. González, and E. Romero, "A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images," *J. Biomed. Inform.*, vol. 42, no. 2, pp. 296–307, 2009.

[21] J.-Y. Yeh, Tai-ShiWu, M.-C. Wu, and D.-M. Chang, "Applying Data Mining Techniques for Cancer Classification from Gene Expression Data," 2007.

[22] P. Sebastiani, M. F. Ramoni, V. Nolan, C. T. Baldwin, and M. H. Steinberg, "Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia," *Nat. Genet.*, vol. 37, no. 4, pp. 435–440, 2005.

[23] S. Min *et al.*, "Deep learning in bioinformatics," *Brief. Bioinform.*, vol. 31, no. 3, p. bbw068, 2016.

[24] A. Bustamam, H. Tasman, N. Yuniarti, Frisca, and I. Mursidah, "Application of k-means clustering algorithm in grouping the DNA sequences of hepatitis B virus (HBV)," *AIP Conf. Proc.*, vol. 1862, 2017.

[25] A. Serra, P. Galdi, and R. Tagliaferri, "Machine learning for bioinformatics and neuroimaging," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 2018.

[26] U. Alon *et al.*, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci.*, vol. 96, no. 12, pp. 6745–6750, 1999.

[27] C. Allayous, S. Clémençon, B. Diagne, R. Emilion, and T. Marianne, "Machine Learning Algorithms for Predicting Severe Crises of Sickle Cell Disease," pp. 1–11, 2008.

[28] L. P. Coelho, E. Glory-Afshar, J. Kangas, S. Quinn, A. Shariff, and R. F. Murphy, "Principles of bioimage informatics: Focus on machine learning of cell patterns," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6004 LNBI, pp. 8–18, 2010.

[29] R. S. Olson, W. La Cava, Z. Mustahsan, A. Varik, and J. H. Moore, "Data-driven Advice for Applying Machine Learning to Bioinformatics Problems," 2017.

[30] A. C. Müller and S. Guido, *Introduction to machine learning with Python : a guide for data scientists*. 2016.

[31] O. Nashiru, C. Huynh, S. Doumbia, J. C. Kissinger, R. D. Isokpehi, and O. O.

Olorunsogo, “Building bioinformatics capacity in West Africa,” *African J. Med. Med. Sci.*, vol. 36 Suppl, pp. 15–18, 2007.

[32] Ö. T. Bishop *et al.*, “Bioinformatics Education-- Perspectives and Challenges out of Africa.,” *Brief. Bioinform.*, vol. 16, no. 2, pp. 355–364, 2015.

[33] T. K. Karikari and J. Aleksic, “Neurogenomics: An opportunity to integrate neuroscience, genomics and bioinformatics research in Africa,” *Appl. Transl. Genomics*, vol. 5, pp. 3–10, 2015.

[34] I. Merelli, H. Pérez-Sánchez, S. Gesing, and D. D’Agostino, “Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives,” *Biomed Res. Int.*, vol. 2014, 2014.

[35] A. Djikeng, S. Ommeh, S. Sangura, I. Njaci, and M. Ngara, “Genomics and Potential Downstream Applications in the Developing World,” pp. 335–356, 2012.

[36] O. O. Ojo and M. Omabe, “Incorporating bioinformatics into biological science education in Nigeria: prospects and challenges.,” *Infect. Genet. Evol.*, vol. 11, no. 4, pp. 784–787, 2011.

[37] “EBioKit.” [Online]. Available: http://www.ebiokit.eu/. [Accessed: 14-Jul-2018].

[38] T. K. Karikari, “Bioinformatics in Africa: The Rise of Ghana?,” *PLoS Comput. Biol.*, vol. 11, no. 9, 2015.

[39] S. Ranganathan, “Bioinformatics Education-Perspectives and Challenges,” *PLoS Comput. Biol.*, vol. 1, no. 6, 2005.

[40] N. J. Mulder *et al.*, “Development of Bioinformatics Infrastructure for Genomics Research,” *Glob. Heart*, vol. 12, no. 2, pp. 91–98, 2017.

[41] R. M., D. V. J., S. H., and N. S.A., “Ethical issues in genomic research on the African continent: Experiences and challenges to ethics review committees,” *Hum. Genomics*, vol. 8, no. 1, p. no pagination, 2014.