REVIEW OF MACHINE LEARNING APPLICATIONS AND DATASETS IN CLASSIFICATION OF ACUTE LEUKEMIA

Jaishree Ranganathan

Department of Computer Science,
Middle Tennessee State University, Murfreesboro,TN, USA

ABSTRACT

Cancer is an extremely heterogenous disease. Leukemia is a cancer of the white blood cells and some other cell types. Diagnosing leukemia is laborious in a multitude of areas including heamatology. Machine Learning (ML) is the branch of Artificial Intelligence. There is an emerging trend in ML models for data classification. This review aimed to describe the literature of ML in the classification of datasets for acute leukemia. In addition to describing the existing literature, this work aims to identify different sources of publicly available data that could be utilised for research and development of intelligent machine learning applications for classification. To best of the knowledge there is no such work that contributes such information to the research community.

KEYWORDS

Machine Learning, Oncology, Data Repository, Leukemia, Cancer.

1. INTRODUCTION

Artificial Intelligence (AI) is the field of computer science that focuses on intelligent systems that perceive the environment and takes action to maximize the chance of achieving the goal [1] [2]. Machine learning (ML) is the subfield of AI that uses statistical algorithms. Deep learning (DL) is subfield of ML utilizing artificial neural networks, inspired by the way the human brain processes information. These methods have revolutionized the fields of image classification, speech recognition, and many other domains [3]. Utilizing the power of artificial intelligence (AI), with its sub-disciplines machine learning and deep learning is inherent in the era of big health care data [4][5]. It is estimated that AI in health care would see a compound annual growth rate of 50.2 percent. By 2025 approximately 36.1 billion dollars would be invested in AI-related health products [6].

Data is no more a single entity, meaning different sources contribute to the final knowledge. For instance, a physician might consider a patient's personal history, genomic sequences, prior medications, treatments, and hospitalization information for evaluation. Electronic health records (EHR) provide massive amounts of data. EHR's contain quantitative data, qualitative data, transactional data that could help guide clinical decisions [5]. Deriving insights from all potentially connected sources could be an overwhelming task for humans [7].

Artificial Intelligence (AI) plays a major role in many of the specialty fields in medicine including radiology, dermatology, ophthalmology, cardiology, and pathology [8]. Many AI-based medical devices and algorithms are being approved by the United States Food and Drug Administration (FDA) [9]. Cancer falls under extremely heterogeneous disease. Thus, cancer research continues to be on top of the list for the research community. Artificial intelligence has been used in assessing the degree of aggressive activity of cancers to predict the course of the disease and prognosis. It also provides potential guidelines to determine modes of treatment such as immunotherapy, chemotherapy, radiotherapy. Artificial intelligence, especially deep learning has been at the forefront of cancer image analysis, cancer genomics [10].

Leukemia is the cancer of the early blood forming cells. Philadelphia (Ph) Chromosome is a chromosomal abnormality when chromosome 9 breaks off and bonds to a section of chromosome 22. This break can affect the tumor suppressor genes. This change is sometimes one of the causal factors of Acute Myeloid Leukemia (AML), Chronic Myeloid Leukemia (CML), Acute Lymphoid Leukemia (ALL).

Haematology is the study of the physiology of the blood. Haematology is the most important component in Leukemia diagnosis. When performed manually by experts, it is a time consuming and labor-intensive process. Leukemia is a life-threatening cancer disease. There is no tolerance for errors. Automated intelligent processing systems are the critical applications in such scenarios [11].

This study focuses on the review of literature pertaining to the use of machine learning and deep learning models for classification in datasets for acute myeloid and lymphoid leukemia. There are not many publicly available datasets for research. This study aims to identify available data sources that can help in the research for developing classification models, training, and validation.

The rest of this paper is organized as follows section 2 - literature of ML models using acute lymphoid leukemia data, section 3 - literature of ML models using acute myeloid leukemia data, section 4 – datasets, section 5 - discussion and finally conclusion and references.

2. LITERATURE OF ML USING ACUTE LYMPHOID LEUKEMIA DATA

Acute Lymphoid Leukemia (ALL) is the type of cancer triggered by immature lymphocytes in the bone marrow. Most of the studies in the literature used blood smear images [12][13][14][15][16][17][18][19][20][27] and some of them use bone marrow samples [21][48]. The major focus of the studies is to build a model to classify healthy cell vs cancerous (leukemia) cells [12][19][20]. However, some of them further classify the leukemia cells into subtypes [21][13][14]. The subtypes are based on FAB (French, American, and British) classification. The Leukemia experts divided it into three subtypes (L1, L2, L3) based on the structure of cells. There is a mix of traditional machine learning classifiers and neural network models. Table 1. gives an overview of the studies in the literature for machine learning models for acute lymphoid leukemia data classification.

2.1. Deep Learning Classifiers

Only a few of the studies employ a neural network model, especially a convolutional neural network (CNN) [12][13][20][21]. The CNN models use pre-trained AlexNet. Achieved an accuracy of 96.06% - 97.78% for subtype classification and approx. 94% accuracy to classify between healthy and leukemic cells.

2.2. Traditional Machine Learning Classifiers

The majority of the papers use Support Vector Machine (SVM) for classification. They achieve accuracy in the range of 74% to 97%.

Table 1. ML Studies based on Acute Lymphoid Leukemia.

Reference	**Type of Data**	**Pre-Processing**	**Classification Model**	**Accuracy**
Di Rubertoet.al. [12]	Blood Smear Images	Hue Saturation Value; Blob detection; Segmentation watershed	Convolutional NeuralNetwork (CNN)	94.1% - Leukemia classification
Shafique andTehsin [13]	Blood Smear Images	Data Augmentation	Transfer Learning – DeepCNN	99.50% for normal vs cancerous cells;96.06 % for subtypes
Wang et. al.[48]	Bone marrow samples	Feature selection	Decision tree; Naïve Bayes; Support VectorMachine	Explained in terms of percentage [48]
Rawat et. al.[14]	Blood Smear Images	Segmentation; Feature Extraction(PCA)	Hybrid HierarchicalClassifiers	Overall classification accuracy97.6%
Laosai et. al.[27]	Blood smear images	Segmentation; Feature extraction	Support Vector Machine	92%
Bigorra et. al.[15]	Blood Smear Images	Segmentation - Spatial Kernel fuzzy c-means; Feature extraction(PCA)	Support Vector Machine	~ 74% forLBC
Rawat et. al.[16]	Blood Smear Images	Segmentation	Support Vector Machine	72% - 86.7%
Reta et. al. [17]	Blood Smear Images	Segmentation	Multiclass classifier	ALL: ~94%
Umamaheswari& Geetha [18]	Blood Smear Images	Segmentation; Feature Extraction	K-Nearest Neighbour	96.25%
Putzu et. al.[19]	Blood Smear Images	Segmentation; Feature Extraction	Support Vector Machine	93%
Prellberg and Kramer [20]	Blood smear images	Image flipping	ResNeXt CNN	F1 score – 88.91%
Rehman et.al.[21]	Bone Marrow Images	Segmentation	CNN	97.78%

Table 2. ML Studies based on Acute Myeloid Leukemia.

Reference	Type of Data	Pre-Processing	Classification Model	Accuracy
ko et.al. [22]	MulticolorFlow Cytometry	NA	Support Vector Machine	84.6% - 92.4%
Warnat-Herresthalet.al. [23]	Gene ExpressionData	NA	L1 – Regularized Logistic Regression; NeuralNetworks	95% - 97%
Kazemi et. al.[24]	Blood Microscopic Images	Image processing segmentation – kmeans	Support Vector Machine	10-fold cross validation
Gal et. al.[25]	Gene ExpressionData	Feature Extraction(PCA)	KNN; SVM; RandomForest	AUC 0.73 – 0.84
Agaian et. al.[26]	Blood Images	CIELAB color features; segmentation; feature extraction	Support Vector Machine	98%
Laosai et. al.[27]	Blood smearimages	Segmentation; Feature extraction	Support Vector Machine	92%
Dundar et. al.[28]	Flow Cytometry	Dirichletprocess gaussian mixture model	Non-parametric Bayesian algorithm	AUC 0.99 and 1.00
Manninen et.al. [29]	Flow Cytometry	Feature Generation	LR LASSO; LDA	100%
Biehl et. al. [30]	Flow cytometry	Feature vectors	GMLVQ	AUC 1.0
Matek et.al.[31]	Blood MicroscopicImages	Digitised usingoil immersion	Convolutional NeuralNetwork	Precision and Sensitivity94%

3. LITERATURE OF ML USING ACUTE MYELOID LEUKEMIA DATA

Acute Myeloid Leukemia (AML) starts in the bone marrow and quickly moves into the blood. Itis also denoted as acute myelocytic leukemia, acute myelogenous leukemia, acute granulocytic leukemia, and acute non-lymphocytic leukemia. The details of studies included in the literature is shown in Table 2.

Studies included in the literature use a multitude of data including flow cytometry, gene expression data, blood microscopic images for classification of AML as shown in Table 2. As in the case of ALL, majority of the studies use support vector machine (SVM) for classification of AML. According to the literature we could see that accuracy of the algorithms are in the range of 84% - 98% and AUC in the range of 0.73 to 1.00. Similar to the studies of ALL, segmentation is the major pre-processing step performed along with feature extraction using principal component analysis and feature generation.

4. DATASETS

This study identifies some of the publicly available datasets based on the literature and provides a high-level overview. This study discusses the below datasets for use in machine learning models and not for other specific purposes.

4.1. Acute Lymphoblastic Leukemia – Image Dataset (ALL-IDB)

ALL-IDB [32] is a publicly available dataset of microscopic images of blood samples in jpg format. The images are 24-bit color depth, and with a resolution 2592 x 1944. This dataset consists of two versions namely ALL-IDB1 and ALL-IDB2 [43][44][45].

ALL-IDB1: This specific set of images allow testing of both the segmentation and classification capability of the algorithms.

ALL-IDB2: This dataset is designed for classification algorithms.

4.2. BioGPS – Dataset Library

BioGPS [33] is a gene annotation portal. Such information could be utilized depending on the specific area of research. Structural comparison of protein binding is important in drug design. Gene information could help understand the protein structure and binding sites [38].

4.3. SMC – Blood Image Dataset (IDB)

This blood image dataset is available as part of the work proposed in [34] in 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC). The dataset is made available via MATLAB central.

4.4. Surveillance, Epidemiology, and End Results Program (SEER) Database

The SEER database (National Cancer Institute [46]): data could be used for overall analysis based on various features including age, geography. A lot of external factors could also be the reason for cancer. Pediatric Acute lymphoblastic leukemia (ALL) is analyzed in [41] to understand the widely affected population, and other associated factors.

4.5. AML Morphology Dataset

This dataset consists of expert-labeled single-cell images. The images are taken from peripheral blood smears. The dataset contains blood smears of 100 patients diagnosed with Acute Myeloid Leukemia, and 100 patients without signs of any hematological malignancy [31][39][40].

4.6. National Center for Biotechnology Gene Expression Omnibus

Gene expression omnibus database [35][37][50][51] is supported by National center for Biotechnology Information (NCBI) [49]. This database provides access to gene expression and other functional genomics data sets. This database offers web-based tools that helps to locate data relevant to specific interests, and in visualization of the data [36]. Gene expression data could be useful as gene mutation is one of the major factors in cancer patients. For instance [42] use gene expressions cancer data for feature selection process.

5. DISCUSSION

The literature discussed in section 2 focused on the ML classification models for ALL subtypes (L1, L2, L3) or generic classification of healthy cells vs cancerous cells. Deep learning is at the forefront of image classification. Limited research in the literature uses convolutional neural networks for ALL classification. It is also important to note that majority of the studies utilize the ALL-IDB dataset [32] classified by expert oncologists. Some of the studies use data from clinical settings for training and classification. These data are collected from clinical center or a minimal number of patients. It is necessary for more diverse datasets for such intelligent applications development and testing. The machine learning models yielded high accuracy. Especially in this scenario, the dataset is not diverse and is based on minimal data. Same applies for the studies included in section 3 for acute myeloid leukemia (AML).

Studies in the literature use supervised learning models. These models can be overlong to train especially in the event of big data. Also, another major disadvantage of supervised classificationis the labels require domain expertise. In the context of health care, the reason behind switching to automatic intelligent machines is to reduce manual labor and assist the existing process. Unsupervised learning models discover the inherent structure of unlabeled data by learning on their own from the data. However, some human intervention is required in the process for validating the output when using unsupervised models. Deep learning algorithms might allow ability to develop more accurate intelligent systems because of their intrinsic nature to learn data. Use of unsupervised learning models and deep learning algorithms could help in development of more intrinsic systems [47].

6. CONCLUSIONS

This study provides a review of literature that applied or utilised machine learning models for leukemia classification. The study also discusses multiple publicly available data sources for machine learning research. It is highly difficult to find datasets for machine learning especially health care data. Most of the studies used their own data collected from a clinical or laboratory setting. To the best of the knowledge there is no such paper that provides a review including the publicly available datasets. Based on the review of the existing literature it is evident that the existing studies only focus on limited data for classification. Thus, future studies could consider utilizing diverse data. As discussed in section 5, utilizing more of deep learning algorithms could prove to be meaningful and efficient in the healthcare settings and yield better results.

REFERENCES

[1] J. McCarthy, “What is artificial intelligence?,” 2007.
[2] Wikipedia, “Artificial intelligence,” 2021.[Online; accessed 25-september-2021].
[3] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” nature, vol. 521, no. 7553, pp. 436–444,2015.
[4] A. L. Beam and I. S. Kohane, “Big data and machine learning in health care,” Jama, vol. 319,no. 13, pp. 1317–1318, 2018.
[5] T. B. Murdoch and A. S. Detsky, “The inevitable application of big data to health care,” Jama, vol. 309, no. 13, pp. 1351–1352, 2013.
[6] J. Bresnick, “Artificial intelligence in healthcare spending to hit $36 b,” Health IT Analytics,2018.
[7] G. M. Weber, K. D. Mandl, and I. S. Kohane, “Finding the missing link for big biomedical data,” Jama, vol. 311, no. 24, pp. 2479–2480, 2014.
[8] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” Nature medicine, vol. 25, no. 1, pp. 44–56, 2019.
[9] S. Benjamens, P. Dhunnoo, and B. Meskó, “The state of artificial intelligence-based fda- approved medical devices and algorithms: an online database,” NPJ digital medicine, vol. 3, no. 1, pp. 1–8,

2020.

[10] H. Shimizu and K. I. Nakayama, "Artificial intelligence in oncology," Cancer science, vol. 111, no. 5, p. 1452, 2020.

[11] N. Abbas and D. Mohamad, "Automatic color nuclei segmentation of leukocytes for acute leukemia," Research Journal of Applied Sciences, Engineering and Technology, vol. 7, no. 14, pp. 2987–2993, 2014.

[12] C. Di Ruberto, A. Loddo, and G. Puglisi, "Blob detection and deep learning for leukemic bloodimage analysis," Applied Sciences, vol. 10, no. 3, p. 1176, 2020.

[13] S. Shafique and S. Tehsin, "Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks," Technology in cancer research & treatment, vol. 17, p. 1533033818802789, 2018.

[14] J. Rawat, A. Singh, H. Bhadauria, J. Virmani, and J. S. Devgun, "Classification of acutelymphoblastic leukaemia using hybrid hierarchical classifiers," Multimedia Tools and Applications, vol. 76, no. 18, pp. 19057–19085, 2017.

[15] L. Bigorra, A. Merino, S. Alférez, and J. Rodellar, "Feature analysis and automatic identification of leukemic lineage blast cells and reactive lymphoid cells from peripheral blood cell images," Journal of clinical laboratory analysis, vol. 31, no. 2, p. e22024, 2017.

[16] J. Rawat, A. Singh, H. Bhadauria, and J. Virmani, "Computer aided diagnostic system for detection of leukemia using microscopic images," Procedia Computer Science, vol. 70, pp. 748–756, 2015.

[17] C. Reta, L. Altamirano, J. A. Gonzalez, R. Diaz-Hernandez, H. Peregrina, I. Olmos, J. E. Alonso, and R. Lobato, "Segmentation and classification of bone marrow cells images using contextual information for medical diagnosis of acute leukemias," PloS one, vol. 10, no. 6, p.e0130805, 2015.

[18] D. Umamaheswari, & S. Geetha, "A framework for efficient recognition and classification of acute lymphoblastic leukemia with a novel customized-knn classifier", Journal of computing and information technology, vol. 26, no. 2, pp. 131-140, 2018.

[19] L. Putzu, G. Caocci, and C. Di Ruberto, "Leucocyte classification for leukaemia detection usingimage processing techniques," Artificial intelligence in medicine, vol. 62, no. 3, pp. 179–191, 2014.

[20] J. Prellberg, & O. Kramer, "Acute lymphoblastic leukemia classification from microscopic images using convolutional neural networks". In ISBI 2019 C-NMC Challenge: Classification inCancer Cell Imaging (pp. 53-61). Springer, 2019.

[21] A. Rehman, N. Abbas, T. Saba, S. I. u. Rahman, Z. Mehmood, and H. Kolivand, "Classification of acute lymphoblastic leukemia using deep learning," Microscopy Research and Technique, vol. 81, no. 11, pp. 1310–1317, 2018.

[22] B.-S. Ko, Y.-F. Wang, J.-L. Li, C.-C. Li, P.-F. Weng, S.-C. Hsu, H.-A.Hou, H.-H. Huang, M. Yao, C.-T. Lin, et al., "Clinically validated machine learning algorithm for detecting residual diseases with multicolor flow cytometry analysis in acute myeloid leukemia and myelodysplastic syndrome," EBioMedicine, vol. 37, pp. 91–100, 2018.

[23] S. Warnat-Herresthal, K. Perrakis, B. Taschler, M. Becker, K. Baßler, M. Beyer, P. Günther, J. Schulte-Schrepping, L. Seep, K. Klee, et al., "Scalable prediction of acute myeloid leukemia using high-dimensional machine learning and blood transcriptomics," Iscience, vol. 23, no. 1, p. 100780, 2020.

[24] F. Kazemi, T. A. Najafabadi, and B. N. Araabi, "Automatic recognition of acute myelogenous leukemia in blood microscopic images using k-means clustering and support vector machine,"Journal of medical signals and sensors, vol. 6, no. 3, p. 183, 2016.

[25] O. Gal, N. Auslander, Y. Fan, and D. Meerzaman, "Predicting complete remission of acute myeloid leukemia: machine learning applied to gene expression," Cancer informatics, vol. 18, p.1176935119835544, 2019.

[26] S. Agaian, M. Madhukar, and A. T. Chronopoulos, "Automated screening system for acute myelogenous leukemia detection in blood microscopic images," IEEE Systems journal, vol. 8, no. 3, pp. 995–1004, 2014.

[27] J. Laosai and K. Chamnongthai, "Acute leukemia classification by using SVM and K-Means clustering," 2014 International Electrical Engineering Congress (iEECON), pp. 1-4, 2014 doi: 10.1109/iEECON.2014.6925840.

[28] M. Dundar, F. Akova, H. Z. Yerebakan, and B. Rajwa, "A nonparametric bayesian model for joint cell clustering and cluster matching: identification of anomalous sample phenotypes with random effects," BMC bioinformatics, vol. 15, no. 1, pp. 1–15, 2014.

[29] T. Manninen, H. Huttunen, P. Ruusuvuori, and M. Nykter, "Leukemia prediction using sparse logistic regression," PloS one, vol. 8, no. 8, p. e72932, 2013.

[30] M. Biehl, K. Bunte, and P. Schneider, "Analysis of flow cytometry data by matrix relevance learning vector quantization," PLoS One, vol. 8, no. 3, p. e59401, 2013.

[31] C. Matek, S. Schwarz, K. Spiekermann, and C. Marr, "Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks," Nature Machine Intelligence, vol. 1, no. 11, pp. 538–544, 2019.

[32] R. D. Labati, V. Piuri, and F. Scotti, "All-idb: The acute lymphoblastic leukemia image database for image processing," in 2011 18th IEEE International Conference on Image Processing, pp. 2045–2048, IEEE, 2011.

[33] C. Wu, X. Jin, G. Tsueng, C. Afrasiabi, and A. I. Su, "Biogps: building your own mash-up of gene annotations and expression profiles," Nucleic acids research, vol. 44, no. D1, pp. D313–D316, 2016.

[34] M. Mohamed, B. Far, and A. Guaily, "An efficient technique for white blood cells nuclei automatic segmentation," in 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 220–225, 2012.

[35] T. Barrett, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C.L. Robertson, N. Serova, S. Davis, A. Soboleva, "NCBI GEO: archive for functional genomics data sets -update", Nucleic Acids Res., vol. 41, issue. D1, pp. D991-995, 2013 https://doi.org/10.1093/nar/gks1193

[36] E. Clough and T. Barrett, "The gene expression omnibus database," in Statistical genomics, pp. 93–110, Springer, 2016.

[37] Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 30(1), 207-210.

[38] L. Siragusa, S. Cross, M. Baroni, L. Goracci, and G. Cruciani, "Biogps: navigating biological space to predict polypharmacology, off-targeting, and selectivity," Proteins: Structure, Function, and Bioinformatics, vol. 83, no. 3, pp. 517–532, 2015.

[39] Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., and Prior, F. "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," Journal of digital imaging, vol. 26, no. 6, pp. 1045-1057, 2013. DOI: 10.1007/s10278-013-9622-7.

[40] Matek, C., Schwarz, S., Marr, C., & Spiekermann, K. (2019). A Single-cell Morphological Dataset of Leukocytes from AML Patients and Non-malignant Controls [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/tcia.2019.36f5o9ld

[41] D. E. McNeil, T. R. Coté, L. Clegg, and A. Mauer, "Seer update of incidence and trends in pediatric malignancies: acute lymphoblastic leukemia," Medical and pediatric oncology, vol. 39, no. 6, pp. 554–557, 2002.

[42] P. Patel, K. Passi, and C.K. Jain, "Efficacy of non-negative matrix factorization for feature selection in cancer data", International journal of data mining & knowledge management process (IJDKP), vol. 10, no. 4, pp. 1-20, 2020. DOI: 10.5121/ijdkp.2020.10401

[43] F. Scotti, "Robust Segmentation and Measurements Techniques of White Cells in Blood Microscope Images", in Proceedings of the 2006 IEEE Instrumentation and Measurement Technology Conf. (IMTC 2006), Sorrento, Italy, pp. 43-48, April 24-27, 2006. ISSN: 1091- 5281. DOI: 10.1109/IMTC.2006.328170

[44] F. Scotti, "Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images", in Proceedings of the 2005 IEEE Int. Conf. on Computational Intelligence for Measurement Systems and Applications (CIMSA 2005), Giardini Naxos - Taormina, Italy, pp. 96-101, July 20-22, 2005 DOI: 10.1109/CIMSA.2005.1522835

[45] V. Piuri, F. Scotti, "Morphological classification of blood leucocytes by microscope images", in Proc. of the 2004 IEEE Int. Conf. on Computational Intelligence for Measurement Systems and Applications (CIMSA 2004), Boston, MA, USA, pp. 103-108, July 12-14, 2004. ISBN: 0-7803-8341-9. DOI: 10.1109/CIMSA.2004.1397242

[46] The website of the National Cancer Institute: https://www.cancer.gov

[47] H. T. Salah, I. N. Muhsen, M. E. Salama, T. Owaidah, & S. K. Hashmi, "Machine learning applications in the diagnosis of leukemia: Current trends and future directions". International journal of laboratory hematology, vol. 41, no. 6, pp. 717-725, 2019.

[48] Y. Wang, I. V. Tetko, M. A. Hall, E. Frank, A. Facius, K. F. Mayer, & H. W. Mewes, "Gene selection from microarray data for cancer classification—a machine learning approach". Computational

biology and chemistry, vol. 29, no. 1, pp. 37-46, 2005.

[49] NCBI Resource Coordinators (2016). Database resources of the National Center for Biotechnology Information. Nucleic acids research, 44(D1), D7–D19. https://doi.org/10.1093/nar/gkv1290

[50] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, R. Edgar, "NCBI GEO: mining tens of millions of expression profiles--database and tools update", Nucleic Acids Res. 2007 Jan;35(Database issue):D760-5. doi: 10.1093/nar/gkl887. Epub 2006 Nov 11. PMID: 17099226; PMCID: PMC1669752.

[51] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, R.N. Muertter, R. Edgar, "NCBI GEO: archive for high-throughput functional genomic data", Nucleic Acids Res. 2009 Jan;37(Database issue):D885-90. doi: 10.1093/nar/gkn764. Epub 2008 Oct 21. PMID: 18940857; PMCID: PMC2686538.