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 heterogeneous 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 hematology. 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
International Journal of Data Mining & Knowledge Management Process (IJDKP),
Vol.11, No.6, November 2021

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.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Type of Data</th>
<th>Pre-Processing</th>
<th>Classification Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Di Ruberto et al. [12]</td>
<td>Blood Smear Images</td>
<td>Hue Saturation Value; Blob detection; Segmentation watershed</td>
<td>Convolutional Neural Network (CNN)</td>
<td>94.1% - Leukemia classification</td>
</tr>
<tr>
<td>Shafique and Tehsin [13]</td>
<td>Blood Smear Images</td>
<td>Data Augmentation</td>
<td>Transfer Learning – DeepCNN</td>
<td>99.50% for normal vs cancerous cells; 96.06% for subtypes</td>
</tr>
<tr>
<td>Wang et al. [48]</td>
<td>Bone marrow samples</td>
<td>Feature selection</td>
<td>Decision tree; Naïve Bayes; Support Vector Machine</td>
<td>Explained in terms of percentage [48]</td>
</tr>
<tr>
<td>Rawat et al. [14]</td>
<td>Blood Smear Images</td>
<td>Segmentation; Feature Extraction (PCA)</td>
<td>Hybrid Hierarchical Classifiers</td>
<td>Overall classification accuracy 97.6%</td>
</tr>
<tr>
<td>Laosai et al. [27]</td>
<td>Blood smear images</td>
<td>Segmentation; Feature extraction</td>
<td>Support Vector Machine</td>
<td>92%</td>
</tr>
<tr>
<td>Rawat et al. [16]</td>
<td>Blood Smear Images</td>
<td>Segmentation</td>
<td>Support Vector Machine</td>
<td>72% - 86.7%</td>
</tr>
<tr>
<td>Reta et al. [17]</td>
<td>Blood Smear Images</td>
<td>Segmentation</td>
<td>Multiclass classifier</td>
<td>ALL: ~ 94%</td>
</tr>
<tr>
<td>Umamaheswari &amp; Geetha [18]</td>
<td>Blood Smear Images</td>
<td>Segmentation; Feature Extraction</td>
<td>K-Nearest Neighbour</td>
<td>96.25%</td>
</tr>
<tr>
<td>Putzu et al. [19]</td>
<td>Blood Smear Images</td>
<td>Segmentation; Feature Extraction</td>
<td>Support Vector Machine</td>
<td>93%</td>
</tr>
<tr>
<td>Prellberg and Kramer [20]</td>
<td>Blood smear images</td>
<td>Image flipping</td>
<td>ResNeXt CNN</td>
<td>F1 score – 88.91%</td>
</tr>
<tr>
<td>Rehman et al. [21]</td>
<td>Bone Marrow Images</td>
<td>Segmentation</td>
<td>CNN</td>
<td>97.78%</td>
</tr>
</tbody>
</table>
Table 2. ML Studies based on Acute Myeloid Leukemia.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Type of Data</th>
<th>Pre-Processing</th>
<th>Classification Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ko et.al. [22]</td>
<td>Multicolor Flow</td>
<td>NA</td>
<td>Support Vector Machine</td>
<td>84.6% - 92.4%</td>
</tr>
<tr>
<td>Cytometry</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Warnat-Herresthal et. al. [23]</td>
<td>Gene Expression Data</td>
<td>NA</td>
<td>L1 – Regularized Logistic Regression; Neural Networks</td>
<td>95% - 97%</td>
</tr>
<tr>
<td>Images</td>
<td>kmeans</td>
<td></td>
<td></td>
<td>validation</td>
</tr>
<tr>
<td>Gal et. al. [25]</td>
<td>Gene Expression Data</td>
<td>Feature Extraction(PCA)</td>
<td>KNN; SVM; Random Forest</td>
<td>AUC 0.73 – 0.84</td>
</tr>
<tr>
<td>Agaian et. al. [26]</td>
<td>Blood Images</td>
<td>CIELAB color features;</td>
<td>Support Vector Machine</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>segmentation; feature</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>extraction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laosai et. al. [27]</td>
<td>Blood smear images</td>
<td>Segmentation; Feature</td>
<td>Support Vector Machine</td>
<td>92%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>extraction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dundar et. al. [28]</td>
<td>Flow Cytometry</td>
<td>Dirichlet process gaussian</td>
<td>Non-parametric Bayesian algorithm</td>
<td>AUC 0.99 and 1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mixture model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manninen et. al. [29]</td>
<td>Flow Cytometry</td>
<td>Feature Generation</td>
<td>LR LASSO; LDA</td>
<td>100%</td>
</tr>
<tr>
<td>Biehl et. al. [30]</td>
<td>Flow cytometry</td>
<td>Feature vectors</td>
<td>GMLVQ</td>
<td>AUC 1.0</td>
</tr>
<tr>
<td>Matek et. al. [31]</td>
<td>Blood Microscopic</td>
<td>Digitised using oil immersion</td>
<td>Convolutional Neural Network</td>
<td>Precision and</td>
</tr>
<tr>
<td>Images</td>
<td></td>
<td></td>
<td></td>
<td>Sensitivity 94%</td>
</tr>
</tbody>
</table>

3. LITERATURE OF ML USING ACUTE MYELOID LEUKEMIA DATA

Acute Myeloid Leukemia (AML) starts in the bone marrow and quickly moves into the blood. It is 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 classification is 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

