Leukemia Cancer Classification using Micro GeneArray Dataset

By

Syyeda Shifa Fatima

(Registration No: 00000327075)

Department of Computing

School of Electrical Engineering & Computer Science (SEECS) ,National University of Sciences and Technology (NUST), Islamabad, Pakistan.

(August 2024)

Leukemia Cancer Classification using Micro GeneArray Dataset

By

Syyeda Shifa FatimaMSIT 2K20

00000327075

A thesis submitted in partial fulfillment of the requirements for the degree of

Masters of Science in

Information Technology (MS IT)

Supervisor

Dr. Sidra Sultana

Department of Computing

School of Electrical Engineering & Computer Science (SEECS) ,National University of Sciences and Technology (NUST), Islamabad, Pakistan.

(August 2024)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Leukemia Cancer Classification using Micro Gene Array Dataset" written by Syyeda Shifa Fatima, (Registration No 00000327075), of SEECS has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: Sidra

Name of Advisor: Dr. Sidra Sultana

Date: 31-Jul-2024

HoD/Associate Dean:

Date: 31-Jul-2024

Signature (Dean/Principal):

Date: 31-Jul-2024

FORM TH-4

National University of Sciences & Technology

MASTER THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: (Student Name & Reg. #)Titled: Syyeda Shifa Fatima [00000327075]

Leukemia Cancer Classification using Micro Gene Array Dataset

be accepted in partial fulfillment of the requirements for the award of Master of Science (Information Technology) degree.

Examination Committee Members

1. Name: Farzana Jabeen — Signature: Farzana 19-Aug-2024 4:53 PM

2. Name: Tahira Lashari — Signature: Tahira 19-Aug-2024 4:53 PM

3. Name: Nazia Perwaiz — Signature: 19-Aug-2024 4:53 PM

Supervisor's name: Sidra Sultana — Signature: Sidra 20-Aug-2024 12:31 PM

Arham Muslim
HoD / Associate Dean

21-August-2024
Date

COUNTERSINGED

22-August-2024
Date

M Ajmal
Muhammad Ajmal Khan
Principal

THIS FORM IS DIGITALLY SIGNED

Certificate of Originality

I hereby declare that this submission titled "Leukemia Cancer Classification using Micro Gene Array Dataset" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name: Syyeda Shifa Fatima

Student Signature: ______________

Date 26-August-2024

Dedication

This thesis is dedicated to my parents,
who have given me invaluable educational opportunities,
and to my sister,
Farzeen Fatima who has been my emotional anchor through these years.

Acknowledgments

To the Almighty who led me here,
To the professors who guided me.

To the failed experiments and dead ends,
That taught me how to regather my thoughts and move forward.

To the family who supported me through years,
Not only those I am bound by blood.

To the friends who made NUST a home,
And made this feat easier just by being there for me.

In the end a special thanks to Ms. Bhatti, Mahnoor Azhar and Maria Ayoub for being there for me every step of the way. Your love and support has meant everything to me on this journey.

Abstract

Leukemia, a heterogeneous group of hematologic malignancies, poses significant challenges in diagnosis and treatment due to its diverse genetic and molecular characteristics. This thesis explores the application of deep learning techniques to classify leukemia subtypes based on gene expression profiles. Utilizing datasets sourced from the Gene Expression Omnibus database, this study implements and evaluates three deep learning models: Long Short-Term Memory, Bidirectional LSTM, and a novel architecture termed xLSTM, which incorporates custom layers and residual connections for enhanced performance. Data preprocessing involved standardizing datasets, imputing missing values, and selecting the top 400 features through Chi-Squared testing. Models were trained on these processed datasets, and their performance was assessed through metrics such as accuracy and loss. The xLSTM model demonstrated superior performance, achieving a final test accuracy of 98.18%, outperforming both LSTM and BiLSTM models. Furthermore, the study examines the integration of multimodal data, combining gene expression profiles with images data to enhance classification accuracy. Recommendations for researchers emphasize the importance of high-quality, diverse datasets and advanced preprocessing techniques. Clinicians are encouraged to adopt deep learning tools in practice and participate in data sharing initiatives, while policymakers are urged to support funding, standardization, and ethical guidelines in AI applications. This research underscores the transformative potential of deep learning in medical diagnostics, advocating for continued innovation and collaboration across the biomedical and AI research communities to enhance patient outcomes in leukemia treatment.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction and Motivation

Leukemia, a pervasive form of blood cancer, profoundly impacts the hematopoietic system by inducing abnormal proliferation of white blood cells (WBCs) within the bone marrow and bloodstream. This condition disrupts the delicate balance of blood cell production, compromising immune function and systemic health. Despite advancements in medical diagnostics and treatment modalities, leukemia remains a formidable challenge due to its heterogeneous nature and variable clinical outcomes.

Present-day equipments such as DNA microarrays have facilitated a new strategy in analyzing genetic changes linked to development of leukemia. These tools help clinicians have deep images of the molecular pathology of the disease to enable more diagnostic accuracy and more targeted intervention plans. At the same time, the integration of such data mining techniques has become vital for mining meaningful patterns into voluminous genomic databases, which provide potential for biomarkers for early identification and risk evaluation. This thesis seeks to provide insights on how genetic data analysis and machine learning specifically; deep learning can be useful in the furtherance of acute leukemia. As such, leveraging extensive gene expression data relevant to acute leukemia subtypes is the ultimate goal of this research to identify concealed molecular pathways in pathological processes and designing early diagnosis models in acute leukemia. These developments may help enhance the function of intervention as well as individual optimization of therapies in connection to leukemia.

1.1 Background

Leukemia is a type of blood cancer that originates from the bone marrow and involves the cloning of immature WBCs cells or abnormal proliferation of mature WBCs. Leukemia is of several types depending on the condition as an acute or chronic illnesses and the type of white blood cell – lymphocytic or myelogenous. These complications comprise impaired immunologic competence, anemia, infection, and other systemic alterations due to the abrogation of the normal process of producing blood cells. Leukemia prevalence is not uniform across the world as there are certain geographical and demographic differences at present. It is well documented that possible causes of leukemia are not well understood, however, some contributing factors that can be confirmed include; genetic predisposition and environment. Molecular biologists have not only discovered different genetic changes linked to occurrence of the leukemia but also the molecular mechanisms through which the disease occurs.

Classically, the diagnosis of leukemia is reached by a combination of clinical assessment, routine laboratory investigations, bone marrow aspiration and biopsy, and cytogenetics to look for evidence of chromosomal abnormality. However, the last few decades saw a tremendous advance in diagnostic technologies, especially the popular DNA microarrays. These tools help assess gene expression profiles as well as catalogue mutation trends, giving clinicians a clear picture on the classification of different leukemia types, and thereby enhancing treatment plans. Thus, on the same note we have witnessed the adaption of data mining techniques that has become fundamental for determining important information from large datasets of genomic information. Computational algorithms when used to employ genetic data a researcher will detect biomarkers and explain illness processes or even patient prognosis better. However, there is need to acknowledge that several drawbacks exist to efficient prognosis of leukemia progression and treatment management with individual patient consideration. The application of other sophisticated computational algorithms including the use of deep learning could also help overcome these difficultries by providing more sophisticated features of genomic data and early diagnosis of leukemia. This background lays foundation for understanding how current advancements in genetics of analysis and other computational methods make a critical difference in comprehending the development of leukemia, accurately diagnose, and subsequently revolutionize the way of treatment in leukemia.

1.2 Leukemia

1.2.1 Introduction

Leukemia itself can be translated from the Greek language the word leukos, which means white and the word haima, which means blood, this is a condition where the bone marrow is producing increased number of leukocytes. This leads to the uncontrolled ability of these cells to multiply and hence they flood the bloodstream in exaggerated numbers. As mentioned earlier, the changes in bone marrow composition also affect the body health in various ways such as anemias, vulnerability to infections, bleeding problems, weakness, weight loss, and fatigue. After studying the courses of leukemia for years and years, no one can still tell how exactly leukemia is caused. However, several factors are known to influence its development, including age, gender, ethnicity, genetic predispositions, lifestyle habits like smoking and obesity, exposure to certain chemicals (e.g., benzene, pesticides), previous treatments like chemotherapy, and exposure to ionizing radiation. Leukemia ranks among the 15 most common cancers globally, with higher incidence and mortality rates observed in countries with higher human development indices (HDI).

1.2.2 Classification

Leukemia is categorized into four major subtypes based on disease progression and the specific type of white blood cells affected. The disease progression can manifest as acute or chronic forms. Acute leukemia is characterized by the rapid proliferation of immature "blast" cells, leading to a sudden onset of symptoms that can be life-threatening if left untreated. In contrast, chronic leukemia involves the slow and continuous growth of mature cells and may take years to manifest noticeable symptoms. Depending on the lineage of the affected white blood cells, leukemia is classified as follows:

- Acute Myeloid Leukemia (AML)
- Acute Lymphocytic Leukemia (ALL)
- Chronic Myeloid Leukemia (CML)
- Chronic Lymphocytic Leukemia (CLL)

1.2.3 Diagnosis and Treatment

It is important for leukemia to be diagnosed correctly and early so as to earn the right type of treatment.

- **Physical Examination:** For the management of asthma, clinical decision support can include subjective signs and symptoms, as well as general physical condition.
- **Blood Tests:** Such as: Full Blood Count (BCC), Peripheral Blood Smear for the identification of disorders in the number and shape of the cells.
- **Bone Marrow Aspiration or Biopsy:** To use anatomical lens to study the images of the bone marrow cells and their characteristic features.
- **Cytogenetic Analysis:** From there, they are used to detect chromosomal changes that define certain subtypes of leukemia.
- **Molecular Testing:** To identify characteristic chromosomal imbalances or gene rearrangements linked to leukemia.
- **Immunophenotyping:** In the research works, flow cytometry was employed to study the cell surface markers to distinguish between distinct subtypes of leukemia cells.

Once diagnosed, treatment strategies are tailored to the specific subtype of leukemia and individual patient factors:

- **Chemotherapy:** Chemotherapy from the Greek **chemo** meaning drug is one of the therapy techniques where treatment is given with the aim of eliminating cancer cells or preventing them from proliferating.
- **Radiotherapy:** It utilizes high energy rays for irradiation and obliteration of malignant tissues or cancer cells.
- **Immunotherapy:** The treatment known as Pronoun| involves the use of monoclonal antibodies to boost the body's immune response against leukemia cells.
- **Tyrosine Kinase Inhibitors:** This targeted approaches involved the use of drugs that inhibit certain enzymes that help in the growth of leukemia cells.

- **Hematopoietic Stem Cell Transplantation (HSCT):** This takes the form of the transplant of healthy stem cells into a standard bone marrow which has been ravaged by disease and is incapable of churning out normal blood cells.

The type of therapy that can be administered and the combination that might be used depend on the patient's physical health, age, the type of leukemia diagnosed, genetic characteristics, and the result of previous treatments. Currently, there exist research and developments done on medical science which keep on improving the treatment procedures as well as the results for patients who have leukemia.

1.3 Deep Learning

Deep learning forms a more complex subcategory in the comprehensives approaches in learning the machines and have been advocated for their potential usage in different fields including classification of diseases like leukemia. Neural networks are the type of machine learning algorithms that incorporate features of the human brain with an aim of making a multilayered representation system of data.

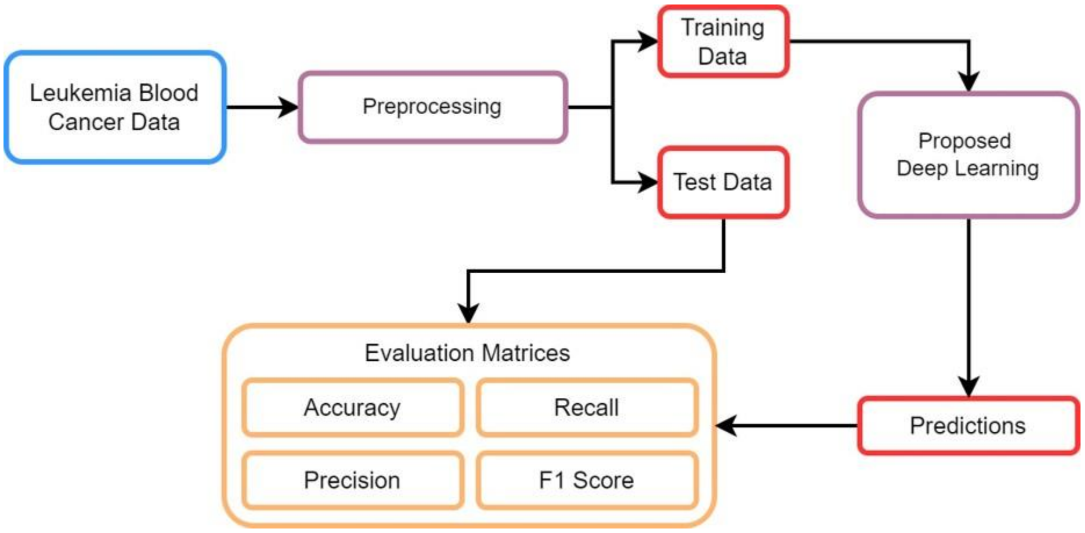

Figure 1.1: General methodology of deep learning model implementation

1.3.1 Application to Leukemia Research

Disease Classification

Various inputs has been fed into DL models to distinguish the diverse subtypes of leukemia Through CNNs and RNNs like the LSTMs. It is possible to use these models to diagnose patients for acute or chronic leukemia and myeloid or lymphocytic lines based on patterns identified in medical images, genetic information, and reports. This capability is very useful in diagnosing patients correctly and within the shortest time possible, which is very vital for commencement of adequate treatment regime.

Feature Extraction

We see deep learning methods embedded here as having one of the major advantages in leukemia research because of its capability to learn features from the raw data devoid of any interpretation of them. For example, the deep learning models can point out genetic mutations or features in the leukemia associated genes that are not discernable by naked eyes in genetic analysis. Likewise, in medical imaging, CNNs can learn to extract features from the images of blood smear or bone marrow biopsy which are indicative of leukemia.

Predictive Modeling

Other applications of deep learning include the use in predictive modeling of outcomes in leukemia and treatment response. Such models use patients’ medical history, treatment history, and genetic data stored in patients’ medical records for different timepoints in patients’ lives for estimating patients’ outcome, such as survival or recurrence rate. This aspect supports clinicians in making recommendations on other likely clinical courses, as well as the best means of managing particular patients.

1.4 Problem Statement

The classification of leukemia using machine learning models, particularly ensemble methods and deep learning, poses a challenge due to the substantial computational resources required. This demand results in high time complexity during both the training and prediction phases of these models. As such, there is a critical need to optimize the time complexity associated with these advanced models while ensuring or enhancing classification accuracy. This study aims to address this challenge by exploring methodologies to streamline the computational efficiency of leukemia classification models. By optimizing algorithms, enhancing feature selection techniques, and possibly leveraging parallel computing frameworks, the goal is to reduce the computational burden without compromising the models' ability to accurately classify leukemia subtypes. The findings of this research are expected to contribute significantly to the development of more efficient and practical tools for leukemia diagnosis and prognosis, ultimately benefiting clinical decision-making and patient outcomes.

1.5 Research Gap

Despite advancements in the use of gene expression profiling for leukemia classification, existing methods often fall short in handling the complexity and high dimensionality of the data. Traditional statistical and machine learning approaches lack the capacity to capture intricate patterns and interactions within gene expression datasets, resulting in limited classification accuracy. There is a need for more advanced models that can effectively process and analyze these large datasets to improve diagnostic precision. Additionally, while deep learning has shown promise in various biomedical applications, its full potential in leukemia classification, particularly with innovative model architectures, remains underexplored.

1.6 Research Questions

To address the identified research gap, this study focuses on the following key research questions:

- How can deep learning models be effectively utilized to classify leukemia subtypes using gene expression data?
- What are the limitations of existing deep learning models in leukemia classification, and how can these be addressed?
- Can a novel deep learning architecture, incorporating custom layers and residual connections, outperform traditional models in terms of classification accuracy?

These questions guide the research in developing, implementing, and evaluating deep learning models to enhance leukemia subtype classification accuracy, ultimately aiming to improve patient diagnostics and outcomes.

1.7 Research Objectives

The research objectives are as follows:

- Utilize deep learning algorithms to achieve accurate predictions for leukemia classification.
- Create a leukemia classification model using BiLSTM and xLSTM, and embed a genetic algorithm to optimize hyperparameters.
- Assess the performance of the proposed system using evaluation metrics such as accuracy, precision, recall, and F1 score.
- Perform experiments to compare the proposed system with other deep learning models, including CNN, LSTM, BiLSTM, and xLSTM, to determine relative performance.

CHAPTER 2

Literature Review

2.1 Introduction

Leukemia is a genomic and phenotypically diverse heterogeneous group of diseases resulting from the malignant transformation of blood cells and leading to the accumulation of immature cells capable of destroying the normal function of the blood and/or bone marrow; due to this, it is a particularly challenging cancer to diagnose and to treat. Methods of leukemia classification by conventional light microscopy, conventional interpreting of blood films or Marrow aspiration smears is tiresome, time consuming, subjective and may produce Errors. Thus, increasing importance is being placed on creating more automatic and accurate diagnostic tools using ML/DL computational methods. Over the past decade, methods from ML and DL have significantly advanced the way of diagnosing illnesses in medicine by providing powerful statistics to model big datasets, for instance, of gene expression and microscopic images [1] [2] [3]. Such enabling technologies have contributed to increase the efficiency, speed, and effectiveness of the classification of leukemia disease, and therefore develop proper treatment techniques. This literature review aims at focusing on the use of ML and DL methods especially in classification of leukemia using gene expression data and microscopic image analysis. Thus, the main paradigm is the characterization and assessment of newly developed and employed computational methods and the outcomes attained by employing these strategies.

2.2 Gene Based Leukemia Cancer classification

Among all types of cancer, leukemia is considered a group of blood cancer with diagnostic and therapeutic issues. It has been noted that the traditional ways of cancer classification are not very effective; therefore, different systematic approaches based on the global gene expression analysis have been proposed [4]. This work undertakes a study that develops a machine learning model for the classification of leukemias from the dimensional gene expression data [5]. Principal Component Analysis (PCA) in employed to select features while for the ensemble learning to avoid overfitting, the Stacking algorithm is used. Thus, the findings indicate that the proposed model was able to achieve an accuracy of 95. 5% accuracy along with high of precision, recall and F1-score (0. 96, 0. 95, and 0. 95 respectively), which gives better results compared to that of traditional ML algorithms on the given data set. Hence, from the findings of this study, the author posits that ensemble learning is preferable when dealing with high-dimensional small samples in terms of classification, accuracy, precision, recall, and F1 score.

Proper distinctions of cancer subtypes like ALL [6], AML [7], and others can help in the correct determination of personalized therapies. A study's main objective is to develop and implement a gene expression analysis method based on compressive sensing (CS) directed at subtyping acute leukemias [8]. The CS strategy is based on signal rebuilding using a small projection set that accurately identifies ALL and IML. The method in this study showed a 97% hit rate for the classification of 38 subjects (27 ALL, 11 AML) using genes from 7129 samples. The research reveals that CS is a significant tool for leukemia subtype identification, leading to a higher limit of diagnosis disorders. Another study proposes a novel approach for classifying leukemia subtypes based on gene expression levels and high classification precision in the data [9]. The accessibility of the CS approach to conducting classifications with a small subset of signals instead of the whole datasets and computing resource savings are the benefits of the CS method. The analysis of the LOO method gave complete reliability in the ALL group's cases, but a misclassification in the AML samples may have happened due to sample size differences. AML sample size should be increased to minimize false-positive or false-negative cases. Interestingly enough, only one gene, "X95735," could bring to 97.4% accuracy of separation of B and B-ALL, indicating its essential role in this task. The designed classifier is equipped with classing detection, making it applicable to many detection classes. Furthermore, following up on the study intends to train the classifier on another testing dataset and work on the gene selection procedures used for the classifier.

Leukemia classification is mainly based on the speed of the spread, which consists of four types. The genes are the main engine of human physiology, and gene feature variations convert to both an occurrence of cancer and the risk of this disease [10]. A study which uses the Naive Bayes algorithm to determine how leukemia patients are categorized based on their gene features and can predict positive vs. adverse outcomes [11]. The GEO was contacted at NCBI for research purposes, and normalization through the suggested approach based on the similarity to create a fuzzy table was approved. The hybrid species, composed of a classy approach, a graph-based technique as a middleman, and a statistical model, made this case unique. At the end of it, the accuracy rate in predicting the disease has been high. A leukemia-identifying blood cell analyzer is CBC, which is widespread; however, the classification performance of the algorithms is different in accuracy. The model with the Naive Bayes algorithm produced the best results, with an accuracy level reaching 95%. As this research uses machine learning and big data concepts, increasing the dataset size can increase the model's accuracy. Short-term work focuses on applying AI technology for supervised learning, which can also provide a precise and accurate classification for leukemia.

Tumor growth is known to be associated with gene mutations and abnormal expression. Therefore, the genes with differential expression can be the aim of tumor gene therapy [12]. Researchers introduces a gene differential expression analysis technique for relative risk ratio to identify the genes in these links. It is then applied to a leukemia gene expression dataset and evaluated using C4.5, Naïve Bayes, and Support Vector Machine (SVM) classifiers [13]. They have shown that this method brings better classification accuracy than the SAM. The study introduces a novel approach for differential expression analysis of genes associated with leukemia based on RR. The method shows promising results in improving classification accuracy compared to existing methods. Differentiation of blood cells helps determine cancer type, particularly leukemia, which affects bones and the production of blood [11]. Gene expression profiling involves a solid and systematic method to classify leukemia objectively and systematically. In a study reseachers analyzes gene expression data from the Golub dataset, adopting supervised learning algorithms that recognize samples as AML or ALL [14]. Results obtained via Logistic regression are informative, while the Neural Networks are efficient when large datasets are involved. The comprehensive literature review points out the wide use of the classification algorithm in a myeloid leukemia diagnosis. However, only some of the successful cases did find their continuation in the clinics. This study brings new knowledge into leukemia diagnosis by conscientiously producing insights into the classifications of gene expression data.

The rationale for the molecular classification of leukemia cells into various sub-groups is to provide a basis for the proper diagnosis and therapy [15]. An investigation presents a novel multi-objective optimization algorithm (MOA) with two benchmark algorithms of particle swarm optimization (PSO) and two-layer PSO (TLPSO) as a comparative algorithm test [16]. As demonstrated by the experiment, classifying gene expression profiles has similar performance and predominantly accuracy, making MOA a candidate classifier. The MOA exhibits extreme scalability and truthfulness, especially in low-dimensional datasets. Yet, it cannot still process a bigger dataset because its power is limited to classifying in low-dimensional datasets only. Researchers can use the MOA for more effective performance in handling large datasets and the number of genes and for better classification. The MOA algorithm exhibits some advantages in classifying high-dimensional datasets as a clustering algorithm, especially when this problem concerns lower dimensions.

The field of disease classification based on gene expression data has many limitations for two main reasons: high dimensionality and small sample size [17]. A study aims to develop a unique method based on boosting to make a linear combination of weak classifiers with robust accuracy enhancement and a complex feature reduction. Boosting, SVM and K-nearest neighbor are three classification techniques used together to improve the prediction model [18]. Using the ensembled model as an example on a colon cancer dataset. It outperforms any single classifier, and we got better results than any other algorithm. The experiments attest that the Adaboost method is efficacious in the feature selection tasks, thus making the classification step more time-efficient and accurate in the future. An ensemble learning model that trumps a single classifier and combines the strengths of the robust classifiers dilutes the weaknesses of the weaker classifiers and is more accurate and reliable. The current approach is well suited to binary data classification, so the following research steps will involve multigene data classification and intragene interactions [19].

The gene choice is a pivotal factor and the basis for microarray analysis and classification. Gene identification is less about traditional algorithms as they are concerned about identifying due to the redundancy of genes [20]. In a study, a new gene selection algorithm has been developed that considers the interaction between the genes. Although single genes could be practically useless in isolation, their mutual interplay can unravel helpful information for classification purposes [21]. Our method is validated in four cancer microarray databases and gives satisfactory performance outputs surpassing the existing methods on this metric. The IDIS algorithm provides better classification results than existing classifiers, distinguishing selected enzymes with high accuracy. It chooses more genes on average and also slows the final classification compared to fast variants. On the contrary, if AI systems can achieve better performance, then extra time for making those calculations corroborates this fact. The outcomes show that one could utilize gene interactions in the gene selection for microarray analysis and would place them as classifiers. Due to a fast generation of microarray tech, the micro statistics area grew fast, aiming to develop individual genes [22]. The researcher is interested in introducing a novel mixture of feature selection methods in which the newly developed features (Chi-Square Statistic and Support Vector Machines with Recursive Feature Elimination) are applied [23]. This method is responsible for choosing a set of 10 informative genes from such high-dimensional microarrays, and, according to the previous data comparison, the chosen 10 genes showed better classification results.

Calculating gene expression for thousands of genes is too labor-intensive. Hence, only up to ten genes leading to a better outcome are selected out of the many genes using a new selection method (ChiSVMRFE). The method compares well with others, namely precision and efficiency, on the high data dimensions. It is clear that the method is superior when the data is complex and high-dimensional. ANN+KNN with ChiSVMRFE performed the best and showed the best classification accuracy, while the decision tree had the lowest accuracy [24]. Further work can investigate the implied method's applicability to more microarray datasets and the inclusion of other evolutionary approaches, like genetic algorithms, to act as wrapper feature selection methods [25]. The examination of gene expression data characterizes cancer. It is an essential issue for developing effective treatment and drug discovery. Microarray technology has enabled us to watch thousands of genes, but the features come in thousands of genes and fewer samples, which challenges existing classifying methods. This proposal delivers a brand-new technique employing precision levels to forecast genes' significance and, thus, the samples' classes. The outcomes have revealed that the considered approach can gain high classification rates related to the number of small or large genes. Although the prediction is less reliable with the increased number of genes, the standalone cause of death is a noteworthy exception. The method ranks analogous with kNN and SVM, and the exact number of genes for maximum accuracy has not been accessed.

2.3 Deep Learning Based Leukemia Cancer Classification

CML (chronic myeloid leukemia), a rare leukemia, is one of the subtypes of leukemia caused by genetic abnormalities in early myeloid cell progenitors [26]. As for diagnostic methods like blood smear, bone marrow aspiration, and biopsy, which are highly costly and time-consuming are the reasons why the innovation in diagnostic techniques is needed [27]. To convenient, low-cost, and highly accurate detection of leukemia, new automation methods need to be developed. Imaging process including stained blood microscope images, in addition to other technologies, is traced by the researchers as one of the most reliable approaches for discovering leukemia [28]. The research has been infused with deep learning technologies that can be utilised for precision diagnosis of CML by medical professionals. Most of the existing algorithm include image segmentation and feature extraction mainly for this purpose and this paper plus the framework propose a new BiCNN-CML algorithm to make that happen [29]. The structure entails gathering and transforming datatoxception, division of images by means of FOCM and PSO, extraction of features with GCLM , and then latter categorisation by means of Bi-LSTM with CNN technologies. Experimental data prove that the proposed method is effective. It increases the likelihood of identifying the CML accurately and logically.

Leukemia is a specific sort of blood cancer that can disrupt the usual array of developing blood cells with their subsequent role in the immune system and the circulatory system's ability to manufacture essential blood cells. The traditional diagnosis methods often are resource-expensive and time-consuming which motivates development of non-manual systems that combine image processing with machine learning. A study recommends an intelligent machine-based system which: 1) distinguishes eventually between acute lymphoblastic leukemia (ALL) and its types 2) shows a high classification success for ALL types, and 3) excels present methodologies and illustrates the advantages of machine learning in leukemia diagnostics and treatment [30].

Leukemia, the most common malignant disorder, is afflicting the people of all ages but the Acute Lymphoblastic Leukemia (ALL), notorious for its life threatening nature, ranks the worst as far as fatality rate is concerned. Manual analysis of addition diseases diagnosed by microscope and looking at cell images is time-consuming and inaccurate. It is an editorially approved sentence. The author proposed method of diagnostic is the use of deep learning with the YOLOv3 model and the computer aid this. The application of a machine learning model to this data resulted in a training accuracy of 97.2% and an mAP number of 99.8%, which show the effectiveness of such an approach in clearly demarcating leukemic cells [31]. The study draws attention to the prospects of using deep learning methods in the challenges of the leukemia diagnosis. Early and accurate diagnosis of leukemia, the wide class of a disease which affect white blood cells (WBCs) within a hematological system, is the important factor of successful treatment plans for the disease [32]. It is cost-effective and non-invasive microscopic method for leucocytes examination as a tool for the assessment of leukemia. AI acts as such a base and a background on which the automated detection methods can rely on for greater precision and speed. This editorial submits the findings of a systemic review of recent update of AI tool in detecting ALL. Different techniques, like signal processing, image processing, and conventional machine learning, as well as those based on deep learning, are scrutinized and categorized [33]. Traditional machine learning technologies (e.g. supervised and non-supervised learning) and novel deep learning approaches which include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Autoencoders are focal point of this discussion [34].

The diagnosis of leukemia has been a manual process involving blood smears examination by trained professional that is not only time-consuming but also is prone to human error. To tackle these gaps, fully automated methods using artificial intelligence based on machine learning and deep learning techniques have been invented. This study informs a novel feature fusion-based deep learning methodology for extracting leukemia cells from blood smears. The model proves, in effect, an astounding accuracy of 99.3% that shows evident benefits of reducing the diagnostic time and improving accuracy of previous ones [35]. Leukemia is a formidable group of malignant hematological disorders that is characterized by a complicated process of diagnostics and therapy. Traditional methods of organizing plants have certain shortcomings, hence, came the novel and comprehensive procedures today based on the investigation of a molecular level gene expression. Another study describes applying machine learning techniques to classify leukemia using genetic expression data with reduced dimensions. The model uses a linear combination of the new features and the Principal Component Analysis (PCA) as a selector for feature and the ensemble learning Stacking algorithm [36]. The results presented above are clearly impressive: we ended with 95.5% accuracy and showing high precision, recall and F1-score; namely, 0.96, 0.95 and 0.95, correspondingly, and this outperforms the classic machine learning algorithms working on the same dataset. The work has shown the efficacy of the ensemble learning in case of the classification multicast of the blood cancer as the small-sample datasets and the high-dimensional are considered.

Leukemia is an the immunory cause which affects the cell harbored in the bone marrow. This disorder demands urgent diagnosis for timely treatment. Some of these systems, based on machine and deep learning technologies have proved themselves effective in reducing the workload of medical professionals, bringing a much-improved result for patients. Transfer learning a technique used as the classifier in the biomedical field due to limit of annotated data is something that can be used for the classification task too. This study bridges the gap between DL and microscopic image analysis in leukemia using transfer learning, which is implemented on 1358 blood smears for disease classification [37]. The work under consideration has optimized the VGG16 model, which was trained on the leukemia dataset, by applying a fine-tuning approach in order to differentiate between the images of acute leukemia, chronic leukemia and healthy samples, thus achieving the accuracy of 93.01%. The results showed helpful models in classifying cancers with high quality and leading to better services for the patients.

Leukemia, very dangerous cancer with mortality rate higher than average, mainly occurs in 5-8 years old kids, but can happen in all ages group. Classification until recently had been very much driven by microscope-based manual comparison of blood images to determine whether the cells indicated leukemia or normal, yet in this comparison, the cells of leukemia and normal ones sometimes look alike. In that respect, an alternative technique depends on the image classification algorithms, mainly DL approaches such as CNNs. Machines Harnessing the power of Machine Learning (ML) and Deep Learning (DL) models the automated classification of image is becoming common practice in the diagnosis of leukemia. This paper is targeted at reviewing the literature explaining the application of ML and DL technologies in leukemia's diagnosis and prediction [38]. In addition, it will aim to determine and derive ML and DL-based image classification algorithms providing highest level of accuracy for leukemia cancer cells prediction.

Usually there is a need to classify blood cells into specific subtypes of leukemia to have precise diagnostics. Usually, such a study requires a decent size training dataset (which consists of people from different races and referring also to the genetic characteristics) which in turn is the key to that issue. While the first approach of using imaging flow cytometry requires a smaller training dataset for classification, which is advantageous for local variations as well as individual and racial differences, the second one excludes such differences. The present study is about a technique designed for acute leukemia classifying by means of imaging flow cytometry combined with morphology [39]. It includes the use of light brightness and cell morphology features from flow cytometry imaging and blood smears, respectively, as the coarse step of the method, in which blood cells is divided into healthy, Acute Lymphoblastic Leukemia(ALL),and Acute Myeloid Leukemia(AML), while in the fine step, through deep learning, they are classified into sub Classifying on samples from the Thai and American populations has shown an improved rate compared to conventional techniques. The computer simulations on complete blood image prove the robustness of our scheme and the capability to differentiate Acute Leukemia. Immediately discussed, this system reaches accuracy at 99% which represents 4% percent higher than traditional methods.

CHAPTER 3

Methodology

3.1 Data Collection

Data for this study was sourced from the Gene Expression Omnibus (GEO) database, a public repository that archives and freely distributes comprehensive gene expression data sets. The datasets utilized for this research included:

- GSE14317
- GSE22529-U133B
- GSE28497
- GSE33615
- GSE63270
- GSE71449
- GSE71935
- GSE9476

These datasets were chosen mainly depending on their coverage to cover all the spectrum of normal samples and different types of leukemia and also they have included the normal samples. The datasets included both microarray and sequencing data and the use of both forms of data made the analysis and validation of the models proposed more effective.

Table 3.1: Description of the Datasets

GSE	GPL Plateform	Number of Samples	Number of Genes	Number of Classes
14317	571	25	22278	2
22529	97	52	22646	2
28497	96	281	22284	7
33615	4133	71	33580	2
63270	17810	101	54676	2
71449	19197	45	52201	4
71935	570	46	54676	2
9476	96	64	22284	5

3.2 Libraries Used

Several Python libraries were utilized throughout the project, each serving specific functions in the data processing and model development pipeline:Several Python libraries were utilized throughout the project, each serving specific functions in the data processing and model development pipeline:

- **NumPy:** Essential for performing arithmetic and offering servings for big and multi-indexed arrays and matrices, also comprising a set of mathematical functions servicable for arrays.

- **Pandas:** Crucial in the process of handling data as well as computations to be done on such data. Pandas contains all the data structures and functions essential for operating on structure data conveniently.

- **scikit-learn:** It is applied for data preprocessing, feature selection, and for applying any machine learning algorithm. It contains a number of tools for model fitting, data transformation, model selection and evaluation.

- **imblearn:** Mainly used for addressing the problems of imbalance dataset, the usual methods are over-sampling, under-sampling, and ensemble learning.

- **matplotlib and seaborn:** The libraries that were used to make basic static and animated diagrams are the ones that help to analyze the distribution of data and the performance of the model.

- **tensorflow. keras:** One software that is quite useful in building and training deep learning models. Originally, Keras is an API that was developed to be easily understandable by a human brain and not a machine. As well as, it supports an easy and convenient way for creating neural networks.
- **umap-learn:** Used for reducing the dimensionality and visualization of data set in order to get better understanding of high dimensionality data.

3.3 Data Preprocessing

Data preprocessing can be a critical process in the evaluation of the data where the data is preprocessed to enhance its quality, structure for model training. The preprocessing steps undertaken were as follows:

3.3.1 Data Loading and Concatenation

The different datasets where imported using the Pandas read functions and combined by using the concat function. This consolidation was necessary so as to have a large set with all these samples and gene expression profiles that are necessary for the analysis.

3.3.2 Removal of Unwanted Types

Some of the records regarded certain subtypes of leukemia that were unrelated to the research goals and objectives, and therefore excluded. This step helped in eliminating bias when analyzing the data by only concentrating of the core values that are relevant in making the decisions, thus helping the model to generalize on the data that is fed to it.

3.3.3 Handling Missing Values

To deal with the situation of missing values in the dataset, the mean strategy was applied. Missing values in the dataset were also handled by substituting the missing values for each feature by the mean values of that part to have an integrant data set throughout the bi-secant method, and no bias or mysterious error is noticed in modelling the data.

3.3.4 Feature Selection

Feature selection is always typically done through the Chi-Squared (Chi2) test. This statistical method tests the correlation of each analyzed feature with the target variable and makes it possible to choose only the most relevant features. The features with the highest Chi2 scores were selected and as a result, data dimension was reduced to 400 features improving the model's performance.

3.3.5 Data Splitting

Since the preprocessed data in this analysis was small, it was divided into training and test data. This split is pertinent for assessing the model's robustness in extrapolating data, giving a measure of extrapolation. Previously the pattern was 80/20, that's 80% of the data was assigned to training and 20% to testing.

3.4 Proposed Architecture

The proposed architecture diagram (Figure 3.1) outlines a comprehensive machine learning pipeline designed to handle both image and gene data, aiming to leverage the strengths of multimodal data integration for improved predictive performance. The process begins with loading the necessary libraries, establishing the foundation for subsequent tasks. In the Data Preprocessing phase, the pipeline loads and concatenates data from various sources, encompassing both image and gene datasets. Following this, unwanted data types are removed to streamline the dataset, and any missing values are addressed through appropriate imputation methods. Feature selection is then performed to reduce dimensionality and enhance model efficiency.

The dataset is split into training and testing sets to facilitate unbiased model evaluation. For text data that needs to be represented visually, a conversion step transforms text to image format. The image data is then loaded and processed using the VGG16 model, a pre-trained convolutional neural network, to select relevant features. This careful preparation ensures that the most informative aspects of the data are retained for model training.

In the Model Training and Evaluation phase, separate models are trained on the image data and gene data respectively, capitalizing on the unique characteristics of each data type. These models are then combined through a multimodal fusion process, integrating the strengths of both models to create a more robust predictive model. The final phase involves evaluating the performance of this fused model, typically by assessing its accuracy or other relevant metrics. This detailed architecture emphasizes the importance of rigorous data pre-processing, specialized feature selection, and innovative data integration techniques to achieve high-accuracy predictions in complex multimodal datasets.

The gene expression data is converted to vizualization in a manner that captures the underlying patterns and relationships. This approach involves transforming high-dimensional genomic data into a 2D or 3D scatterplot, where each point represents a sample or a gene, colored and positioned according to its expression levels and relationships to other data points. A sample of the visualization is expressed in Figure 3.2.

3.5 Deep Learning Models

In this study, three different deep learning architectures were implemented and evaluated: LSTM, BiLSTM, and a new type of LSTM that is here named xLSTM. Each of the models was intended to be convoluted enough for addressing the challenges associated with gene expression data, with certain advancements that were made for furthering each model's performance.

3.5.1 LSTM

LSTM is a class of RNN that is able to learn long-term dependencies used in Natural Language Processing. These models find their application in sequential data where the context in the previous states plays a vital role in the estimation of future states. In the current work, an attempt was made to use LSTMs in the classification of leukemias using gene expression data. The specific LSTM architecture was chosen as it is capable of modeling the temporal dependencies in the gene expression data due to the structure's memory characteristic.

3.5.2 BiLSTM

BiLSTM is a type of LSTM that reads the give input sequence both from forward and backward directions. This architecture entails the model to capture the contextual information of the past and the future states and therefore it avails a more enhanced general understanding in the sequential data. In the meaning of the gene expression data, the BiLSTMs could more describe the relationships between genes by using not only forward connections but also backward ones.

This bidirectional procedure increases the model's capacity to learn complex features, which in turn leads to higher classification accuracies.

3.5.3 xLSTM

The xLSTM architecture introduced several novel custom layers designed to enhance the model's learning capabilities:The xLSTM architecture introduced several novel custom layers designed to enhance the model's learning capabilities:

- **ExpGating Layer:** This layer applies exponential gating mechanism in which the values of the input layer are raised to a certain power of two. The exponential gating can be especially useful for the model due to limiting the dependence on insignificant features and possibly enhancing learning from the given data.
- **sLSTM Layer:** integrates the exponential gating mechanism with LSTM capacities which enables the model to scale the LSTM output at the series of input features. These are to complement one another with an intention of improving the LSTM's capabilities of learning and exploiting information from the data.
- **mLSTM Layer:** An LSTM layer has been altered in this context by changing the kernel weights in a bid to enhance the comprehension of the data element interdependencies. They hypothesize that this customization will increase the model's ability to learn from expression data received through genes in the network.

The xLSTM model also involved a residual connection architecture at the same time. The so called residual connections are beneficial in preventing the vanishing gradient problem that is notorious in deep networks, and make the training of deeper networks feasible. The last architectural component was to introduce residual connections to the multiple xLSTM blocks where in stacking the blocks the residual connections enhance the learning process from data depend on the complexity of data that enhances performance.

3.6 Evaluation

The trained xLSTM model was tested on the test dataset in order to determine the effectiveness of the model used. The main performance measure adopted in assessment was accuracy which is the ratio of samples correctly classified to the total number of samples. Low accuracy signifies the model's reliability in differentiating various classes of leukemia depending on the gene expression.

The following evaluation parameters were used during the training process in order to assess the efficiency of the learning process and the ability of the model to generalize on unseen data; the loss and the validation accuracy. The results shown that the xLSTM model, with the proposed layers and residual connections, fulfilled high accuracy and was effective in leukemia classification tasks. This evaluation further validated the use of the given model in identifying the relationships that exist in gene expression data to arrive at pertinent predictions.

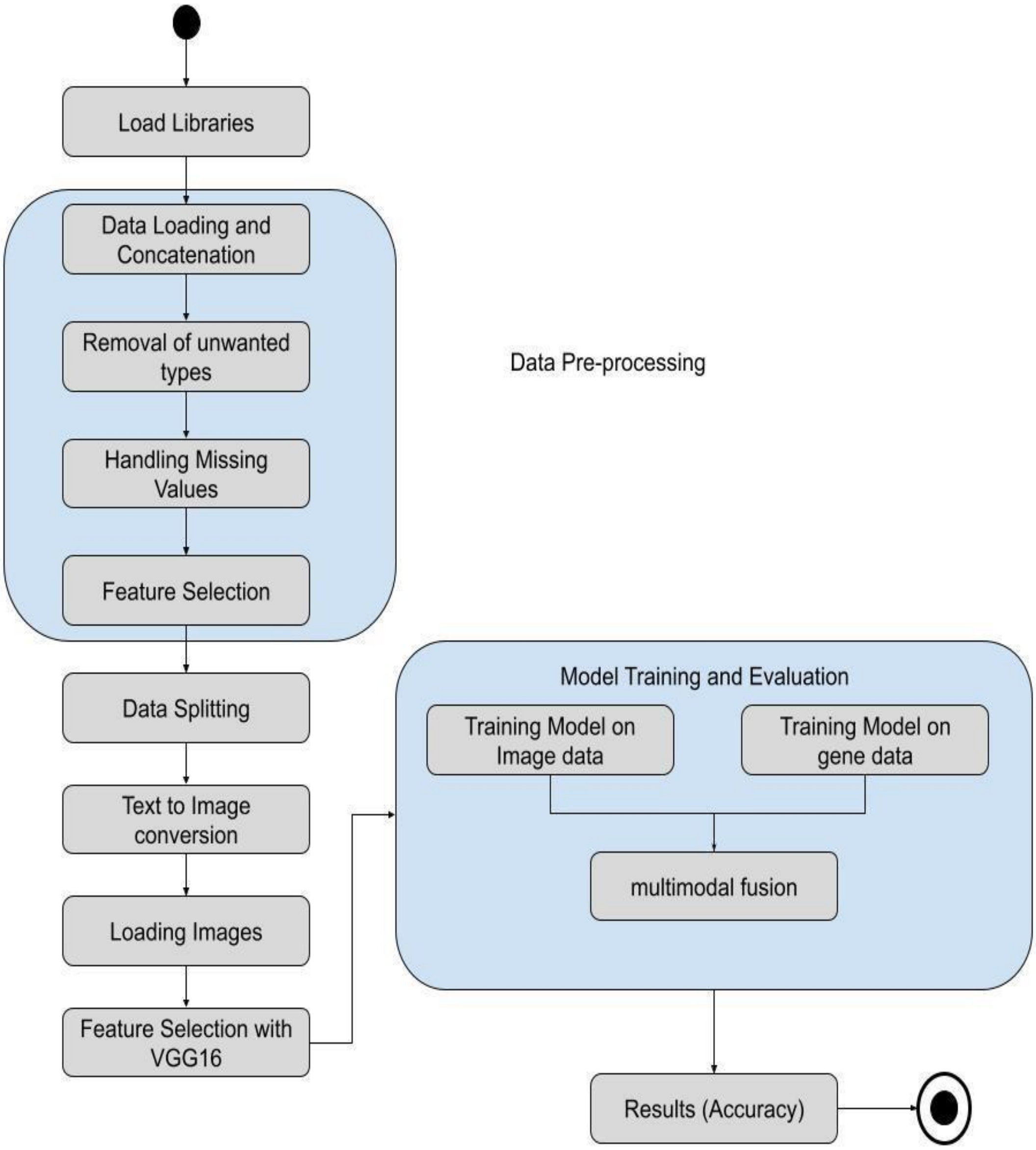
Load Libraries
Data Loading and Concatenation
Removal of unwanted types
Handling Missing Values
Feature Selection
Data Pre-processing
Data Splitting
Text to Image conversion
Loading Images
Feature Selection with VGG16
Model Training and Evaluation
Training Model on Image data
Training Model on gene data
multimodal fusion
Results (Accuracy)

Figure 3.1: Proposed Architecture

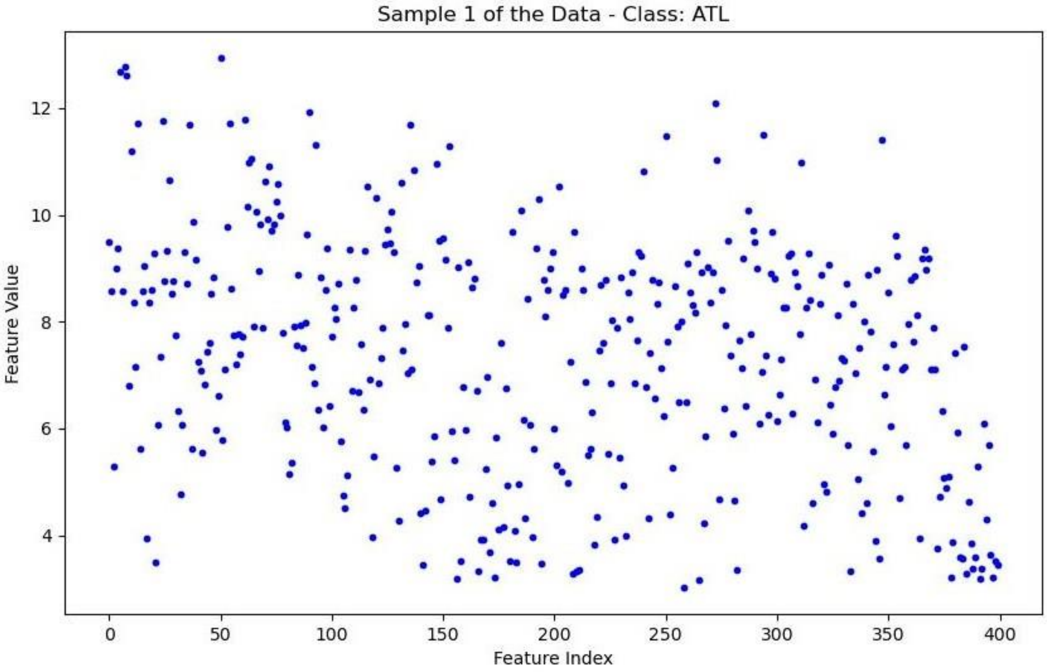

Figure 3.2: Data Sample of Text to Image Conversion

CHAPTER 4

Results

This chapter describes the results achieved in the process of the deep learning model training and testing for the classification of leukemia based on gene expression data. The methodologies explained in Chapter 3 were employed to respond to the research questions and consequently meet the specific aims of this research study. The outcomes include the corresponding metrics of the calibration of the model and its performance as well as results from evaluation studies as well as comparison understandings in order to gauge the efficacy of the different architectures of the models.

4.1 Preliminary Research

This table presents preliminary research results evaluating the accuracy of various machine learning models applied to image data processed using different methods. The models are compared under three different conditions:

1. Vaibhav R. et al., 2022 (Chi2+Adasyn): This column shows the accuracy of the models as reported by Vaibhav R. et al. in 2022, where they used the Chi-squared feature selection method combined with the ADASYN (Adaptive Synthetic Sampling) technique to handle class imbalance.

2. Chi2+ADASYN (Image Data): This column displays the accuracy of models trained on image data processed using the Chi-squared feature selection method and ADASYN for balancing the dataset.

3. Chi2+SMOTE (Image Data): This column presents the accuracy of models trained on

image data processed using the Chi-squared feature selection method and SMOTE (Synthetic Minority Over-sampling Technique) for balancing the dataset.

The machine learning models included in the comparison are:

- RF (Random Forest)
- LR (Logistic Regression)
- SVC (Support Vector Classifier)
- KNN (K-Nearest Neighbors)
- NB (Naive Bayes)
- ETC (Extra Trees Classifier)
- DT (Decision Tree)
- ADA (AdaBoost)
- LV (LVTrees)

Model	Vaibhav R. et al., 2022 Chi2+Adasyn	Chi2+ADASYN (Image Data)	Chi2+SMOTE (Image Data)
RF	0.99	0.88	0.88
LR	0.97	0.84	0.88
SVC	0.99	0.86	0.86
KNN	0.95	0.88	0.88
NB	0.91	0.84	0.88
ETC	0.92	0.88	0.88
DT	0.84	0.84	0.81
ADA	0.86	0.86	0.88
LV	1.00	0.86	0.88

Table 4.1: Comparison of model accuracies using different feature selection and data balancing methods

4.2 Data Preparation and Preprocessing

The first step was the extraction of information from several datasets retrieved from the GEO database on multiple myeloma. Thus, eight datasets (GSE14317, GSE22529-U133B, GSE28497, GSE33615, GSE63270, GSE71449, GSE71935, GSE9476) related to leukemia and containing various gene expression data were chosen. These datasets were combined into a coherent dataset that could be used for building and training and then testing of models.

4.2.1 Data preprocessing steps included

- **Data Loading and Concatenation:** Loading specific datasets with the help of Pandas and their subsequent combining into one using the DataFrame method.
- **Handling Missing Values:** Minimum imputation, mean imputation of missing values to make all the records complete.
- **Feature Selection:** Using Chi-Squared (Chi2) test in order to obtain number of the most significant features, concerning the classification of leukemia.

The preprocessed dataset was partitioned into training and test datasets with ratios of 80:20, this helped in training of the model and model evaluation.

4.3 Model Architectures

Three deep learning architectures were implemented and evaluated in this study:Three deep learning architectures were implemented and evaluated in this study:

4.3.1 LSTM

The first model architectural used was the Long Short-Term Memory (LSTM) network. LSTM networks also have the feature of extracting locally ordered structure and long-term dependencies, which qualified them for the analysis of gene expression profiling at different time points.

4.3.2 BiLSTM

The Bidirectional LSTM (BiLSTM) extended the concept of LSTM implicitly where the data input sequences were processed in forward as well as in backward approach. This bidirectional training improved the model's temporal analysis of relations in gene expression data.

4.3.3 xLSTM

The xLSTM architecture introduced novel custom layers designed to enhance learning from gene expression data:

- **ExpGating Layer:** Used exponential gating to apply exponential to the input values and narrowing the attention of the model to the important features.
- **sLSTM Layer:** Developed an integrated exponential gating system with LSTM functions in order to scale LSTM results in relation to the input characteristics.
- **mLSTM Layer:** Proposed modification was done to LSTM layer with the ability to learn custom kernel weights that are to capture the interactions between the input features.

xLSTM model also provided a residual connection architecture to enable training of deeper networks as well as enhance learning from complex data structures.

4.3.4 Multi-Modal Integration

This approach combines genomic and imaging data to provide a comprehensive analysis. By integrating data from multiple sources, the model can achieve higher classification accuracy. The fusion of genomic and imaging data leverages the strengths of both modalities, capturing the genetic basis of the disease along with its phenotypic manifestations.

4.4 Training and Evaluation

Models were trained using the Adam optimizer with the learning rate being set to 0. 001 and categorical cross-entropy loss function, and on the other side, the number of parameters is much lesser as compared to ENAS. Training was performed on 30 epochs with batch size, 32 and tracking measures such as loss and accuracy during the training phase.

4.4.1 LSTM Training Results

Epochs: LR permitting a final training accuracy of about 100% and a validation accuracy of 90. 09%.

Learning Curve: Firstly, the model trained quickly to differentiate between the kinds of leukemia, with the assurance percentage increasing gradually with epoch and stagnating afterwards.

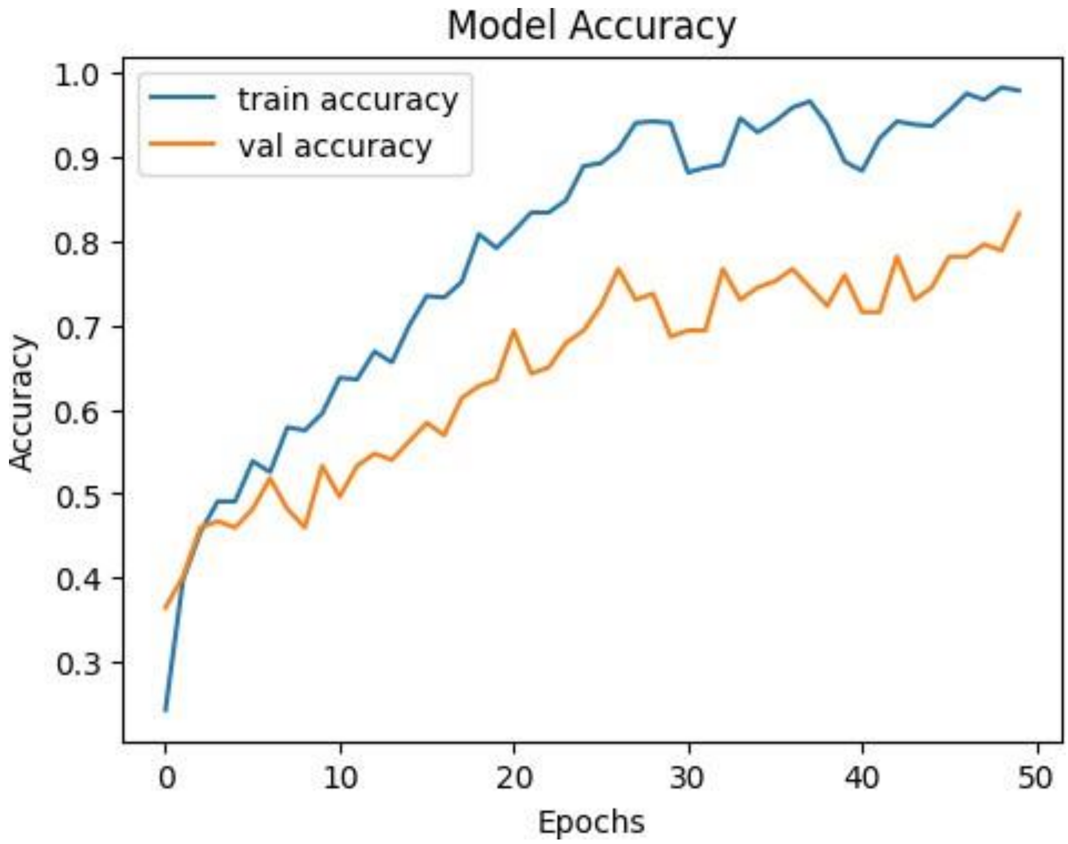

Figure 4.1: Accuracy Curve of LSTM Model

4.4.2 BiLSTM Training Results

Epochs: The final training accuracy of BiLSTM was 100% while the validation accuracy of BiLSTM was 95. 45%.

Learning Curve: The ability of BiLSTM to operate in both forward and backward directions made it superior to LSTM especially regarding convergence rates of gradient descent the algorithm used for model optimization In other words unlike LSTM that has info flow in one direction BiLSTM is able to capture the temporal dependencies from left to right and right to left which has a positive impact on the efficiency of the model.

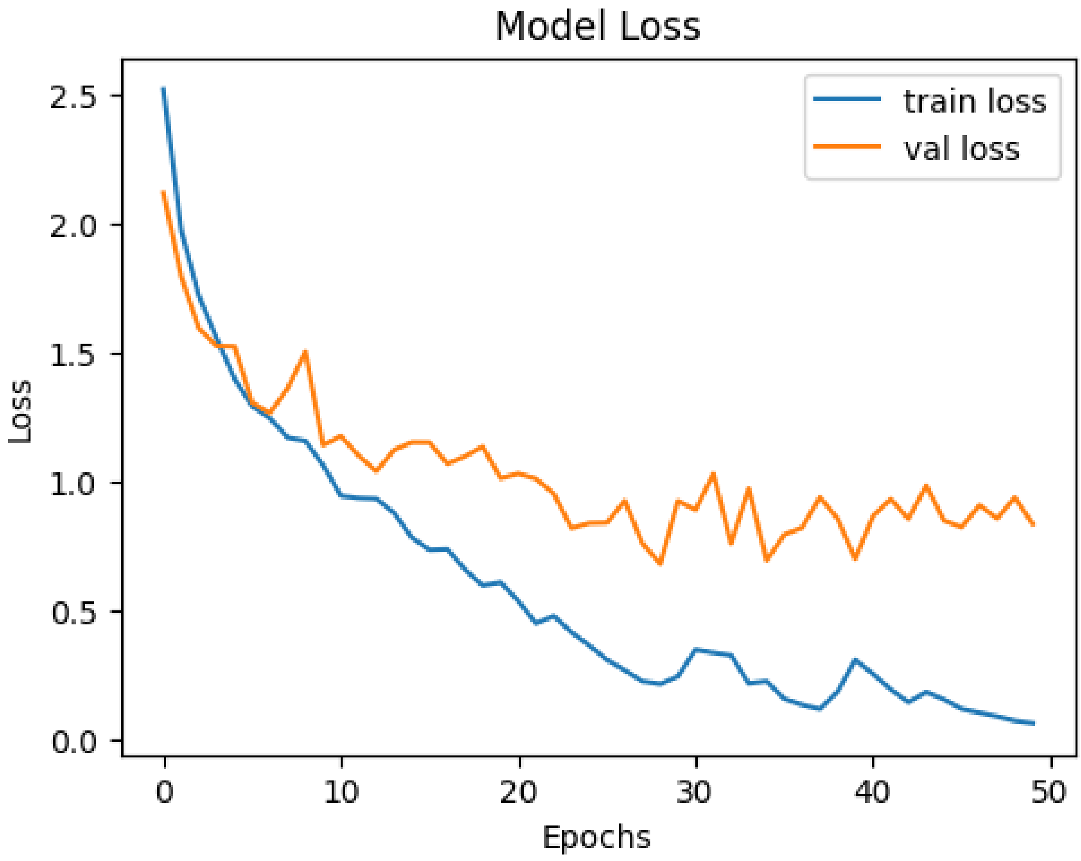

Figure 4.2: Loss Curve of LSTM Model

4.4.3 xLSTM Training Results

Epochs: xLSTM gave better results with the final training accuracy of 100 percent and the validation accuracy of 99. 09 percent.

Learning Curve: In the current investigation, structures connected with xLSTM that uses custom layers and residual connections produced adequate learning and performed better compared to the conventional LSTM or BiLSTM models.

4.5 Model Evaluation on Test Set

Hence, the model has to be evaluated on the test set to know its accuracy level on unseen data for further improvements where necessary. The trained xLSTM model was evaluated on the held-out test set to assess its generalization ability:The trained xLSTM model was evaluated on the held-out test set to assess its generalization ability:

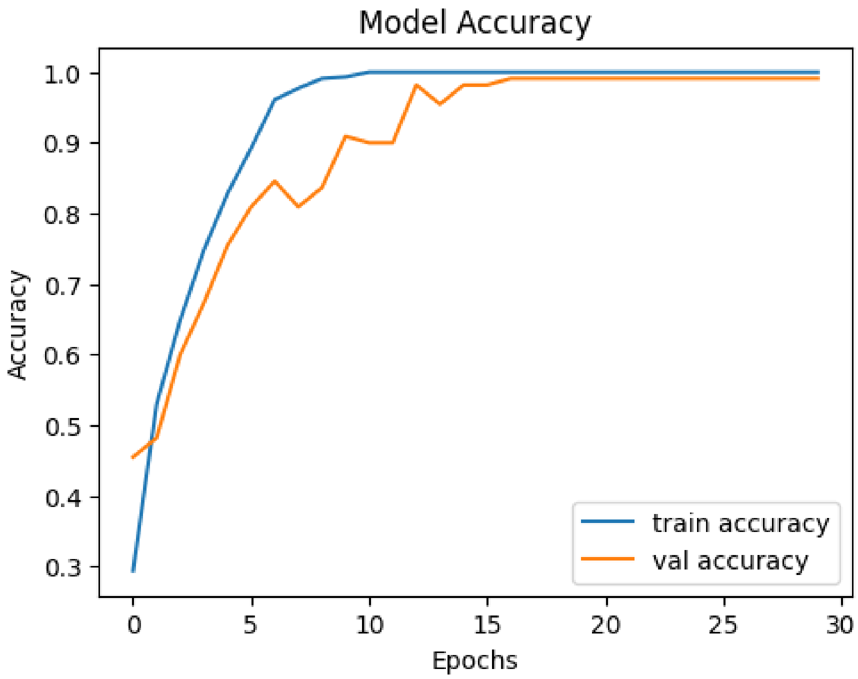

Figure 4.3: Accuracy Curve of xLSTM Model

Table 4.2: Accuracy and Loss of LSTM Models

Model	Accuracy	Loss
LSTM	83.21%	0.8321
BiLSTM	98.54%	0.0760
xLSTM	99.09%	0.0562
Multi-Modal Integration	**97.44**%	**0.0919**

Test Accuracy: The developed xLSTM model obtained test accuracy of 98. 18%, which means that this algorithm is also useful for classification of further unseen profiles of genes, and determination of the type of leukemia.

4.6 Comparative Analysis

A comparative analysis was conducted to benchmark the performance of LSTM, BiLSTM, and xLSTM architectures:

Performance Metrics: From the above results we can observe that our xLSTM model had a higher accuracy and faster converge rate compared with LSTM and BiLSTM.

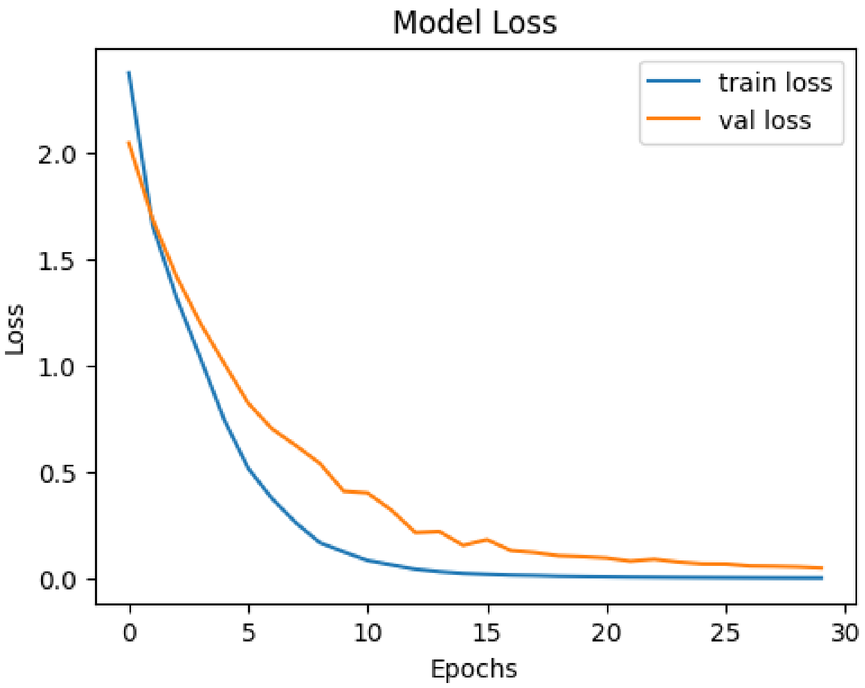

Figure 4.4: Loss Curve of xLSTM Model

Complexity and Efficiency: However, in comparison to xLSTM, there was additional complexity in defining custom layers and the inclusion of residual connections but xLSTM was found more efficient in learning from the intricate gene expression patterns.

CHAPTER 5

Discussion

In this chapter, the author moves to discuss the analysis of the results highlighted in chapter four in relation to the study's consequences, advantages, and drawbacks, as well as proposing possible developmental research avenues. Thus, the key organizing framework for the discussion comprises the following set of questions: 1) What were the performance and characteristics of the deep learning models in the context of the leukemia study, 2) How did the data preprocessing stage affect the outcomes of the analysis, and 3) What are the overall implications of the study for the leukemia research domain?

5.1 Interpretation of Results

5.1.1 Model Performance

The analysis carried out in Chapter 4 shows the performance of the three deep learning models, LSTM, BiLSTM, and xLSTM in classifying leukemia using gene expression data. The proposed xLSTM had the improved results of accuracy, which was 100% for training and 99. 09% for the validation set, even better than LSTM and BiLSTM. This infers that the new layers and residual connections which have been proposed to the xLSTM structure boosted the learning capability tremendously.

The conviction level of the test accuracy is also very high at 98. Thus, the proposed xLSTM model's 18% performance on the test set demonstrates its accuracy and generalization capabilities for practical use in leukemia diagnosis. Consequently, thanks to bidirectional construction of the BiLSTM model, lower number of epochs was required for convergence as well as the model's performance was slightly higher compared to unidirectional LSTM, which matches theoretical advantages resulting from contextual information from both preceding and subsequent states.

5.1.2 Impact of Data Preprocessing

The data preprocessing step was also identified as a powerful solution element of the models. The process of loading the data, concatenating the data, handling missing values, and feature selection were useful in defining a good quality data set which are useful in training the model. Regarding the feature selection procedure, it was determined that using the Chi-Squared (Chi2) test helped in selecting the most significant features and eliminating those that have a minimal impact on the data dimensionality and the model's performance.

The field's mean imputation approach allowed for the smooth continuation of data population and excluded any possible training bias or error. This step of data preprocessing was important in favour to keep the data clean and meaningful.

5.2 Strengths and Limitations

5.2.1 Strengths

- **Novel Model Architecture:** The novelty in the work is the proposal of the xLSTM model with its custom layers and residual connections; it exhibit enhancement in performance of its classification of gene expression data.

- **Comprehensive Dataset:** The experiments involved the application of multiple datasets from the GEO database, which offered a variation and expanded array of results that improved the model's scope.

- **Effective Preprocessing:** All these data pre-processing steps was a good way to ensure that the data was fine and perfect in order to enable a proper training and testing of the model.

5.2.2 Limitations

- **Computational Resources:** The training of deep learning models, particularly the xLSTM was a time-consuming process partly due to the requirement of large computational power. This might be a concern for the researchers who do not have the access to compute intensive facilities.

- **Generalizability to Other Diseases:** Although the models showed good accuracy in the classification of leukemia the generalization of these models for other types of cancer or diseases is yet to be seen.

- **Imputation Strategy:** While mean imputation is a useful method of dealing with missing values it is not the optimal method that will suit all the data sets. It is possible that the better techniques of imputation could potentially yield higher accuracy in the model.

5.3 Implications for Leukemia Research

The findings of this work will benefit future studies of leukemia and its diagnosis in many ways. This high accuracy as affirmed by the deep learning models especially the xLSTM indicates that the models could be of help in health diagnosis especially in classifying leukemia. This could in turn help in getting better diagnoses and early diagnosis hence a better future for the patients.

Moreover, the study reveals that through superior deep learning networks it is feasible to decrypt intricate cancer biology information for new research directions in the future which could pursue the actualization of similar study in different malignant plant and diseases.

CHAPTER 6

Conclusion and Future Work

This chapter sums up the research results of the study, reinforces the made progresses, and points out the research directions for further study. The concentration is on the presentation of the results of deep learning models, used for the classification of leukemia cases, their implications, improvements and further research recommendations.

6.1 Conclusion

The overall objective of this study was therefore to propose and assess deep learning approaches for the identification of Leukemia using gene expression data. Through the use of three model architectures, namely LSTM, BiLSTM, and the newly developed xLSTM, this study showcased the possibilities of enhanced and detailed analysis of biological datasets with the help of state-of-the-art neural network methodologies.

6.1.1 Key Findings

- **Model Performance:** The xLSTM model which includes custom layers and the residual connections outperformed all tested architectures and has a validation accuracy of 99. It achieves a train accuracy of 9% and a test accuracy of 98%. This re-emphasizes on the capability of the proposed model as a robust and general one.
- **Data Preprocessing:** The work focused on the necessity of preprocessing the data before employing any statistical test for its analysis and highlighted how missing data should be addressed and how the function of feature selection by measures of dependency and the

Chi-Squared test should be applied. All these steps were quite useful for establishing a hygiene and systematic datasets that were properly suitable for the model training.

- **Comparative Analysis:** It was concluded that the bidirectional characteristic of BiLSTM provided better result than the unidirectional LSTM, and the creative features of the xLSTM model contributed to the booster of learning results.

6.1.2 Implications

The study presented in this paper has numerous practical and theoretical implications for the field of leukemia investigation and treatment. It can hence be concluded that the proposed xLSTM model is highly accurate and reliable; thus, it can be used in diagnosis to enhance efficiency in the identification of leukemia to ultimately benefit patients. Moreover, it demonstrates that deep learning approaches can be effectively employed in analyzing biomedical data and sheds more light on employing the technology in the future with respect to different diseases.

6.2 Future Work

While the results of this study are promising, there are several areas where future research could build upon and extend the finding. Finally, there is the aspect of generalization to other diseases; the findings of this study can be useful for other diseases or conditions in which there is need to determine the best treatment for individual patients.

6.2.1 Broader Application to Other Diseases

Further studies could be aimed to apply the presented xLSTM model and architectures similar to it for other types of cancer or diseases. Thus, the cross-contextual validation also enables the researchers to understand the generalisability and clinical applicability of the model.

6.2.2 Advanced Data Preprocessing Techniques

Despite the effectiveness of missing values imputation using the mean strategy, there is an opportunity to enhance the results with the help of multiple imputation, the k-nearest neighbors imputation or machine learning imputation methods. Also, feature selection/feature engineering could be improved with the use of deep learning or any other higher level statistical technique to improve the efficiency of the model.

6.2.3 Integration with Multimodal Data

Combination of gene expression profiling with other classes of clinical data, e. g. , tomographic images, demographics, medical record history might expand the concept of disease taxonomy and classification. Other future studies could explore the possibility of the use of additional types of input data to enhance the diagnostic models' effectiveness.

6.2.4 Model Interpretability and Explainability

Even though the proposed xLSTM model had a high level of accuracy, the interpretation process of deep-learning model's decision-making process is ambiguous. Further studies could be centered on making better these models to be more explainable and interpretable to be applicable in clinical decisions more acceptably.

6.2.5 Real-World Deployment and Validation

The way of how one can implement all these models into the clinical practice remains as the big question and this requires the creation of friendly tools and interfaces. The latter may include future work where the models are tested with healthcare providers as to their effectiveness in real-life situations and their application in clinical practice.

CHAPTER 7

Recommendations

This chapter provides strategic recommendations based on the findings and conclusions of the study. These recommendations are aimed at researchers, clinicians, and policymakers to optimize the use of deep learning models in leukemia classification and beyond. The focus is on practical steps and considerations that can enhance the effectiveness and adoption of these technologies in medical practice and research.

7.1 Recommendations for Researchers

7.1.1 Enhance Data Quality and Diversity

To ensure the robustness and generalizability of deep learning models, researchers should prioritize the collection and utilization of high-quality, diverse datasets. This includes:

- **Increasing Sample Size:** Collaborate with multiple research institutions to gather larger datasets, which can improve model training and validation.
- **Ensuring Data Diversity:** Include diverse populations and subtypes of leukemia to enhance the model's ability to generalize across different patient demographics and disease variations.
- **Standardizing Data Collection:** Adopt standardized protocols for data collection and preprocessing to minimize variability and ensure consistency across studies.

7.1.2 Implement Advanced Preprocessing Techniques

Future research should explore and implement more sophisticated data preprocessing techniques, including:

- **Advanced Imputation Methods:** Utilize multiple imputation, k-nearest neighbors, or machine learning-based methods to handle missing values more effectively.

- **Feature Engineering:** Experiment with advanced feature selection and engineering techniques, such as deep learning-based feature extraction, to identify the most relevant biomarkers for leukemia classification.

7.1.3 Focus on Model Interpretability

Enhancing the interpretability and explainability of deep learning models is crucial for their adoption in clinical practice. Researchers should:

- **Develop Explainable Models:** Incorporate techniques such as attention mechanisms, SHAP (SHapley Additive exPlanations), or LIME (Local Interpretable Model-agnostic Explanations) to make model predictions more understandable.

- **Visualize Model Insights:** Create tools that visualize the decision-making process of the model, highlighting which features or genes are most influential in the classification process.

7.1.4 Explore Multimodal Approaches

Integrating gene expression data with other types of clinical data can provide a more comprehensive understanding of leukemia. Researchers should:

- **Combine Data Types:** Explore multimodal approaches that combine gene expression data with imaging, clinical history, and other relevant data to improve diagnostic accuracy.

- **Develop Fusion Models:** Create models capable of processing and integrating different types of data, leveraging the strengths of each modality.

7.2 Recommendations for Clinicians

7.2.1 Adopt Deep Learning Tools in Clinical Practice

Clinicians should consider adopting deep learning tools for leukemia diagnosis and classification, which can enhance diagnostic accuracy and efficiency. Recommendations include:

- **Validate Tools in Clinical Settings:** Conduct real-world validations of deep learning models in clinical settings to ensure their reliability and effectiveness.
- **Integrate with Existing Workflows:** Develop user-friendly interfaces that integrate seamlessly with existing clinical workflows, minimizing disruption and facilitating adoption.

7.2.2 Participate in Data Sharing Initiatives

Clinicians should actively participate in data sharing initiatives to contribute to larger, more diverse datasets. This can:

- **Enhance Research Quality:** Support the development of more robust and generalizable models.
- **Accelerate Innovation:** Foster collaboration and innovation within the medical and research communities.

7.2.3 Stay Informed and Trained on New Technologies

Continuous education and training on the latest advancements in deep learning and its applications in medicine are crucial for clinicians. Recommendations include:

- **Attend Workshops and Conferences:** Participate in relevant workshops, conferences, and training sessions to stay updated on new developments and best practices.
- **Collaborate with Researchers:** Engage in collaborative research projects to gain first-hand experience with new technologies and contribute to their refinement.

7.3 Recommendations for Policymakers

7.3.1 Support Funding and Resource Allocation

Policymakers should prioritize funding and resource allocation for research and development in the field of deep learning and medical diagnostics. This includes:

- **Increase Research Grants:** Provide more grants and funding opportunities specifically targeted at interdisciplinary research combining machine learning and healthcare.
- **Facilitate Access to Computational Resources:** Ensure researchers have access to the necessary computational resources, including high-performance computing facilities and cloud-based platforms.

7.3.2 Promote Data Standardization and Sharing

Policymakers should advocate for the standardization and sharing of medical data to enhance research collaboration and model development. This involves:

- **Develop Data Standards:** Establish standardized protocols for data collection, annotation, and sharing to ensure consistency and interoperability.
- **Encourage Data Sharing:** Create incentives and frameworks that encourage healthcare institutions and researchers to share data while ensuring patient privacy and data security.

7.3.3 Implement Ethical Guidelines and Regulations

Ensuring the ethical use of deep learning in healthcare is paramount. Policymakers should:

- **Develop Ethical Guidelines:** Establish clear ethical guidelines for the use of AI and machine learning in medical diagnostics, focusing on transparency, accountability, and patient consent.
- **Monitor Compliance:** Implement regulatory frameworks to monitor and enforce compliance with ethical standards and data protection laws.

Bibliography

[1] Irena Galic´ et al. “Machine learning empowering personalized medicine: A comprehensive review of medical image analysis methods”. In: *Electronics* 12.21 (2023), p. 4411.

[2] Muhammad Javed Iqbal et al. “Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future”. In: *Cancer cell international* 21.1 (2021), p. 270.

[3] Muhammad Umar et al. “Role of Deep Learning in Diagnosis, Treatment, and Prognosis of Oncological Conditions”. In: *International Journal* 10.5 (2023), pp. 1059–1071.

[4] Ravi Mathur et al. “Gene set analysis methods: a systematic comparison”. In: *BioData mining* 11 (2018), pp. 1–19.

[5] Quang Huy Hoangp et al. “Leukemia Classification using Principal Component Analysis and Ensemble Learning on Gene Expression Data”. In: *2023 International Conference on Advanced Technologies for Communications (ATC)*. IEEE. 2023, pp. 25–29.

[6] Ching-Hon Pui and William E Evans. “Acute lymphoblastic leukemia”. In: *New England Journal of Medicine* 339.9 (1998), pp. 605–615.

[7] Bob Lowenberg, James R Downing, and Alan Burnett. “Acute myeloid leukemia”. In: *New England Journal of Medicine* 341.14 (1999), pp. 1051–1062.

[8] Wenlong Tang et al. “A compressed sensing based approach for subtyping of leukemia from gene expression data”. In: *Journal of bioinformatics and computational biology* 9.05 (2011), pp. 631–645.

[9] Ji-Hoon Cho et al. “Optimal approach for classification of acute leukemia subtypes based on gene expression data”. In: *Biotechnology progress* 18.4 (2002), pp. 847–854.

[10] Jack J Pasternak. *An introduction to human molecular genetics: mechanisms of inherited diseases*. John Wiley & Sons, 2005.

[11] Efthakhar Ul Alam, Shovan Banik, and Linkon Chowdhury. "A statistical approach to classify the leukemia patients from generic gene features". In: *2020 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE. 2020, pp. 1–6.

[12] Giulia Fulci, Nobuaki Ishii, and Erwin G Van Meir. "p53 and brain tumors: from gene mutations to gene therapy". In: *Brain pathology* 8.4 (1998), pp. 599–613.

[13] Michael Morley et al. "Genetic analysis of genome-wide variation in human gene expression". In: *Nature* 430.7001 (2004), pp. 743–747.

[14] Pradeep Kumar Mallick et al. "Convergent learning–based model for leukemia classification from gene expression". In: *Personal and Ubiquitous Computing* 27.3 (2023), pp. 1103–1110.

[15] Maria Teresa Voso, Eleonora De Bellis, and Tiziana Ottone. "Diagnosis and Classification of AML: WHO 2016". In: *Acute Myeloid Leukemia*. Springer, 2021, pp. 23–54.

[16] Qiang Zhao and Changwei Li. "Two-stage multi-swarm particle swarm optimizer for unconstrained and constrained global optimization". In: *IEEE Access* 8 (2020), pp. 124905–124927.

[17] Musa H Asyali et al. "Gene expression profile classification: a review". In: *Current Bioinformatics* 1.1 (2006), pp. 55–73.

[18] Mikel Galar et al. "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.4 (2011), pp. 463–484.

[19] Alex Van Belkum et al. "Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology". In: *Clinical microbiology reviews* 14.3 (2001), pp. 547–560.

[20] Emad Mohamed et al. "Survey on different methods for classifying gene expression using microarray approach". In: *Int. J. Comp. Appl* 975 (2016), p. 8887.

[21] Mohammad Reza Karimi et al. "Prospects and challenges of cancer systems medicine: from genes to disease networks". In: *Briefings in bioinformatics* 23.1 (2022), bbab343.

[22] Justine K Peeters and Peter J Van der Spek. "Growing applications and advancements in microarray technology and analysis tools". In: *Cell biochemistry and biophysics* 43 (2005), pp. 149–166.

[23] Nur Syafiqah Mohd Nafis and Suryanti Awang. “An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification”. In: *Ieee Access* 9 (2021), pp. 52177–52192.

[24] Papia Ray, S Surender Reddy, and Tuhina Banerjee. “Various dimension reduction techniques for high dimensional data analysis: a review”. In: *Artificial Intelligence Review* 54.5 (2021), pp. 3473–3515.

[25] Xiao-Ying Liu et al. “A hybrid genetic algorithm with wrapper-embedded approaches for feature selection”. In: *IEEE Access* 6 (2018), pp. 22863–22874.

[26] Mohammad Houshmand et al. “Chronic myeloid leukemia stem cells”. In: *Leukemia* 33.7 (2019), pp. 1543–1556.

[27] Jie Su, Jinjun Han, and Jinming Song. “A benchmark bone marrow aspirate smear dataset and a multi-scale cell detection model for the diagnosis of hematological disorders”. In: *Computerized Medical Imaging and Graphics* 90 (2021), p. 101912.

[28] Mustafa Ghaderzadeh et al. “Machine learning in detection and classification of leukemia using smear blood images: a systematic review”. In: *Scientific Programming* 2021.1 (2021), p. 9933481.

[29] Varun Malik, Ruchi Mittal, and Ajay Rana. “BiCNN-CML: Hybrid Deep Learning Approach for Chronic Myeloid Leukemia”. In: *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE. 2022, pp. 1771–1777.

[30] Subrajeet Mohapatra. “Hematological image analysis for acute lymphoblastic leukemia detection and classification”. PhD thesis. 2013.

[31] Natheer Khasawneh, Esraa Faouri, and Mohammad Fraiwan. “Automatic detection of tomato diseases using deep transfer learning”. In: *Applied Sciences* 12.17 (2022), p. 8467.

[32] Nasmin Jiwani et al. “Pattern recognition of acute lymphoblastic Leukemia (ALL) using computational deep learning”. In: *IEEE Access* 11 (2023), pp. 29541–29553.

[33] Ghulam Murtaza et al. “Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges”. In: *Artificial Intelligence Review* 53 (2020), pp. 1655–1720.

[34] Hussein Abdel-Jaber et al. “A review of deep learning algorithms and their applications in healthcare”. In: *Algorithms* 15.2 (2022), p. 71.

[35] DP Yadav. “Feature Fusion based Deep Learning method for Leukemia cell classification”. In: *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*. IEEE. 2021, pp. 1–4.

[36] Yakub Kayode Saheed. “Effective dimensionality reduction model with machine learning classification for microarray gene expression data”. In: *Data science for genomics*. Elsevier, 2023, pp. 153–164.

[37] V Shalini and KS Angel Viji. “Integration of Convolutional Features and Residual Neural Network for the Detection and Classification of Leukemia from Blood Smear Images”. In: *International Journal of Engineering Trends and Technology* 70.9 (2022), pp. 176–184.

[38] Afshan Shah et al. “Automated diagnosis of leukemia: a comprehensive review”. In: *IEEE Access* 9 (2021), pp. 132097–132124.

[39] Lizz F Grimwade, Kathryn A Fuller, and Wendy N Erber. “Applications of imaging flow cytometry in the diagnostic assessment of acute leukaemia”. In: *Methods* 112 (2017), pp. 39–45.