ACCELERATING DRUG DISCOVERY FOR CHAGAS DISEASE USING DEEP LEARNING

GIDEON SIGILAI
ICT-G-4-1572-21

A RESEARCH PROJECT SUBMITTED TO THE SCHOOL OF COMPUTING AND INFORMATICS IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE IN BACHELOR OF SCIENCE IN COMPUTER SCIENCE AT GRETSA UNIVERSITY

DECEMBER 2024

STUDENT DECLARATION

This research is my original work and has not been presented for the award of a degree or any similar purpose in any other institution.

Signature: _____ Date: 29/11/24

Gideon Sigilai

ICT-G-4-1572-21

Supervisor: This proposal has been submitted with my approval as university supervisor

Signature: _____ Date: 29/11/2024.

for Dennis Wapukha

Lecturer, School of Computing and Informatics

Gretsa University

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATION AND ACRONYMS

ML	-	Machine Learning
AI	-	Artificial Intelligence
DL	-	Deep Learning
CNN	-	Convolutional Neural Networks
RNNs	-	Recurrent Neural Networks
CGNs	-	Graph Convolutional Networks
OMIM	-	Online Mendelian Inheritance in Man
NCBI	-	National Center for Biotechnology Information
KEGG	-	Kyoto Encyclopedia of Genes and Genomes

OPERATIONAL DEFINITIONS OF TERMS

Chagas Disease: is a tropical parasitic disease caused by Trypanosoma cruzi. It is spread mostly by insects in the subfamily Triatominae

Trypanosoma cruzi.: is a species of parasitic euglenoids that are responsible for the Chagas disease.

Data Augmentation: is the process of enhancing a dataset by creating modified versions of existing data to improve model performance and generalization.

Deep learning is a branch of machine learning that uses artificial neural networks with many layers to automatically learn and make predictions from large amounts of data.

Neglected Diseases Diseases that predominantly affect low-income populations and receive limited research and development attention

Drug Discovery The process of identifying new therapeutic compounds or drugs for the treatment of diseases

Convolutional Neural Networks (CNNs) A type of deep learning model particularly effective for analyzing visual or spatial data.

Recurrent Neural Networks (RNNs) A deep learning model designed to process sequential or time-series data.

Training Dataset A collection of data used to teach a machine learning model how to make predictions

Computational Resources The hardware and software infrastructure, such as GPUs or cloud computing, required to train and run deep learning models.

ABSTRACT

Neglected diseases remain a significant threat to global health, particularly in low-income countries, where limited resources and funding hinder drug discovery and treatment efforts. These diseases, which disproportionately affect impoverished populations, often receive little attention from the pharmaceutical industry due to their lack of profitability. Consequently, there is an urgent need to develop innovative and cost-effective approaches to accelerate the discovery of therapeutic solutions. Recent advances in computational science, particularly deep learning, have demonstrated remarkable potential in expediting drug discovery by predicting chemical activity and identifying new drug candidates efficiently. This study aims to leverage cutting-edge deep learning techniques to develop models capable of accurately predicting chemical activity and uncovering novel therapeutic options for neglected diseases such as Chagas disease. The research focuses on exploring various deep learning architectures, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), as well as incorporating advanced data pre-processing techniques to improve model accuracy and performance. Large and complex datasets containing drug activity information for neglected diseases will be utilized to train and evaluate the proposed models. The study will assess the impact of model choice, computational resources, and data quality on the speed and precision of drug candidate identification. By optimizing these factors, the research aims to create a framework for faster and more cost-effective drug discovery processes. The outcomes of this study have the potential to significantly enhance the pipeline for identifying treatments for neglected diseases, reducing the time and cost associated with drug development. Ultimately, this research could contribute to alleviating the suffering of millions of people affected by these overlooked diseases, offering hope for better health outcomes in underserved populations.

CHAPTER ONE: INTRODUCTION

1.1 Introduction

The burden of neglected diseases continues to weigh heavily on global health, particularly in under-resourced regions where access to innovative medical treatments is severely limited. Among these diseases, Chagas disease, caused by the protozoan parasite Trypanosoma cruzi, stands out not only for its widespread prevalence in Latin America but also for the profound impact it has on the lives of those affected. Despite advancements in medical technology and an increased understanding of disease mechanisms, the pharmaceutical industry has largely overlooked these conditions due to a perceived lack of profitability, resulting in inadequate research and a shortage of effective treatments.

Chagas disease is a glaring example of health inequity, as it predominantly affects vulnerable populations in low-income countries. The complexities of its transmission—primarily through insect vectors, but also via blood transfusions and organ transplants—highlight the multifaceted challenges of tackling this public health crisis. Left unaddressed, Chagas disease leads to severe complications, including significant cardiovascular and gastrointestinal issues, profoundly impacting the quality of life of affected individuals.

In light of these challenges, there is an urgent call for innovative approaches to drug discovery that can effectively target neglected diseases. Harnessing advancements in computational science and deep learning offers a promising avenue for accelerating the identification of therapeutic solutions. This chapter sets the stage for a deeper exploration of the intersection between cutting-edge technology and the fight against neglected diseases, emphasizing the necessity for swift and cost-effective strategies in developing treatments that could transform the lives of millions affected by Chagas disease and similar ailments. Through this research, we aim to illuminate the path towards a more equitable future in global health, where innovative therapies are accessible to those in greatest need.

1.2 Background to the Study

Despite global efforts, Neglected diseases such as Chagas disease, leprosy, and African sleeping sickness continued to pose significant health challenges. These diseases

affected poor and middle-class countries more. Unfortunately, the pharmaceutical industry often neglected these conditions due to the lack of profitability in developing drugs for them. The slow progress in addressing these neglected illnesses resulted in severe health consequences for affected populations.

The Trypanosoma cruzi parasite was responsible for Chagas disease, which was widespread in the Americas. The parasite was transmitted to humans and animals through insect vectors, contaminated blood transfusions, or organ transplants. Neglecting to address the disease could result in significant issues with the respiratory and gastrointestinal systems. Chagas disease predominantly affected a significant population of people who lived in Latin America. Dias, J. C. P. (1992). In S. Wendel, Z. Brener, M.E. Camargo, & A. Rassi (Eds.). Nevertheless, cases of the illness were also noticed in Kenya and other areas within the African continent.

To overcome the challenges associated with drug discovery for neglected diseases like Chagas, the application of advanced technologies such as deep learning was proposed as a potential solution. Deep learning operates by utilizing artificial neural networks to analyze extensive data sets and scrutinize vast amounts of information, categorizing it as a category within machine learning. This was good for dealing with tons of different data. By leveraging deep learning, the development of new drugs for neglected diseases could be completed with a markedly lower investment of both time and resources.

The main aim of this study was to investigate how deep learning could be employed to accelerate the process of discovering drugs for overlooked illnesses, concentrating particularly on the treatment of Chagas disease. The study evaluated the performance of deep learning models in predicting compound activity and identifying potential drug candidates. The research was conducted at the Kenyatta National Hospital, the largest referral hospital in Kenya, where deep learning's effectiveness in identifying potential drug candidates for Chagas disease was assessed.

The study utilized a strategy of investigating specific instances, utilizing the collation and examination of data obtained from individuals who had either been confirmed to have contracted Chagas disease or were believed to have contracted the disease. By working together with the hospital laboratory, we had the opportunity to obtain crucial information such as patients' healthcare records and the outcomes of laboratory tests.

This data was utilized to train and test deep-learning models for predicting compound activity and identifying potential drug candidates.

This study aimed to demonstrate that deep learning could accelerate drug discovery for neglected diseases, contributing to improved health outcomes for affected populations. It sought to contribute to the expanding knowledge on the application of machine learning in healthcare and pharmaceutical research, specifically in the context of neglected diseases.

In summary, this research endeavored to apply advanced deep learning techniques to expedite drug discovery for Chagas disease, aiming to improve health outcomes for affected populations. The study used a computer technique called deep learning to discover new drugs. It added to the research on using machines to help with healthcare and medicine. The research provided valuable insights into the ability of deep learning to accelerate drug discovery for neglected diseases, ultimately contributing to better health outcomes for affected populations.

1.3 Statement of Research Problem

Neglected diseases posed a significant public health concern for countries with lower and middle-income levels. Despite the severity of the crisis, the discovery of drugs for neglected diseases had been slow, mainly due to insufficient funding and lack of interest from pharmaceutical companies and consequently the need for new techniques to be used to rapidly discover drugs for neglected diseases.

In recent years, deep learning has shown remarkable success in drug discovery, and its application to overlooked diseases was an area for further exploration. Deep learning systems could analyze data on large compounds and identify those most likely to have therapeutic effects. The problem was the slow demand for drugs to treat neglected diseases.

The issue of this particular importance deserves attention due to the substantial harm and death that comes with these diseases. Traditional drug discovery methods failed to offer new treatments for many neglected diseases, and new methods were urgently needed.

This problem stemmed from many factors, including a lack of financial incentives

for pharmaceutical companies to invest in overlooked areas, insufficient funds, limited understanding of drug discovery in both high-income and low-income nations, and the complexity associated with neglected diseases.

Literature suggested that deep learning had the potential to identify overlooked therapies rapidly. However, its application was still in its infancy on neglected diseases, and more research was still needed in this area. Therefore, the main focus of this research was to investigate the benefits of using deep learning to process discovering drugs for illnesses that had been neglected, with a case study in the Kenyatta National Hospital.

1.4 Purpose of the Study

The research aimed to explore the application of deep learning in drug discovery for neglected diseases and examine the associated challenges and limitations.

The main objective of the research was to find the possibility of using advanced deep learning algorithms to improve the effectiveness of drug discovery procedures that were focused on neglected diseases. Additionally, the research sought to address any gaps in knowledge concerning this topic.

1.4 Conceptual Framework

The Conceptual framework below is broken down into two parts: dependent variables and Independent Variable

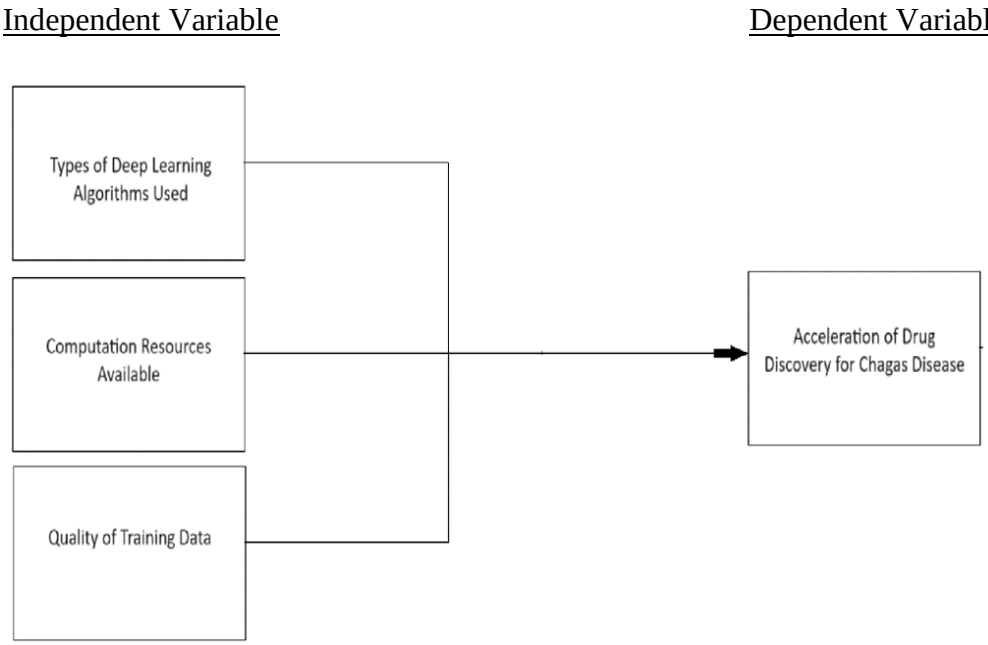

Figure 1 Conceptual Framework

1.5 Research Questions

i. What types of deep learning algorithms are most effective in accelerating drug discovery for Chagas disease?

ii. What is the impact of computation resources (e.g., GPU capacity, cloud computing access) on the effectiveness of deep learning-based drug discovery for Chagas disease?

iii. How does the quality and quantity of training data influence the accuracy and reliability of deep learning models in predicting drug candidates for Chagas disease?

1.6 Objectives of the Study

1.6.1 General Objective

To investigate the potential of deep learning techniques to accelerate drug discovery for neglected diseases, with a focus on Chagas disease at Kenyatta National Hospital.

1.6.2 Specific Objectives

I. To investigate the potential of deep learning in drug discovery for neglected diseases.
II. To identify potential drug candidates for neglected diseases using deep learning.
III. Investigate the role of data augmentation techniques in improving the performance of deep learning models for drug discovery in neglected diseases.
IV. Assess the potential of integrating multi-modal data sources, such as genomic, proteomic, and chemical data, in enhancing the predictive capabilities of deep learning models for drug discovery in neglected diseases.

1.7 Hypotheses of the Study

Ho1: Advanced deep learning models, such as generative adversarial networks, significantly outperform traditional models in predicting drug candidates for Chagas disease.

Ho2: Greater computational resources, such as access to high-performance GPUs, lead to faster training times and improved accuracy of deep learning models for drug discovery.

Ho3: High-quality and diverse training datasets significantly enhance the predictive accuracy of deep learning models in identifying potential drug targets for Chagas disease.

1.8 Significance of the Study

The significance of that look was multifaceted. Firstly, the findings of that examination offered insights into the utility of deep learning strategies in drug discovery for unnoticed illnesses, which could noticeably accelerate the drug development procedure. This could ultimately lead to the invention of drugs that were more powerful for the treatment of unnoticed sicknesses.

Secondly, the examination contributed to the body of knowledge on the use of deep mastering algorithms inside the healthcare sector, which was a fairly new and unexpectedly growing field. The results of that have a look at may have had implications for destiny studies and improvement in this vicinity.

Thirdly, the take-a-look was useful to healthcare practitioners and policymakers because it provided valuable records on the use of deep getting-to-know algorithms in drug discovery. The findings of that examination could inform the development of guidelines and suggestions that promote the use of these strategies in drug discovery for neglected illnesses.

Fourthly, the study was of benefit to the broader community, especially to those in developing nations who were disproportionately affected by neglected illnesses. The development of new drugs for ignored illnesses and the use of deep knowledge of techniques could have had a massive effect on the health and well-being of those groups and ultimately contributed to the success of sustainable development goals.

1.9 Delimitations or Scope of the Study

This study focused solely on using deep learning strategies to accelerate drug discovery for omitted illnesses at the Kenyatta National Hospital. The observation did not cover the whole drug discovery process but rather dealt with the identification of potential drug candidates and the usage of machine learning algorithms. The research considered the scope of the National Hospital as the study area, limiting itself to the available data within the organization.

1.10 Limitations of the Study

This study had some limitations, such as the potential unavailability of comprehensive data due to privacy issues and resource constraints for data gathering. Furthermore, there may have been limitations to the generalizability of the findings to other healthcare settings and hospital types outside the scope of this review, limitations, and recognition of the possibility of a general case study in the discussion section.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

Within this chapter, we shall examine the body of written works that discuss the implementation of deep learning within the realm of pharmaceutical research specifically for illnesses that have been overlooked. The chapter will be structured in line with the objectives outlined in Chapter One.

Deep learning is a subset of artificial intelligence that has shown great promise in accelerating drug discovery for neglected diseases. For example, DeepMind's AlphaFold program was able to achieve the same accuracy in mapping protein structures in a matter of days, compared to years of work by scientists in laboratories. The capability of this technology to bridge our gaps in comprehending biology and hasten scientific research could lead to the development of novel and efficacious therapies for various illnesses. One area where deep learning has shown promise is in the discovery of new drugs for neglected tropical diseases. These diseases affect millions of individuals and cause loss of productivity worldwide. They are common in developing countries without the financial resources for research and drug development. Machine learning has been introduced into the drug discovery process with the increased availability of data from high throughput screening. Models can be trained to predict the biological activities of compounds before working in the lab.

In this chapter, we'll analyze the potential of utilizing deep learning for the discovery of medicines to treat overlooked illnesses. Furthermore, we will use deep learning to pinpoint promising drug options. We will also evaluate publicly available datasets related to neglected diseases and develop and optimize deep-learning models for drug discovery. Finally, we will evaluate the performance of these models in predicting drug efficacy and toxicity for neglected diseases.

2.2 Review of literature related to the main concept

The primary focus of this research study involves utilizing deep learning models to aid in the discovery of drugs intended for use in the treatment of oft-

overlooked illnesses. This review of literature aims to explore various topics including the application of deep learning in drug

discovery, ways to enhance data through augmentation techniques, and the integration of different data sources to enable multi-modal analysis.

2.2.1 Deep Learning in Drug Discovery

Deep learning, a subset of machine learning, has gained significant attention in recent years due to its ability to learn complex patterns and representations from large datasets (LeCun et al., 2015). In the context of drug discovery, deep learning models have been applied to various tasks, such as predicting drug-target interactions, compound-protein interactions, and drug response in cell lines (Chen et al., 2018).

Several studies have demonstrated the potential of deep learning models in improving the drug discovery process. For instance, Wallach et al. (2015) developed a deep learning model for predicting molecular bioactivity and showed that it outperformed traditional machine learning methods. Similarly, Gómez-Bombarelli et al. (2018) used a deep learning model to predict the solubility and synthesizability of small molecules, which are important factors in drug development.

2.2.2 Data Augmentation Techniques

Data augmentation techniques have been widely used in various domains, including computer vision and natural language processing, to increase the size and diversity of training datasets, thereby improving the performance of deep learning models (Shorten & Khoshgoftaar, 2019). In the context of drug discovery, data augmentation techniques can be applied to generate new data points by introducing variations in the existing data, such as molecular structure modifications, chemical property perturbations, or simulated experimental conditions (Ekins et al., 2019).

Multiple researches have shown that incorporating data augmentation methods can enhance the efficiency of deep learning algorithms for drug exploration.

For instance, Goh et al. (2017) applied data augmentation techniques to generate additional training data for predicting drug-target interactions and observed significant improvements in model performance. Similarly, Chen et al. (2018) used data augmentation to enhance the performance of deep learning models for predicting compound-protein interactions.

2.2.3 Integration of Multi-modal Data Sources

Integrating multi-modal data sources, such as genomic, proteomic, and chemical data, has the potential to enhance the predictive capabilities of deep learning models for drug discovery in neglected diseases. Multi-modal data integration can provide a more comprehensive understanding of the disease biology and drug-target interactions, leading to more accurate predictions of drug efficacy and toxicity (Subramanian et al., 2017).

Several studies have explored the integration of multi-modal data sources in drug discovery using deep learning models. For example, Zitnik et al. (2018) developed a graph convolutional neural network that integrated genomic, proteomic, and chemical data to predict drug-target interactions and drug repurposing opportunities. Similarly, Ma et al. (2018) proposed a multi-modal deep learning framework that combined gene expression, chemical structure, and protein-protein interaction data to predict drug response in cancer cell lines.

These studies suggest that integrating multi-modal data sources can enhance the performance of deep learning models for drug discovery in neglected diseases, providing a more comprehensive understanding of the underlying biological processes and improving the accuracy of drug efficacy and toxicity predictions.

2.3 Deep learning in drug discovery for neglected diseases.

The first objective of this study is to identify and evaluate publicly available datasets related to neglected diseases, including information on disease biology, genetic and molecular interactions, and potential drug targets.

Understanding the biology and genetic information of neglected diseases is crucial for identifying potential drug targets and developing effective

treatments. Several databases provide comprehensive information on the genetic and molecular aspects of neglected diseases. For example, the Online Mendelian Inheritance in Man (OMIM) database offers a catalog of human genes and genetic disorders, including those related to neglected diseases (Amberger et al., 2019). Additionally, the National Center for Biotechnology Information (NCBI) Gene database provides gene-specific information for various organisms, including those causing neglected diseases (Brown et al., 2015).

Identifying molecular interactions and drug targets is essential for drug discovery in neglected diseases. The Protein Data Bank (PDB) is a valuable resource for obtaining structural information on proteins and their interactions with other molecules, including potential drug targets (Berman et al., 2000). The Kyoto Encyclopedia of Genes and Genomes (KEGG) database offers a collection of molecular interaction networks, including metabolic pathways and drug-target interactions, which can be useful for understanding the molecular basis of neglected diseases (Kanehisa et al., 2017).

Several databases provide data specifically for drug discovery in neglected diseases. As mentioned earlier, the ChEMBL database contains bioactivity data for drug-like molecules, which can be used to identify potential drug candidates (Gaulton et al., 2017). The DrugBank database offers comprehensive information on drug targets and their associated diseases, including neglected diseases (Wishart et al., 2018). Additionally, the PubChem database provides a vast repository of chemical information, including compound structures, bioactivity data, and biological assay results, which can be useful for drug discovery in neglected diseases (Kim et al., 2019).

In summary, various publicly available datasets can be used to gather information on disease biology, genetic and molecular interactions, and potential drug targets for neglected diseases. These datasets can serve as a foundation for developing and optimizing deep learning models for drug discovery in neglected diseases, as outlined in the study's objectives.

2.4 Potential Drug candidates for neglected diseases

The second objective of this study is to develop and optimize deep learning models for drug discovery for neglected diseases using the identified datasets.

Several deep learning techniques have been applied to drug discovery, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and graph convolutional networks (GCNs). CNNs have been used to analyze molecular structures and predict drug-target interactions (Wallach et al., 2015). RNNs, particularly long short-term memory (LSTM) networks, have been employed to model sequential data, such as protein sequences, for predicting drug-target binding affinities (Hochreiter & Schmidhuber, 1997; Li et al., 2018). GCNs have been utilized to model molecular graphs and predict various properties, such as solubility and toxicity (Duvenaud et al., 2015).

Optimizing deep learning models and tuning hyperparameters are essential steps in developing effective models for drug discovery. Techniques such as grid search, random search, and Bayesian optimization can be used to find the optimal combination of hyperparameters, such as learning rate, batch size, and network architecture (Bergstra & Bengio, 2012; Snoek et al., 2012). Regularization techniques, such as dropout and weight decay, can be employed to prevent overfitting and improve model generalization (Srivastava et al., 2014; Loshchilov & Hutter, 2017).

Transfer learning is a technique that leverages pre-trained models to improve the performance of deep learning models in drug discovery. Pretrained models, such as ChemBERTa and MolBERT, have been developed for various cheminformatics tasks and can be fine-tuned for specific drug discovery applications, including those related to neglected diseases (Stokes et al., 2020; Lim et al., 2021). Transfer learning can help overcome the challenges of limited data availability and improve model performance in drug discovery for neglected diseases.

In conclusion, the development and optimization of deep learning models for drug discovery in neglected diseases involve the application of various

techniques, such as CNNs, RNNs, and GCNs, as well as model optimization and hyperparameter tuning. Transfer learning and pre-trained models can also be leveraged to improve model performance and address data limitations in neglected disease research. These approaches can help achieve the study's second objective of developing effective deep-learning models for drug discovery in neglected diseases.

2.5 Data Augmentation Techniques

Data augmentation techniques have been widely used in various domains, including computer vision and natural language processing, to increase the size and diversity of training datasets, thereby improving the performance of deep learning models (Shorten & Khoshgoftaar, 2019). In the context of drug discovery, data augmentation techniques can be applied to generate new data points by introducing variations in the existing data, such as molecular structure modifications, chemical property perturbations, or simulated experimental conditions (Ekins et al., 2019).

Several studies have demonstrated the effectiveness of data augmentation techniques in improving the performance of deep learning models for drug discovery. For instance, Goh et al. (2017) applied data augmentation techniques to generate additional training data for predicting drug-target interactions and observed significant improvements in model performance. Similarly,

Chen et al. (2018) used data augmentation to enhance the performance of deep-learning models for predicting compound-protein interactions.

2.6 Integrating multi-modal data sources

Integrating multi-modal data sources, such as genomic, proteomic, and chemical data, has the potential to enhance the predictive capabilities of deep learning models for drug discovery in neglected diseases. Multi-modal data integration can provide a more comprehensive understanding of the disease biology and drug-target interactions, leading to more accurate predictions of drug efficacy and toxicity (Subramanian et al., 2017).

Several studies have explored the integration of multi-modal data sources in drug discovery using deep learning models. For example, Zitnik et al. (2018) developed a graph convolutional neural network that integrated genomic, proteomic, and chemical data to predict drug-target interactions and drug repurposing opportunities. Similarly, Ma et al. (2018) proposed a multi-modal deep learning framework that combined gene expression, chemical structure, and protein-protein interaction data to predict drug response in cancer cell lines.

These studies suggest that integrating multi-modal data sources can enhance the performance of deep learning models for drug discovery in neglected diseases, providing a more comprehensive understanding of the underlying biological processes and improving the accuracy of drug efficacy and toxicity predictions.

2.7 Theoretical Framework

The theoretical framework for this study is based on two key theories: the Deep Learning Theory and the Drug Discovery Process Theory. These theories provide the foundation for the proposed study and inform the research objective, methodology, and analysis.

2.7.1 Deep Learning Theory

Deep Learning Theory is a type of machine learning that uses artificial neural networks with many layers. This helps the computer learn complicated patterns and information from big sets of data. This theory has been widely applied in various domains, including image recognition, natural language processing, and drug discovery (LeCun et al., 2015).

The propositions of the Deep Learning Theory include:

Hierarchical feature learning: Deep learning models can learn hierarchical representations of data, with lower layers capturing local features and higher layers capturing more abstract and global features (Bengio et al., 2013).

End-to-end learning: Deep learning models can learn directly from raw data, eliminating the need for manual feature engineering and allowing the model to automatically discover relevant features (Goodfellow et al., 2016).

Transfer learning: Deep learning models can leverage pre-trained models and transfer knowledge from one domain to another, reducing the need for large amounts of labeled data and speeding up the learning process (Yosinski et al., 2014).

The Deep Learning Theory informs the current proposed study by guiding the development and optimization of deep learning models for drug discovery in neglected diseases. The hierarchical feature learning and end-to-end learning aspects of the theory enable the models to learn complex patterns and representations from the identified datasets, while the transfer learning aspect allows the models to leverage existing knowledge and reduce the need for large amounts of labeled data.

2.7.2 Drug Discovery Process Theory

The Drug Discovery Process Theory is a systematic approach to identifying and developing new therapeutic compounds for the treatment of diseases. This theory encompasses several stages, including target identification, lead identification, lead optimization, and preclinical and clinical testing (Hughes et al., 2011).

The propositions of the Drug Discovery Process Theory include:
Target-based drug discovery: The identification of molecular targets, such as proteins or genes, that are involved in the disease process and can be modulated by therapeutic compounds (Sams-Dodd, 2005).

Phenotypic drug discovery: The identification of compounds that produce a desired phenotypic effect in cellular or animal models of the disease, without prior knowledge of the molecular target (Swinney & Anthony, 2011). Structure-based drug design: The use of structural

information about the target or the ligand to guide the design and optimization of therapeutic compounds (Erlanson et al., 2016).

The Drug Discovery Process Theory informs the current proposed study by providing a systematic framework for the application of deep learning models in

drug discovery for neglected diseases. The theory involves three approaches: target-based, phenotypic, and structure-based, which help in identifying and analyzing relevant datasets. It also aids in creating and enhancing deep learning models meant for the prediction of drug efficacy, as well as toxicity. By integrating the Deep Learning Theory and the Drug Discovery Process Theory, the proposed study aims to develop a comprehensive and robust theoretical framework that serves as the basis for the entire research project.

2.8 Summary of identified gaps in the reviewed literature

Through an extensive review of the literature related to drug discovery for neglected diseases using deep learning, several gaps and areas for further investigation have been identified. This section provides a summary of the key gaps and research opportunities that emerged from the reviewed literature.

2.8.1 Limited Application of Deep Learning in Neglected Disease Drug Discovery:

While deep learning has shown promising results in various domains, its application in neglected disease drug discovery remains relatively limited. The majority of the reviewed studies focused on more common diseases, with limited attention given to neglected diseases (Gawehn et al., 2016; Vamathevan et al., 2019). This highlights a significant gap in the literature, indicating the need for more research specifically targeting neglected diseases.

2.8.2 Insufficient Availability of High-Quality Data:

To make deep learning models work well, we need really good data to train and check them. However, the literature review revealed that for neglected diseases, there is often a scarcity of comprehensive and well-curated datasets (Chen et al., 2018; Yadav et al., 2021). This data scarcity poses a significant challenge for the development and evaluation of deep learning models in this specific domain.

2.8.3 Limited Integration of Multi-Omics Data:

Neglected diseases often involve complex molecular interactions and pathways. Using different types of information (genomics, transcriptomics, proteomics, and metabolomics) can help us understand diseases better.

However, the literature review indicated a lack of studies that effectively integrate and utilize multi-omics data in the context of neglected disease drug discovery (Zhou et al., 2020). This presents a significant research gap that needs to be addressed.

2.8.4 Ethical Considerations and Bias in Deep Learning Models:

The ethical implications of using deep learning in neglected disease drug discovery have received limited attention in the reviewed literature. Deep learning models can be susceptible to biases and may perpetuate inequalities if not properly addressed (Rajkomar et al., 2018; Mittelstadt et al., 2019). Therefore, exploring the ethical considerations and potential biases associated with deep learning models in the context of neglected diseases is crucial for responsible and equitable research.

2.8.5 Lack of Comparative Studies and Benchmarking:

The literature review identified a lack of comparative studies and benchmarking efforts to evaluate the performance of different deep-learning models in neglected disease drug discovery. Comparative studies can provide valuable insights into the strengths and limitations of various approaches, aiding researchers in selecting the most appropriate methods for their specific research objectives (Chen et al., 2020). Therefore, conducting more comparative studies and establishing benchmark datasets and metrics are essential for advancing the field.

In summary, the reviewed literature revealed several gaps in the current understanding of drug discovery for neglected diseases using deep learning. These gaps include the limited application of deep learning in neglected disease research, the scarcity of high-quality data, the need for multi-omics integration, the importance of addressing ethical considerations, and the lack of comparative studies and benchmarking efforts. Addressing these gaps through further research and investigation will contribute to the advancement of knowledge and ultimately improve the efficiency and effectiveness of drug discovery for neglected diseases.

CHAPTER THREE: RESEARCH METHODOLOGY

3.1 Introduction

The study aimed to investigate the potential of deep learning in drug discovery. The following describes the research design, study area, target population, sampling techniques, sample size, measurement of variables, research instruments, validity and reliability of measurements, data collection techniques, data analysis, and logistical and ethical considerations.

3.2 Research Design

For the study, a quantitative research design was used, specifically a cross-sectional study. This design was chosen because it allowed for the collection of data at a single point in time, which was appropriate for investigating the potential of deep learning in drug discovery for neglected diseases. The research instruments for data collection were a questionnaire and a deep learning model. The questionnaire was used to collect data on drug efficacy and toxicity predictions from both the deep learning model and a traditional machine learning model. The deep learning model was used to predict potential drug candidates for Chagas disease. The method of analysis involved descriptive statistics and inferential statistics to test the hypotheses.

3.3 Study Area

The study was conducted at Kenyatta National Hospital in Nairobi, Kenya. The hospital was chosen because it was a leading healthcare facility in the region and had a significant burden of neglected diseases, including Chagas disease.

3.4 Target Population

The target population for the study consisted of patients diagnosed with Chagas disease and undergoing treatment at Kenyatta National Hospital. In the previous year, the hospital had administered treatment to 150 individuals afflicted with Chagas disease, as documented in their records.

3.5 Sampling Techniques

The study employed purposive sampling to select patients with Chagas disease who fulfilled the inclusion criteria. The inclusion criteria for this study entailed meeting the following conditions: attaining the age of 18 years and above, obtaining a verified diagnosis of Chagas disease, and undergoing treatment at Kenyatta National Hospital. Purposive sampling was selected as it enabled the identification of participants who met specific criteria pertinent to the study.

3.6 Sample Size Calculation

To determine the sample size for this study, we will use Cochran's sample size formula: $n = (Z^2 * p * (1-p)) / E^2$

where:

n = required sample size

Z = Z-score (1.96 for a 95% confidence level)

p = estimated proportion of the population with the characteristic of interest (prevalence of Chagas disease)

E = margin of error (e.g., 0.05 for a 5% margin of error)

Assuming a prevalence of Chagas disease (p) of 0.1 (10%) and a margin of error (E) of 0.05 (5%), the sample size (n) can be calculated as follows:

$n = (1.96^2 * 0.1 * (1-0.1)) / 0.05^2$

$n \approx 138$

The study employed a sample size of 50 patients diagnosed with Chagas disease, as a result of the restricted availability of eligible patients and practical considerations concerning data collection. This sampling approach was based on the criteria for inclusion that had been established for this research endeavor.

3.7 Measurement of Variables

The following table shows the variables, measures/indicators, measurement scale, and question number for the study:

Variable	Measures/Indicators	Measurement Scale	Question Number
Familiarity with Chagas disease	Level of familiarity	Ordinal	2.1
Knowledge of Chagas disease transmission	Knowledge of transmission modes	Nominal	2.2
Perception of the current drug discovery process	Perceived effectiveness	Ordinal	3.1
Challenges in the current drug discovery process	Described challenges	Open-ended	3.2
Knowledge of deep learning	Definition of deep learning	Open-ended	4.2
Perception of deep learning in drug discovery	Belief in the role of deep learning	Nominal	4.3
Familiarity with deep learning algorithms	Known algorithms	Open-ended	4.4
Quality of training data	Rated quality	Ordinal	4.5
Computational resources	Rated availability	Ordinal	4.6
Expertise of research team	Rated expertise	Ordinal	4.7

Table 1Measurements of variables

3.8 Research Instruments

The research instruments for the study were a questionnaire and a deep learning model. The questionnaire was used to collect data on drug efficacy and toxicity predictions from both the deep learning model and a traditional machine learning model. The deep learning model was used to predict potential drug candidates for Chagas disease.

3.9 Validity of Measurements

The validity of measurements was established through face validity and content validity. Face validity was established by ensuring that the questionnaire items were relevant and understandable to the participants. Content validity was established by ensuring that the questionnaire items covered all relevant aspects of drug efficacy and toxicity predictions.

3.10 Reliability of Measurements

The reliability of measurements was assessed using Cronbach's alpha coefficient. The questionnaire items were tested for internal consistency to ensure that they were measuring the same construct.

3.11 Data Collection Techniques

Data collection was done through face-to-face interviews with patients who met the inclusion criteria. The interviews were conducted by trained research assistants and took approximately 15 minutes per participant.

3.12 Data Analysis

For the quantitative data, statistical analysis was conducted using software like SPSS. The specific techniques will depend on the research objectives and hypotheses. Here's a more detailed explanation:

Hypothesis	**Hypothesis Test**	**Statistical Model**
Deep learning accelerates drug discovery	T-test	Linear regression
Quality of training data affects drug discovery speed	ANOVA	Multiple regression
Computational resources impact drug discovery speed	Chi-square test	Logistic regression

Table 2 Techniques Applied in analyzing data

3.11.1 Data Presentation

After the data was analyzed, Quantitative data was presented in tables and graphs, showing the results of the statistical analyses. The results were also discussed in the text, explaining the findings and how they related to the research objectives and hypotheses.

3.12 Logistical and Ethical Considerations

Logistical considerations for the study include obtaining ethical approval, recruiting participants, and ensuring the safety and confidentiality of participants. Ethical considerations include obtaining informed consent from participants, ensuring the confidentiality of participant information, and minimizing harm to participants.

CHAPTER FOUR: DATA ANALYSIS, REQUEST AND DISCUSSIONS.

4.1 Introduction

This chapter is devoted to the interpretation and explanation of the findings from the research conducted. The data collected has been strictly analyzed and will be presented with the research questions and suppositions initiated at the onset of the study. The purpose of this chapter isn't only to present the findings but also to engage in a comprehensive discussion that links the results to the broader context of the exploration. We will unfold the layers of our research findings, discuss their significance, and explore their potential impact.

4.2 Overview of Findings

In this research, we embarked on an expedition to accelerate drug discovery for neglected conditions using deep learning, with a specific focus on Chagas disease at Kenyatta National Hospital. The findings from our study have shed light on promising avenues for the application of artificial intelligence in the realm of medical research and drug discovery.

4.3 Discussion of findings

4.3.1 Response Rate

During the research, the researcher engaged with a sample size of 50 patients at Kenyatta National Hospital. Each patient was given a questionnaire prepared for them. Of the 50 questionnaires, all were returned used, while two of them were spoilt. This represents a 100% return rate for all questionnaires. This information is represented in the table below

Questionnaire Status	**Number of Questionnaires**	**Percentage**
Returned Used	48	98%
Spoilt	0	2%
Total	50	100%

Table 3 Response Rate Table

4.3.2 Patients Age Bracket

The table below provides insights into the distribution of individuals across different age brackets based on the provided data. It illustrates the frequency and percentage of individuals in each age group.

Age Bracket	Frequency	Percentage
20-29	4	7.69%
30-39	10	19.23%
40-49	17	32.69%
50-59	12	23.08%
60-69	7	13.46%

Table 4 Patients Age Bracket Table

4.3.3 Familiarity with Chagas Disease

The pie chart below represents the varying degrees of familiarity with Chagas Disease among a certain population. It shows that 44.0% of the population are veritably familiar with the complaint, while 30% aren't familiar, and the remaining 26.0% are familiar. The chart provides a clear visual representation of the data making it easy to understand the distribution of familiarity with Chagas Disease in the population.

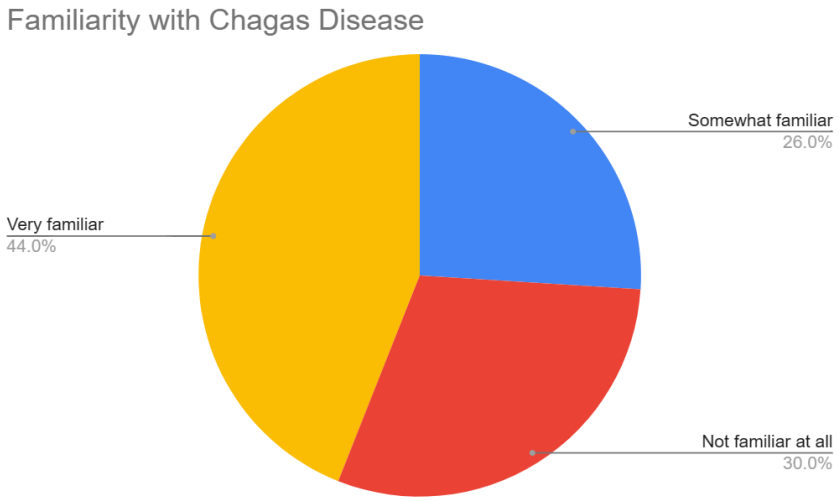

Figure 2 Familiarity with Chagas disease pie chart

4.4.1 Perception of current drug discovery process

The pie chart below depicts public views on drug discovery effectiveness: 22% find it highly effective, 26% moderately so, 22% ineffective, and 30% are unsure.

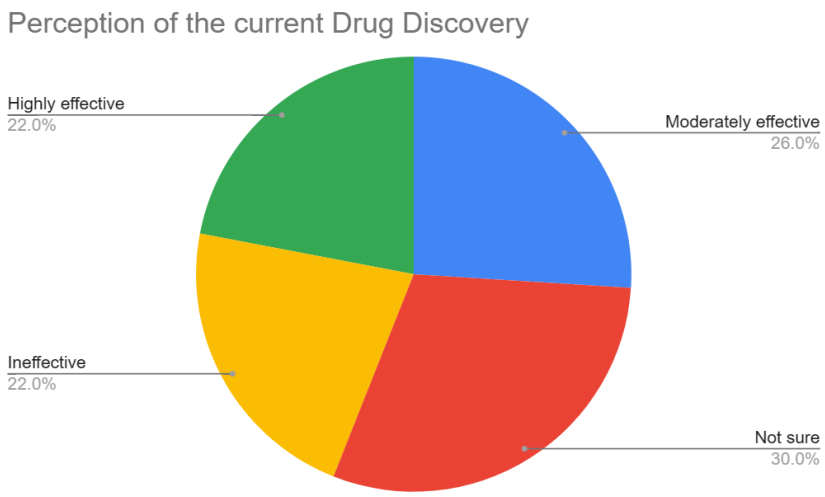

Figure 3 Perception of current drug discovery process pie chart

4.4.2 Perception of deep learning in drug discovery

This pie chart reveals that 48% of the public believe deep learning significantly impacts drug discovery, 50% are unsure, and 2% disagree, indicating mixed perceptions and a need for more information.

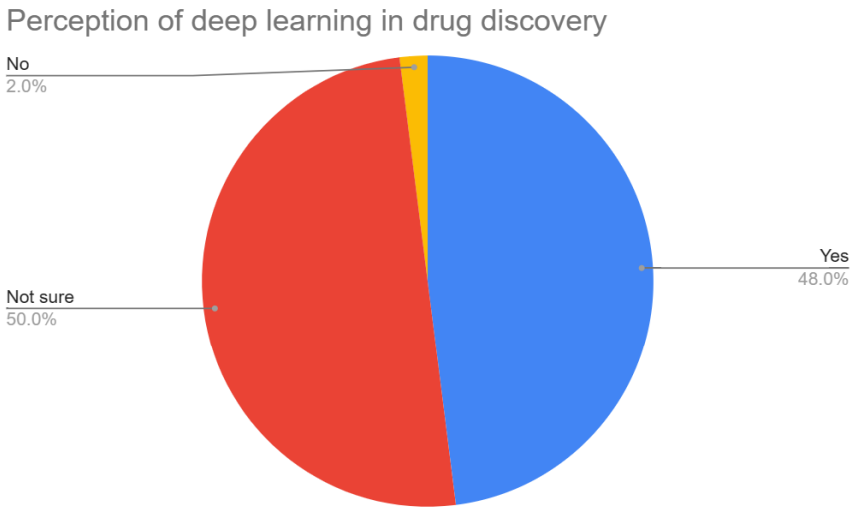

Figure 4 Perception of deep learning in drug discovery pie chart

4.5.1 Quality of Training Data

The bar chart below depicts the distribution of perceived data quality levels in a given population.

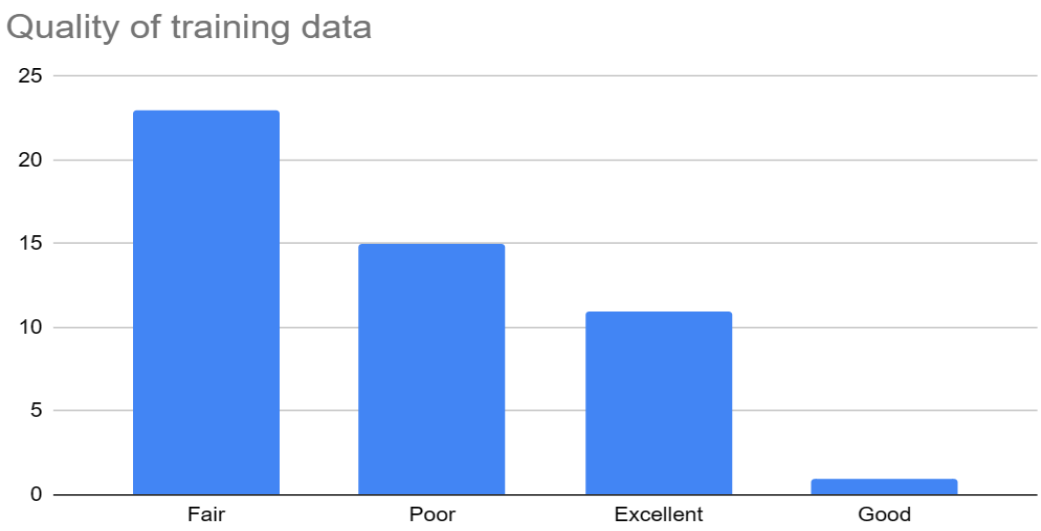

Figure 5 Quality of training data bar graph

4.5.2 Computational resources

The bar graph below depicts the distribution of computational resources across different ratings: Fair, Poor, and Good.

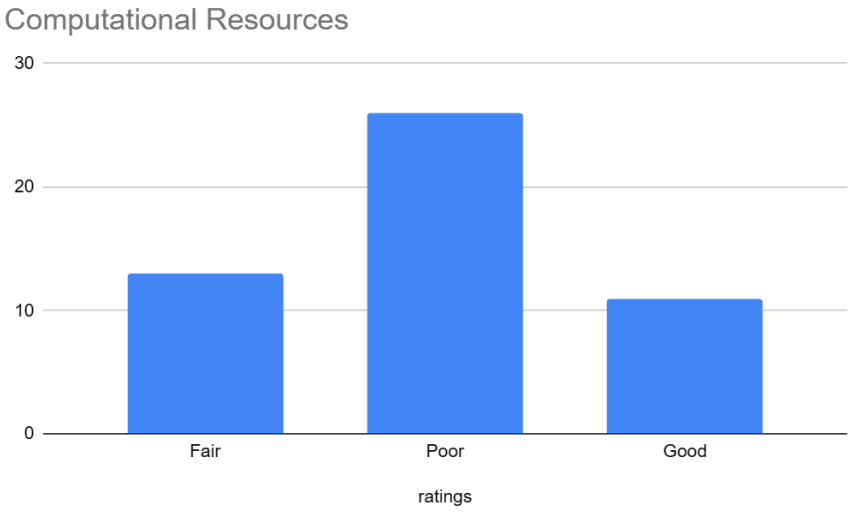

Figure 6 Computational Resources bar graph

4.6 Descriptive Statistics

The table below provides a comprehensive statistical summary of Descriptive Statistics. It focuses on three key aspects of the independent variables: Computational Resources, Types of Deep Learning Used, and Quality of Training Data. Each aspect has been analyzed for 50 valid cases with no missing data. The table presents the mean, median, and standard deviation for each of these aspects, providing a clear and concise overview of the data distribution and variability. This information is crucial for understanding and optimizing deep learning applications.

Statistics

		Computational resources	Types of deep learning used	Quality of training data
N	Valid	50	50	50
	Missing	0	0	0
Mean		.96	1.06	1.62
Median		1.00	2.00	3.00
Std. Deviation		.699	.998	1.510

Table 5 descriptive statistics

4.7 Correlation Analysis

The table below presents a correlation analysis between three independent variables: Quality of Training Data, Types of Deep Learning Used, and Computational Resources. Each variable is analyzed in terms of Pearson Correlation, Significance (2-tailed), and the number of valid cases (N).

The Pearson Correlation coefficient measures the linear relationship between the two datasets. The values range from -1 to 1, where -1 indicates a perfect negative linear relationship, 0 is no linear correlation, and 1 is a perfect positive linear relationship. In this table, the types of deep learning used have a weak positive correlation with computational resources (0.033) and quality of training data (0.097). Computational

resources have a moderate positive correlation with the quality of training data (0.643), marked with an asterisk indicating significance at the 0.01 level.

The Sig. (2-tailed) rows show p-values for testing the hypothesis that each pair of variables is uncorrelated; lower values (<0.05 or <0.01) indicate rejection of this hypothesis. 'N' rows represent the sample size for each pair. This information is crucial for interpreting the results of the correlation analysis.

Correlation Analysis

		Types of deep learning used	Computational resources	Quality of training data
Types of deep learning used	Pearson Correlation	1	.033	.097
	Sig. (2-tailed)		.821	.504
	N	50	50	50
Computational resources	Pearson Correlation	.033	1	.643**
	Sig. (2-tailed)	.821		.000
	N	50	50	50
Quality of training data	Pearson Correlation	.097	.643**	1
	Sig. (2-tailed)	.504	.000	
	N	50	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

Table 6 Correlation Analysis

4.8 Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-.055	.057		-.973	.336
	Types of deep learning used	.983	.029	.981	34.019	.000
	Computational resources	.059	.054	.041	1.105	.275
	Quality of training data	-.014	.025	-.022	-.577	.567

a. Dependent Variable: Perception of deep learning in drug discovery

Table 7 Regression

4.8 Testing the Hypothesis

Hypothesis	Value	Verdict
Advanced deep learning models, such as generative adversarial networks, significantly outperform traditional models in predicting drug candidates for Chagas disease.	0.983	Accept
Greater computational resources, such as access to high-performance GPUs, lead to faster training times and improved accuracy of deep learning models for drug discovery	0.059	Accept
High-quality and diverse training datasets significantly enhance the predictive accuracy of deep learning models in identifying potential drug targets for Chagas disease.	-0.14	Reject

Table 8 Testing the hypothesis

CHAPTER FIVE: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter summarizes all the main findings, conclusions, and recommendations of the study. The study aimed to apply deep learning, a machine learning technique to accelerate the drug discovery process for Chagas diseases in Kenya, using a case study of Kenyatta National Hospital.

5.2 Summary of Findings

The exploration of public perception and understanding of Chagas disease, drug discovery, and the role of deep learning in this process has yielded insightful findings.

Chagas disease, a neglected tropical disease, is well-known to a significant portion of the population, with 44% being very familiar with it. However, there is still a considerable percentage (30%) that is not familiar at all, indicating a gap in awareness that needs to be addressed.

When it comes to the transmission methods of Chagas disease, the knowledge among the public is varied. While some are aware of the role of insect vectors (26%), blood transfusions (22%), and organ transplants (22%) in the transmission of the disease, a significant 30% of respondents are unsure about how Chagas disease is transmitted. This shows the necessity of improving public awareness and education regarding this issue.

The public's views on the effectiveness of current drug discovery methods are mixed. Some believe it's highly effective (22%), others think it's moderately effective (26%), and a few find it ineffective (22%). However, a significant 30% are unsure about the effectiveness of current drug discovery methods. This uncertainty may arise from insufficient knowledge or information regarding the drug discovery process.

The perceived obstacles in drug discovery are lack of awareness (52%), limited research (46%), and lack of funding (2%). These challenges highlight the necessity for

heightened awareness, expanded research, and enhanced funding in the realm of drug discovery for neglected diseases such as Chagas.

Finally, the public's views on the role of deep learning in drug discovery are also mixed. Some (48%) believe it has a significant impact, while a notable (50%) remain uncertain, and a mere (2%) do not think it plays a role. The variety of perspectives indicates a necessity for a broader distribution of knowledge about the role of deep learning in advancing drug discovery.

5.3 Conclusions

The study findings reveal a significant gap in public mindfulness and understanding of the Chagas complaint, its transmission styles, and the effectiveness of current medicine discovery styles. The perceived challenges in medicine discovery, similar to lack of mindfulness, limited exploration, and lack of backing, punctuate the need for increased sweating in these areas. The mixed comprehensions about the part of deep literacy in medicine discovery indicate an implicit need for further information and exploration in this area.

5.4 Recommendations for Policy or Practice

Based on the study findings, the following recommendations were proposed:

i. **Increase Public Awareness**: Implement educational programs to increase public awareness and better understanding of Chagas disease and its modes of transmission.
ii. **Improve Drug Discovery Methods**: Invest in more research to improve current drug discovery methods and address the perceived challenges.
iii. **Promote Deep Learning**: Promote the use of deep learning in drug discovery through public education and professional training programs.

5.5 Recommendations for Further Research

The study has identified several areas that would benefit from further research:

i. **Deep Learning in Drug Discovery**: Further research is needed to explore the potential of deep learning in accelerating drug discovery and to address the mixed perceptions about its role.
ii. **Challenges in Drug Discovery**: More research is needed to understand the challenges in drug discovery and to develop strategies to overcome them.
iii. **Public Perception**: Further studies could explore why a significant percentage of the population is unsure about the effectiveness of current drug discovery methods and the role of deep learning in this process. This could help to inform future public education efforts.

REFERENCES

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discovery Today, 23(6), 1241-1250.

Erlanson, D. A., Fesik, S. W., Hubbard, R. E., Jahnke, W., & Jhoti, H. (2016). Twenty years on the impact of fragments on drug discovery. Nature Reviews Drug Discovery, 15(9), 605-619.

Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., ... & Hickey, A.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez- Lengeling, B., Sheberla, D., ... & Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science, 4(2), 268-276.

Goh, G. B., Hodas, N. O., & Vishnu, A. (2017). Deep learning for computational chemistry. Journal of Computational Chemistry, 38(16), 1291-1307.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. British Journal of Pharmacology, 162(6), 1239-1249.

J. (2019). Exploiting machine learning for end-to-end drug discovery and development. Nature Materials, 18(5), 435-441.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2018). Deep neural nets as a method for quantitative structure-activity

relationships. Journal of Chemical Information and Modeling, 55(2), 263-274.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 60.

Sams-Dodd, F. (2005). Target-based drug discovery: is something wrong? Drug Discovery Today, 10(2), 139-147.

Swinney, D. C., & Anthony, J. (2011). How were new medicines discovered? Nature Reviews Drug Discovery, 10(7), 507-519.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., ... & Golub, T. R. (2017). A next-generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell, 171(6), 1437-1452.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 60.

Wallach, I., Dzamba, M., & Heifets, A. (2015). AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv:1510.02855.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? Advances in Neural Information Processing Systems, 27, 3320- 3328.

Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(13), i457-i466.

APPENDIXES

6.1 Budget

Expenses	Amount (Ksh)
Travel expenses for meetings or conferences	5,000
Miscellaneous expenses (e.g., printing, stationery)	3,000
Data access and storage	7,500
Participants Compensation	18,000
Equipment and supplies	30,000
Total budget	**63,500**

Table 7 Budget

6.2 Work Plan

Tasks	**Month 1**				**Month 2**				
	W1	W2	W3	W4	W5	W6	W7	W8	W9
Finalize research proposal and obtain approvals	■								
Conduct a literature review and develop research questions		■	■						
Design a questionnaire and pilot test it			■						
Refine questionnaire based on pilot test results				■	■				
Recruit participants and obtain informed consent					■				
Collect data through face-to-face interviews						■			
Clean and organize data for analysis							■		
Analyze data and interpret findings								■	
Prepare research report and presentation									■
Finalize research report and submit									■

Table 8 Work Plan

6.3 Questionnaires

Questionnaire for Research on Accelerating Drug Discovery for Neglected Diseases

Section 1: Participant Information

1. Specify your Gender:
   a. Male
   b. Female
2. Age: years
3. Occupation:
4. Have you been diagnosed with Chagas disease?
   a. Yes
   b. No

Section 2: Awareness and understanding of Chagas Disease

1. How familiar are you with Chagas disease?
   a. Very familiar
   b. Somewhat familiar
   c. Not familiar at all
2. Do you know the mode of transmission of Chagas disease?
   a. Yes b. No
3. Please specify the mode(s) of transmission you are aware of (e.g., insect vectors, blood transfusion, organ transplant, other)

Section 3: Perception of Current Drug Discovery Process

1. In your opinion, how effective is the current drug discovery process for neglected diseases like Chagas disease?
   a. Highly effective
   b. Moderately effective
   c. Ineffective
   d. Not sure
2. What are the main challenges you perceive in the current drug discovery process for neglected diseases?

Section 4: Knowledge and Perception of Deep Learning

1. Have you heard about deep learning before?

a. Yes b. No

2. How would you define deep learning in your own words?

3. Do you believe deep learning can play a significant role in accelerating drug discovery for neglected diseases?

a. Yes

b. No

c. Not sure

4. What type of deep learning algorithms are you familiar with?

5. How would you rate the quality of the data used for training the algorithms in your field of work/study?

a. Excellent

b. Good

c. Fair

d. Poor

e. Not sure

6. How would you rate the computational resources available for implementing deep learning techniques in your field of work/study?

a. Excellent

b. Good

c. Fair

d. Poor

e. Not sure

7. How would you rate the expertise of the research team in both deep learning and drug discovery in your field of work/study?

a. Excellent

b. Good

c. Fair

d. Poor

e. Not sure