**Abstract**

Leukemia is a form of cancer that can be a fatal disease, and to rehabilitate and treat it requires a correct and early diagnosis. Standard methods have transformed into automated computer tools for analyzing, diagnosing, and predicting symptoms.

In this work, a comparison study was performed by comparing two different leukemia detection methods. The methods were a genomic sequencing method, which is a binary classification model and a multi-class classification model, which was an images-processing method. The methods had different input values. However, both of them used a Convolutional neural network (CNN) as network architecture. They also split their datasets using​ 3-way cross-validation. The evaluation methods for analyzing the results were learning curves, confusion matrix, and classification report. The results showed that the genome model had better performance and had several numbers of values that​ were correctly predicted with a total accuracy of 98%. This value was compared to the image processing method results that have a value of 81% total accuracy. The size of the different data sets can be a cause of the different test results of the algorithms.

**Glossary**

**Adam** -​ an optimization algorithm used to iterative update the network’s weight during training.

**BCCD -** A MIT ​ license​ dataset with white and red blood cells images​

**CBC -** stands for Cell blood counting and is a method used in cancer detection.​

**Fasta Form -** is nucleotide bases written in plain text based form.​

**Genbank -** A database that preserves genetic material.​

**Tensor operations -** is​ a data structure type used in linear linear algebra where you can calculate vectors and matrices.

**WBC** -​​ stands for white blood cell.​

# 1 Introduction

The practice of medicine is getting modernized every year and continuously moving towards more automated systems that help and improves the healthcare practice to be more productive with treatments and accurate in their assessments [1]. With the use of machine learning, it increases the values and redefines diagnostic methods.

Over the years, cancer-related research has grown and evolved into different fields and have adapted

deep learning methods such as image screening and genome sequencing. Moreover, the new treatments and diagnostic strategies have increased test results’ accuracy for cancer predictive methods [2]. There are tools such as genomic sequencing which can detect and identify patterns in input values and effectively diagnose cancer types, which is a challenging task for physicians to do manually.

Deep learning is a part of Artificial intelligence and is described as a computer that works similar to

the human mind and collects raw data with a logical construct [3]. The Artificial Neural Networks(ANNs) consists of neurons, which is where they accept and store information at each before transferring to the next layer. It builds a complex system with multiple layers. This makes it possible for the system to retrieve information without human interference [3]. A convolutional neural network(CNN) is a good examples of ANN [7].

Advanced methods can be used to help patients detect terminal disorders such as leukemia, which is a fatal disorder and common cancer type amongst children. Leukemia is a form of cancer that begins in blood cells and the bone marrows, where it grows new immature blood cells when the body does not need them. White blood count(WBC) is a routine blood test usually done manually, to search for leukemia cells and can be automated by applying machine learning techniques such as CNN. It is a simple and faster way to perform a test and detect abnormality in the blood [19]. Other practices are genomic sequencing to detect the abnormal markers in coding and non-coding regions along with DNA sequences. This is used to predict or detect cancer from using biomarkers [20].

Genomic sequencing uses DNA sequence as input data, and are composed of nucleotides [4]. Nucleotides have four nitrogen bases adenine, cytosine, guanine, or thymine. They form a base pair that creates a double shaped helix, which is the principal structure for DNA [5].

Despite all the benefits of AI, such as preventing diseases, there are concerns and ethical implications. These concerns revolve around data privacy that could affect the patients safety, but also the safety of their genetic relatives [4]. It also has a positive side in the medical care system, assisting doctors and in giving second opinions to increase the accuracy of the diagnoses. But there are also risk of genetic discrimination [6].

**1.2 Purpose**

The purpose of this project is to do a comparative analysis study on cancer detection models that applies deep learning to address the topic of automated diagnosis. The project also uses two different models used for leukemia detection. The goal of this project is to select two models that apply deep learning to detect leukemia and compare them, to discuss its use of AI. The idea is to highlight, inform, and discuss a few selected applications on cancer predictive AI methods and also the benefits and challenges in cancer detection models.

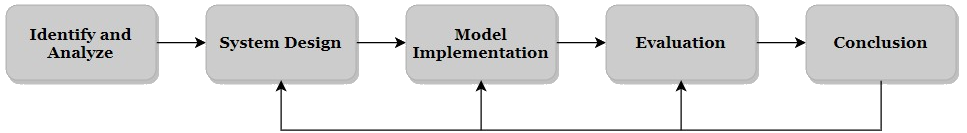
## 2.2 Technological Background

#### 2.1.2 Machine learning

*Machine learning* ​is a part of artificial intelligence, and the idea is generally defined as a software system having the information to learn from experience using a set of tasks. Three essential aspects define how machine learning functions. These aspects are tasks, experience, and performance. Tasks are datasets to train the computer to increase its performance. With time and experience, the computer system can learn and become a refined model that can prognosticate the answer to a topic that it has learned from previous attempts [12]. There are multiple algorithms used in machine learning, but they fall into two categories, supervised learning and unsupervised learning. The supervised learning group also referred to as a method working with a set of training data. The dataset has an input and output object for each example [12]. In an attempt to classify the result, the algorithm needs to work on manually entered answers. This type of working method is heavily dependent on the training data. Therefore, the set needs to be correct for the algorithm to make sense of the data. Unsupervised learning is that the algorithm finds undetected patterns in a massive amount of data. In this type of method, it allows the computer algorithm to execute and see what the outcome patterns are going to be. For that reason, there is no clear answer that is considered right or wrong [12]. In machine learning, there are dependent and independent variables. The independent variables are also referred to as predictor or control input; this holds the values that control the experiment. The dependent variables, otherwise known as output values, are regulated by the independent variables[39].

# 4 Method

This chapter presents a method that is selected for this project. The research process aims to gain more knowledge of the subject around deep learning and its application in the medical world. The experiment phase uses two models to implemented and tested. The research methodology selected is Takeda's General Design Cycle (GDC) because of its simple formatted research design and iterative approach has been modified to fit the thesis, which is shown in figure 4[28]. Each cycle produces a result that is used to compare to the next attempt result. This is to test quality and to improve the research continuously. These are attributes that are essential for the project, where testing needs to be done in multiple ways and compared.



*Figure 4. The Systems Development Research Methodology[45]*

## 4.2 System Design

In the second step, a diagram is designed and represents the projects workflow from collecting the data to testing and evaluating the result. This phase is a creative place to make a drawing of the process and describe the necessary functions that are required. In chapter 5, there is a process model that describes the systems' multiple phases, such as the selection of datasets and preprocessing. All these steps are important in order to prepare the models to be implemented and tested so that the output gives accurate results.

## 4.3 Model Implementation

The third phase consisted of testing the model. All the details for the implementation and testing of both models are explained in this phase. Information and design from previous steps are applied in order to implement the models. The genomic sequence method and the image processing method

uses its datasets to feed the model in order to train and test the models network architecture.

## 4.4 Evaluation

The fourth phase of the process is to observe and evaluate the system, which are an opportunity to improve the result from testing. The results from both tests are presented in section 5.4, and it explained what each curve and matrices indicated. The information in this phase revealed if the value stated in prior testing is according to the experiments results. The output from the result are used to answer the research question concerning the different methods, which are useful for the next phase. In both methods, the learning curve and confusion matrix are used to evaluate and summarize the performance of the models.

## 4.5 Conclusion

The conclusion step reflects on the performance of the two different models that have a similar purpose by doing a comparative analysis. With the use of the theory chapter, the models are discussed and analysed. Their performances are evaluated, and their results' value and quality are compared to result from other previous attempts. This part is the end of the cycle. It brings it together to form a conclusive decision regarding the solution to the research questions. This is presented in chapter 6.