Enhanced Thyroid Disease Prediction Through Machine Learning and Deep Learning: A Comparative Approach

By

Rehnuma Tarannum

Sharmin Akter Reka

Thesis for the Degree of Bachelor of Science
in
Computer Science and Engineering

NATIONAL UNIVERSITY 2022

APPROVAL

The thesis paper titled "Enhanced Thyroid Disease Prediction Through Machine Learning and Deep Learning: A Comparative Approach" submitted by Rehnuma Tarannum (18502003701), Sharmin Akter Reka(18502003735) to the Department of Computer Science and Engineering, Institute of Science, Trade and Technology has been accepted as satisfactory for the partial fulfillment of Bachelor of Science degree requirements for deposit.

Supervisor Approval

Sheikh Tanjil Sharif
Lecturer, Department of Computer Science and Engineering

Departmental Approval

S.K. Mamunur Rashid
Head Of the Department, Department of Computer Science and Engineering

DECLARATION

We declare that our work has not been previously submitted and approved for the award of a degree by this or any other University. As per of my knowledge and belief, in this thesis contains no material previously published or written by another person except where due reference is made in the paper itself.

Signature

Name: Rehnuma Tarannum

Registration No: 18502003701

Signature

Name: Sharmin Akter Reka

Registration No: 18502003735

ACKNOWLEDGEMENT

We would like to express our deepest gratitude to our research supervisor Shekh Tanjil Sharif, Lecturer, Department of Computer Science and Engineering, ISTT, for his insightful guidance, unwavering support, and continuous encouragement throughout all stages of our work. His valuable advice and constant inspiration have been pivotal in the successful completion of this thesis.

We also extend our heartfelt thanks to the Head of the Department (CSE) for providing us with an excellent platform and resources that facilitated the successful completion of our research.

Lastly, we are deeply grateful to all our teachers, friends and family members for their continuous encouragement and support. Their belief in us has been a source of motivation throughout this journey, and we are sincerely thankful for their presence and understanding during challenging times.

Rehnuma Tarannum

Sharmin Akter Reka

Dedicated to our friendship

ABSTRACT

Thyroid diseases have become increasingly prevalent worldwide, affecting millions across diverse populations. The thyroid gland, located in the front of the neck and wrapped around the windpipe, plays a critical role in regulating metabolism through the secretion of thyroid hormones. These hormones influence nearly every system in the body, making the proper functioning of the thyroid essential for overall health. Two major thyroid disorders are hyperthyroidism, which is characterized by excessive hormone production, and hypothyroidism, marked by insufficient hormone secretion. Although both conditions are chronic, they can be effectively managed with appropriate treatment, allowing patients to lead stable lives.

Advancements in medical research and technology, particularly in machine learning and deep learning, are contributing to more accurate diagnoses and personalized treatment plans for thyroid disorders. These technological approaches are improving patient outcomes by providing predictive models for disease progression and response to treatment, which is especially valuable for chronic conditions like thyroid disease. The integration of such technologies offers significant promise for enhancing both the precision and efficiency of thyroid disease management, a critical advancement given the global rise in incidence of these conditions.

This thesis focuses on the prediction and analysis of thyroid disease using two primary approaches: Machine Learning (ML) and Deep Learning (DL). Five machine learning algorithms—Logistic Regression, Decision Tree, Support Vector Machine (SVM) with linear kernel, SVM with RBF kernel, and Random Forest—were applied to predict thyroid disease, and their performance was compared. Additionally, Recurrent Neural Networks (RNN), a powerful deep learning algorithm were implemented to explore the predictive capabilities of deep learning techniques.

The study demonstrates that machine learning techniques, such as Decision Tree and Random Forest, can effectively predict thyroid disease with high accuracy. Deep learning models like RNN also provide competitive performance. The effectiveness of these models was evaluated using performance metrics, including Accuracy, Mean Squared Error (MSE), Root Mean Squared Error (RMSE). A confusion matrix was utilized to assess classification outcomes and error types generated by the models.

The thesis also discusses the limitations and challenges of current prediction models and provides recommendations for future research. These findings contribute to a better understanding of thyroid disease diagnosis using machine learning and deep learning techniques, highlighting areas where improvements in predictive accuracy can be made.

Keywords – Thyroid, Hyperthyroidism, Hypothyroidism, Machine Learning, Deep Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Recurrent, Neural Networks, Confusion matrix, Mean Squared Error (MSE), Root Mean Squared Error (RMSE)

Table of Contents

Contents

List of Tables

Tables

List of Figures

Figures

Chapter 1 Introduction

1.1 Introduction

The thyroid is one of the most diagnosed medical conditions. A healthy existence is necessary for a prosperous society, yet it is a fact that ostensibly invisible diseases affect our families and cause people to suffer. The disease of the thyroid falls within this category. The thyroid dysregulation is one of the prevalent chronic conditions that affect people of all ages. The condition is not as life - threatening as heart disease and cancer, but it may be the underlying cause of diseases with severe complications.

The thyroid gland's production of thyroid hormones serves in the regulation of the body's metabolism. Levothyroxine (abbreviated T4) and triiodothyronine (abbreviated T3) are the two active thyroid hormones produced by the thyroid glands (abbreviated T3). These hormones are necessary for the manufacturing as well as the comprehensive development and oversight of the body's temperature regulation system. Specifically, thyroxin (T4) and triiodothyronine (T3) are the two types of active hormones that the thyroid glands typically produce. Women are more susceptible than men to developing thyroid cancer. Women are more likely than males to get thyroid disease in their forties and fifties (The disorder of thyroid disease primarily happens in the women having the age of 17–54), whereas men are more likely to develop thyroid disease in their sixties and seventies. Whites are more likely than blacks to develop follicular thyroid disease.

It is estimated that approximately 10% of Bangladeshis suffer from clinically apparent thyroid problems. Recently, subclinical hypothyroidisms and hyperthyroidisms have been added to the list of thyroid disorders, bringing the total number of dysthyroid individuals to 20% of the population, with 10% of the population suffering from any sort of thyroid condition. As iodine deficiency disorders began to reduce in the early 1990s of the previous centuries, the absolute and relative number of different thyroid problems has altered over time. In Bangladesh, there were few published publications on the range of thyroid disorders. The range of thyroid problems other than iodine deficits was believed to be the same in Bangladesh as in other Asian countries. According to a study in which the author participated and which reported 35 percent of all thyroid disorders. 50 million of Bangladeshis are afflicted with thyroid disease

where the thirty million of the fifty million Bangladeshis who suffer from thyroid disease are unaware of their condition. At a press conference held by the Bangladesh Endocrine Society (BES) to mark World Thyroid Day at the Sagor-Runi auditorium of Dhaka Reporters Unity, it was revealed that "Females are 10 times more at risk of Thyroid hormone problems than males," BES President Dr. Md Faruk Pathan stated, adding that 20-30% of women suffered from some form of thyroid disease.

Despite the advancement, thyroid-related emergencies such as thyrotoxic crisis, undetected congenital hypothyroidism, pregnancy loss and infertility due to thyroid disorders, inappropriate use of thyroid medications, and late detection of thyroid malignancies are frequently the leading causes of significant morbidity and mortality. Graves' ophthalmopathy is still difficult to treat, and thyrotoxic crisis is nearly quite often fatal. Due to the vast number of new cases each year, it is still difficult to spot or diagnose this disease in its early stages, despite the availability of many technologies. As far as artificial intelligence is concerned, it has made significant contributions to the health industry by examining a wide range of illnesses and providing patients, doctors, lab workers, and others with the most assistance possible.

In this study, we analyzed the methodologies of different researchers for identifying various thyroid illnesses using machine learning and deep learning models.
Machine learning has become a fundamental element of human living, providing intelligent and cost-effective appropriate responses to a variety of challenges. In addition, new technologies have boosted the popularity of deep learning.

Deep learning is a subset of machine learning and a sort of artificial intelligence. It is essentially an AI function that aims to imitate the decision-making patterns of any human intellect, as well as the sorting and processing of data. Deep learning has demonstrated superior performance in recognition, classification, segmentation, and prediction of the future status of time-variable data compared to conventional approaches.

Simultaneously, deep neural networks have gained widespread popularity, particularly following the development of machine learning libraries for the most popular programming languages such as MATLAB, Python, Java, etc.

By reviewing the results and thesis, the tradeoff between machine learning and deep learning methods becomes noticeable.

1.2 Background

Accurate data analysis and use can improve service in numerous industries crucial to human life. Data analysis and data utilization are not synonymous, as their respective purposes differ. In other words, data analysis provides the foundation for data use. In addition, there is a variation in the required expertise for each of the two. The goal of data analysis is to extract the desired information from the data.

Data analysis is crucial for predicting the future state of a certain application, such as identifying the elements that lead to future growth or loss in the business sectors. Data is merely a collection of numbers and letters; therefore, it must be processed to make it easier to gather information. Knowledge of data analysis, statistics, and data processing procedures will enable more in-depth study.

Data utilization refers to the continual use of data to enhance operations and maintenance efficiency and production for the betterment of the company. Utilizing data is one of the key components to achieve data driven. Utilizing data makes it simpler to determine what tactics, strategies and activities should be followed to attain a set of goals. Listed below are a few advantages that data utilization can bring to our research.

1. Helpful for understanding the current state
2. Increase the speed of the decision-making process.

In the medical field, a vast quantity of medical data, including electronic medical records, examination, and imaging data, claim data, and research data, are managed. The healthcare professional requires useful tools for diagnosing medical conditions, analyzing prescription days
and amounts etc.

For example Specifically, in a scenario where data utilization results in preventative thyroid medication and early detection of thyroid disorders. By gathering and analyzing a large number of patients with the same symptoms from a massive amount of medical data, the data is utilized to identify and determine the progression of the

thyroid condition. But there are still numerous outstanding questions, including data analysis, utilization, prediction, and recognition.

The importance of classifiers in providing the health care provider with valuable tools? However, the question arises: which classifier should be used? What metrics are suitable for measurement? How can a proper distribution of the data be determined and predicted so that the classifier does not bias the medical patterns towards a certain class? Then, the most crucial question is whether these actions are effective for a certain ailment.

Machine learning and deep learning have attracted a great deal of attention due to the proliferation of problems and obstacles. One of the most widespread methodologies is machine learning algorithms. The techniques of machine learning contribute to both learning and prediction. Classical forms of machine learning involve varied degrees of human intervention to determine the output's accuracy. Algorithms for Machine Learning are a collection of components that collaborate to discover available patterns in a training dataset and detect unknown patterns in a new dataset.

Unsupervised learning and supervised learning are two approaches for machine learning. In machine learning, the supervised learning method is the most common. Renowned supervised learning methods include logistic regression, decision tree, support vector machine (linear kernel), support vector machine (RBF Kernel), and Random Forest. The evolution of technology has been altered by deep learning. It offers the most advanced prediction software available.

Deep learning models are trained with massive, labeled data sets and uncover patterns by analyzing data with a predetermined logical structure. In the era of Big Data, deep learning is essential for knowledge application and knowledge-based predictions.
Deep learning is utilized in computer vision, natural language processing (NLP), and predictive model development. Text categorization, sentiment analysis, translation, and speech recognition are among the most important NLP applications. Intelligent virtual assistants (Siri, Alexa, and Cortana) as well as adaptable email filters and chatbots are instances of NLP in the real world.

Whereas examples of predictive models driven by artificial neural networks (ANN) or convolutional neural networks (CNN) can be found on the most popular video streaming and e-commerce platforms. The potential of deep learning algorithms to finally learn from their own mistakes. It can evaluate the accuracy of its predictions and outputs and making any necessary adjustments. Deep learning is highly scalable because it has capacity to process vast volumes of data and execute a large variety of calculations in a cost- and time-efficient manner. CNN, LSTM, and RNN are the most well-known algorithms for deep learning.

In our thesis we also try to implement machine learning algorithms like - logistic regression, decision tree, support vector machine (linear kernel), support vector machine (RBF Kernel), Random Forest and deep learning algorithm - Recurrent neural networks. We were able to establish a comparison between the implemented algorithms by using as many algorithms as possible.

1.3 Research Problem

Due to environmental conditions and contemporary lifestyles, people often experience a various human disease nowadays. The majority of people in Bangladesh have a low standard of living and a limited education. In addition, the country has suffered a variety of infections caused by the natural spread of a wide range of pathogenic microbes. People in Bangladesh frequently consume enormous dosages of antibiotics not prescribed by doctors. Because of this, the level of pressure of bacterial infections to antibiotics is higher in our country. The high population density, lack of awareness of personal hygiene, and insufficient water distribution systems have all been linked to the advent of various diseases.

The world today suffers from several chronic diseases that cause death for many of the world's population. One of the most frequent disorders is thyroid gland disease, which is a very complex infection caused by high levels of (thyroid-stimulating hormone) or by difficulties with the thyroid organ itself. The existence of the thyroid gland and its diverse illnesses have been identified and medicated for ages. In the late nineteenth and early twentieth centuries, there were a variety of contemporary thyroid treatments and studies

These Diseases predicting has been one of the most difficult challenges in the medical industry in recent years. Analysis of medical data is essential for the development of new medical theories and the prevention of certain diseases. Diagnosing a patient's condition is seen as a crucial yet complex task that must be completed precisely and effectively.

The primary problem in modern healthcare is the provision of high-quality services and accurate, effective diagnoses. The medical field generates an increasing amount of data each year. The records of a huge amount of medical data compiled by medical professionals are available for analysis and knowledge extraction. Due to the increase in the amount of data in the medical and healthcare fields, early patient treatment has benefited from the proper analysis of medical data.

Using machine learning techniques, this huge amount of information or records in healthcare may be managed. The application of machine learning is expanding dramatically across all corporate and scientific fields, including medicine.

Classification models are ideally suited for classifying and distinguishing data classes. Under the supervised learning scheme, classification methods commonly utilized in machine learning research are applied to datasets pertaining to medical conditions. In the healthcare industry, machine learning is one such topic that is gaining steady recognition.
In healthcare, machine learning aids in the analysis of thousands of data points, the prediction of outcomes, the provision of timely risk scores, and the allocation of resources with precision, among other applications.

Machine learning examples are as follows:

1. Disease Identification and Diagnosis
2. Medical Imaging Diagnosis
3. Personalized Medicine
4. Digital Health Records
5. Clinical Research and Trials
6. Epidemic Prediction.

Deep learning has innovative applications in healthcare. Deep learning collects a vast quantity of data, such as patient records, medical reports, and insurance information, and applies neural networks to achieve the best results. Notable MDL uses include:

1. Medical imaging
2. Genomics analysis
3. Prescription analysis

Both of these methods play a key role in clinical healthcare systems. This method both minimizes misdiagnoses caused by human error and maximizes time utilization. There are machine learning (ML) and deep learning (DL) algorithms that can be implemented in the healthcare industry to reduce the amount of time required by doctors, surgeons, and other medical professionals for accurately evaluating, predicting, and diagnosing diseases.

The research aims to the following sectors:

- Choosing a dataset
- Preprocessing of data
- Using supervised machine learning to analyze a dataset.
- Using deep learning algorithms to analyze a dataset.
- Build classifiers in a distributed environment
- Perform comparative analysis

1.4 Hypothesis

The advent of machine learning (ML) technologies has revolutionized the field of medical diagnostics, particularly in the prediction and management of chronic diseases. The increasing complexity of healthcare data and the necessity for accurate, timely diagnoses have led to the exploration of innovative computational methods.

This thesis hypothesizes that implementing advanced machine learning algorithms, specifically Logistic Regression, Random Forest, and Support Vector Machines, in conjunction with robust data preprocessing and feature selection techniques, will significantly improve the accuracy and reliability of thyroid disease predictions compared to conventional diagnostic methods.

1. Accuracy of Machine Learning Algorithms: We propose that machine learning models, particularly ensemble methods like Random Forest, will outperform traditional clinical assessment methods in terms of predictive accuracy for thyroid disease. This hypothesis is grounded in the expectation that these models can effectively capture complex non-linear relationships within the data, leading to better identification of at-risk patients.

Traditional methods often rely on heuristic approaches and established clinical guidelines, which can overlook subtle patterns in data. By leveraging the computational power of machine learning algorithms, we anticipate that our models will discover these patterns and provide more nuanced insights into patient risk factors.

Furthermore, the performance of these models will be evaluated not only on accuracy but also on other metrics such as precision, recall, and F1 score. We expect that machine learning algorithms will not only increase overall accuracy but also enhance the identification of true positives, which is crucial in a clinical setting where missing a diagnosis can have significant consequences for patient health.

2. Impact of Feature Engineering: In addition to the choice of algorithms, this study posits that thoughtful feature engineering plays a critical role in model performance. By integrating hormonal levels (such as TSH, T3, and TT4), demographic data, and lifestyle factors, we aim to develop models that reflect the multifaceted nature of thyroid health. Feature engineering involves not just selecting which variables to include but also transforming and creating new variables that better capture underlying patterns in the data. For example, interactions between certain hormonal levels may provide additional predictive power that is not evident when examining these variables independently.

We will systematically evaluate the contribution of each feature through techniques such as Recursive Feature Elimination (RFE) and correlation analysis. This approach will help identify the most significant predictors of thyroid disease and allow us to refine our models to improve predictive accuracy further.

3. Role of Class Imbalance: In the context of medical diagnostics, particularly with thyroid disease, class imbalance is a significant challenge. Thyroid conditions may present in varying degrees of severity and prevalence, leading to datasets where one class (e.g., healthy individuals) significantly outnumbers another (e.g., those with thyroid disease). This research hypothesizes that employing techniques like the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance within the dataset will lead to an increase in the model's recall and F1 score. Specifically, we expect that these techniques will improve the model's ability to correctly identify patients with thyroid disease, thereby reducing the risk of false negatives.

Class imbalance can skew the performance of machine learning models, leading to high accuracy while still missing a substantial proportion of positive cases. By balancing the classes through SMOTE, we anticipate that our models will become more sensitive to the minority class, thereby enhancing their clinical utility.

4. Adaptive Learning Techniques: In addition, we propose that applying adaptive learning techniques, such as Adaptive Moment Estimation (Adam) during the model training process, will optimize the convergence speed and accuracy of the machine learning algorithms. The use of adaptive learning rates allows the model to adjust based on the error landscape during training, facilitating a more efficient learning process.

We believe that this approach will result in more robust models capable of generalizing well to unseen data, which is a critical requirement in clinical settings where new patients may present different characteristics. The optimization process is crucial not only for improving performance but also for reducing the likelihood of overfitting. By utilizing techniques that promote generalization, we expect to develop models that can adapt to variations in patient data without losing predictive power.

5. Comparative Analysis of Algorithms: Lastly, we hypothesize that a comparative analysis of multiple machine learning algorithms will reveal distinct strengths and weaknesses, providing insights into the most effective strategies for thyroid disease prediction.

By systematically evaluating the performance of these algorithms, we aim to identify best practices that can be adopted in clinical settings. The outcomes of this analysis will contribute to a deeper understanding of how different algorithms perform in relation to one another and under varying conditions, thereby informing future research and clinical practice.

In conclusion, this thesis aims to demonstrate that the integration of machine learning techniques, data preprocessing, and feature engineering will not only enhance the predictive performance for thyroid disease but also contribute to more informed clinical decision-making processes. By establishing a comprehensive framework for utilizing machine learning in the early detection and management of thyroid disorders, we hope to bridge the gap between computational advancements and practical applications in healthcare.

The ultimate goal is to empower clinicians with reliable predictive tools that can lead to timely interventions, improved patient outcomes, and a deeper understanding of thyroid health dynamics.

1.5 Objectives

The incidence of diseases of the thyroid is on the rise. Thyroid hormone combines with numerous signaling pathways, and thus its function is controlled by nutrition and iodine status. Thyroid hormone activates a large number of genes after its conversion from the prohormone thyroxine (T4) to the active form triiodothyronine (T3).

Resistance to thyroid hormone (RTH) was first defined in 1967 as the primary clinical disease associated with reduced nuclear effect of thyroid hormone. It is demonstrated that the impact of thyroid hormone level and action represents is a difficult challenge. We conducted study on thyroid disease due to our curiosity and the numerous diagnostic difficulties associated with it. Although the ingredients necessary for thyroid hormone function are widely understood, the interaction between the multiple pathways has been complicated to understand.

The hypermetabolic and hypometabolic crises are described to as a "thyroid storm." In an effort to minimize the severe complication of "thyroid storm," numerous studies and developments have been executed in this field. Numerous studies in the literature

focus on predicting thyroid disease based on the hormonal parameters of people's tendencies. Thyroid crisis can still occur in patients with uncontrolled hyperthyroidism due to a trigger such as operations, illness, or trauma.

In an effort to model these disease prediction difficulties using statistical techniques and, more recently, machine learning and deep learning techniques, a group extremely devoted researchers' study have come to our assist. Methods such as data mining, big data, deep learning, and machine learning are utilized to diagnose thyroid disease. Methods for data mining and machine learning require labeled data for training.
The value of data and finding knowledge is clarified through the use of data mining and processing techniques. Quantity of data is essential for achieving greater precision. New data science terms include data mining, big data, and statistics, among others. Big data and deep learning have produced new branches of research over time. These two have become highly valuable as a result of the data science field's efforts to increase throughput.

The purpose of this study is to predict thyroid disease so that general public can realize the benefits of scientific advances in computational methodologies. We proposed a general disease prediction structure based on the patient's symptoms. This study's primary focus is the prediction of disease using machine learning approaches.
Clearly, machine learning has the potential to transform the healthcare industry. The expanding number of machine learning applications in healthcare enables us to envision a future in which data, analysis, and technology cooperate to treat countless people without their involvement.

In the near future, it will be commonplace to discover ML-based applications incorporating real-time patient data from various healthcare systems in several regions. Utilizing distributive functions, the processing of patient data to calculate the probability of developing thyroid has been statistically computed.

For the prediction of thyroid, we have established three steps: data collection, preprocessing, and classification. The logistic regression (LR) method is used to categorize data instances for classification purposes. In addition, Adaptive Moment

Estimation (Adam) and an adaptive learning rate optimization technique are used to tune the parameters of LR.

The widespread usage of deep learning in recent days has resulted in a notable increase in prediction and analysis. Therefore, we also implemented a deep learning technique RNN algorithm for high precision. Random Forest (RF), decision tree, support vector machine (linear kernel), and support vector machine (RBF Kernel) techniques are all applied to the same dataset, and all outcomes have an approximation of the modern and most accurate accuracy.

In order to identify the optimal model for thyroid disease identification, the implementation of each algorithm must be carefully evaluated in the context of the above objectives.

Chapter 2 Literature Review

2.1 Literature Review

In recent years, a large amount of work has been dedicated to determining the different disorders of the thyroid. Previous researchers have utilized various data mining techniques. By using a wide range of datasets and algorithms, as well as looking ahead to the work that needs to be done to achieve even better results, the authors demonstrated that they had obtained an adequate approach and certainty to identify disorders similar to thyroid disease. The development of widely used machine learning-based illness prediction mechanism makes use of several different machine learning algorithms, such as Random Forest, Decision Tree, Naive Bayes, SVM, and ANN. There have been several research and review articles concerning thyroid disease have been published. In this chapter, we will therefore present an overview of some of the most important research and studies on the prediction and analysis of thyroid disease. Some of the most prominent articles that inspired us the current study are discussed below:

Obermeyer [1] et al. recommended Machine learning (ML) is a division of artificial intelligence and is infiltrated in the dimensions of scientific research at growing steps. Machine learning facilitates algorithms to review from experience without notably being prioritized. Artificial intelligence (AI) and the progressive starting of AI research in the medical field have enabled people to glimpse the promising future of AI and healthcare integration. The interest in utilizing Artificial Intelligence (AI) and Machine Learning (ML) approaches to performance better and obtain large datasets to enhance healthcare.

Ghahramani [2] et.al proposed that For unsupervised mechanism of learning, only unlabeled data are feasible and the algorithms peeks to asset the analogies and devices, unsupervised learning algorithms may catch the vast number of unlabeled genomics data as input and analyze formerly anonymous assemblage of data. These algorithms may somehow be dominant in previously formerly arrangements in complex data that are not primarily measurable by humans. Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it

closer to the real AI. Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.

Chaubey [3] et al. proposed a study to compare the performance of logistic regression, decision trees and k-nearest neighbor algorithms for predicting and evaluating thyroid disorder in terms of accuracy. UCI Machine Learning repository was used as a data source and the dataset was Divided into three parts: training (70%), validation (15%) and test (15%). For decision tree algorithm, the two main thyroid hormones which are commonly reporting disorder and hence creating further complains at the body namely: triiodothyronine (T3) and total serum thyroxin (T4), used as feature. According to the study, k-NN approach was found better with 96.875% accuracy for the studied dataset. The authors reported that UCI Machine Learning Repository has more than one thyroid disease dataset.

Dogantekin [4] et al. proposed an automatic diagnosis system based on thyroid gland (ADSTG) method. The structure of ADSTG has three stages. The first stage is feature reduction by using Principal Component Analysis method. The second stage is the classification by using Least Square Support Vector Machine classifier. And third stage is the performance evaluation of the proposed method for diagnosis of thyroid disease. It is evaluated by using classification accuracy, k-fold cross-validation, and confusion matrix methods, respectively. The classification accuracy of the proposed method was obtained about 97.67% with 10-fold cross validation.

Menger [5] et al. showed Deep learning approaches have been applied for the prediction of violent incidents by patients. Deep learning has demonstrated superior performance in recognition, classification, segmentation, and prediction of the future status of time-variable data compared to conventional approaches. Simultaneously, deep neural networks have gained widespread popularity, particularly following the development of machine learning libraries.

Prasad [6] et al., the researchers have developed the hybrid architecture system using rough data sets theory and machine learning algorithms to predict thyroid diseases.

Raghuraman et al. in [7] performed comparative thyroid disease diagnosis using Machine learning techniques—Support Vector Machine (SVM), Multiple Linear Regression and Decision Trees, and the highest accuracy of 97.97% was obtained by the decision tree model.

Dharamrajan et al. in [8] applied Support Vector machine (SVM) and Detection tree classifier for thyroid prediction and obtained an accuracy of 97.35 using decision trees.

Dhyan Chandra Yadav in [9] proposed the prediction of thyroid diseases using a decision tree ensemble approach. They implemented two stage approach gives a maximum accuracy of 99.95% which is very good as compared to existing techniques. The thyroid disease dataset consists of 3152 cases, 23 characteristics and finally a class to predict whether the individual is ill or not.They present techniques and experimental set-up used for this task. The workflow was with preprocessing of data, then applying dimension reduction and data augmentation techniques. After this, classifiers are implemented in a distributed environment, and finally comparative analysis was performed. They described the three-dimension reduction techniques. Dimensionality reduction is a method for obtaining the information with lesser number of dimensions from a high dimensional feature space. In machine learning it is very important for the better classification, regression, presentation, and visualization of data to reduce the high-dimensional data collection. In this paper they have shown that dimension reduction and data augmentation can be used very efficiently for achieving high accuracy of disease prediction. To make the size of data large enough for building a deep learning neural network model, they have applied data augmentation.

Ghali [10] et al. (2020) proposed using artificial intelligence-based algorithms and multi-linear regression to predict thyroid regulating hormone balance in human.

Irina and Liviu [11] (2016) propose a data mining technique for classifying thyroid diseases. The study focuses on the categorization of thyroid illness in two common types of thyroid condition (hyperthyroidism and hypothyroidism) in the population [16].

Geetha and Baboo [12] proposed an approach that focuses on thyroid disease classifying. Two of the most common thyroid diseases are hyperthyroidism and hypothyroidism among the public. Dataset of the study was provided from UCI repository. First, the data was pre-processed. The pre-processed data is multivariate in nature. The dimensionality was decreased using Hybrid Differential Evolution Kernel Based Naïve Based algorithm so that the available 21 attributes is optimized to 10

attributes. Then, the subset of data was provided to the Kernel Based Naïve Bayes classifier algorithm to verify the fitness. It was mentioned that detection accuracy was 97.97% [22].

Begum and Parkavi [13] studied the prediction of thyroid illness [24]. The purpose of the study was to predict thyroid disease using different classification methods. The authors were experienced with data mining strategies such as K-Nearest Neighbor (k-NN), Support vector machines (SVM), ID3, and naive Bayes (NB). The UCI dataset consists of 15 attributes. The purpose of the study was to determine the relationship between thyroid hormones T3, T4, and TSH and gender in relation to hyperthyroidism and hyperthyroidism. However, neither the performance nor the outcomes of the approaches and measurements are demonstrated in the paper. The study highlights the significance of employing data mining methods on medical data to improve treatment's speed, accuracy, and cost.

Shroff [14] et al. offered an assessment of previous work with semi-automatic medical diagnosis in general and thyroid disease diagnostics specifically. Medical Diagnosis included the use of classifiers such as Fuzzy Neural Networks, k- Nearest Neighbor, and Decision Tree, whereas the document included the use of Computer-Aided Diagnosis, various Neural Networks, and Support Vector Machine, as well as the impact of feature selection on classification using particle Swarm Optimization and Ant Colony Optimization. The authors proposed conducting an experiment using kNN with all distances (Euclidian, Manhattan, Mahalanobis) on the UCI thyroid dataset [19]. As a result, the authors have noted that early disease detection is crucial for patients' awareness and prevention.

After reviewing the literature on thyroid disease prediction and analysis, it is evident that various techniques such as machine learning, deep learning, data mining, and other forms of artificial intelligence have been applied. However, it has been observed that deep learning, despite its growing popularity, has not been sufficiently explored in the context of thyroid disease detection. On the other hand, machine learning approaches have been more extensively researched and widely implemented in this field. This gap in research suggests an opportunity for further exploration and development of deep learning methods for more accurate and reliable thyroid disease prediction.

Chapter 3 Machine Learning

3.1 Machine Learning Algorithms

In numerous fields of technology, learning has been the focus of extensive research. Learning is the foundation of all human activities, including cooking, fixing broken devices, swinging, playing sports, building a house, and many more. Thus, in order to accomplish any career, it is necessary to learn the profession beforehand so that the brain acquires the required skills to execute the task with no or little error. The human brain has a more complicated learning system than any other intelligent system. Neural cells in the human body transform sensory information into signals that are sent to the brain. After taking in information from the environment, the brain forms behaviors through the application of judgment and inference of events. Despite the outstanding performance of machine learning approaches in games such as Go and chess, they do not appear to provide sufficient insight into how humans make decisions and learn in the real world. One of the most significant aspects of the human learning process is that it is painfully slow. This may be due to the particular information processing characteristics that humans have.

There are two purposes of artificial intelligence. First, AI is focused on making computers intelligent and capable of performing intelligent tasks so that humans are not required to perform them. And secondly, AI is also geared toward simulating humans with computers so that we can learn how humans work and potentially assist them in improving their performance. Machine Learning is the application of artificial intelligence (AI) that enables computers to self-grow, self-modify, and self-learn when presented with new data.

Machine learning is utilized in the training of the program over large databases. In Machine Learning, a computer program is tasked with performing certain tasks, and it is claimed that the machine has learned from its experience if its measured performance in these tasks improves as it obtains more and more experience executing them. Consequently, the machine makes decisions and makes predictions based on available information. A significant advantage of machine learning is that algorithms can learn information directly from input without depending on

predetermined equations as models that can easily analyses even large amounts of dataset.

Depending on the availability of training data types and categories, to implement the most appropriated machine learning method may require the implementation of "supervised learning," "unsupervised learning," or "semi-supervised learning" techniques.

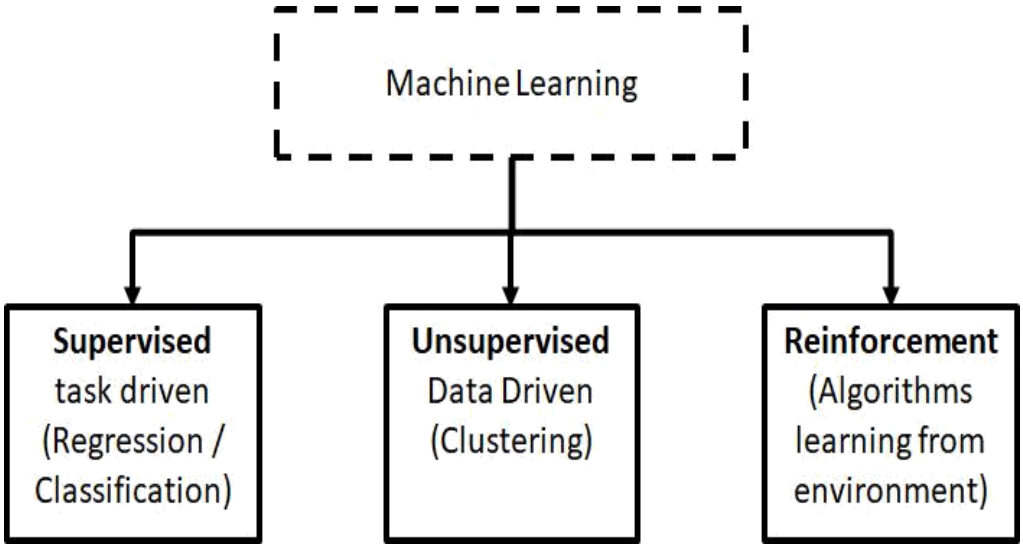

Figure 3.1: Classification of Machine Learning

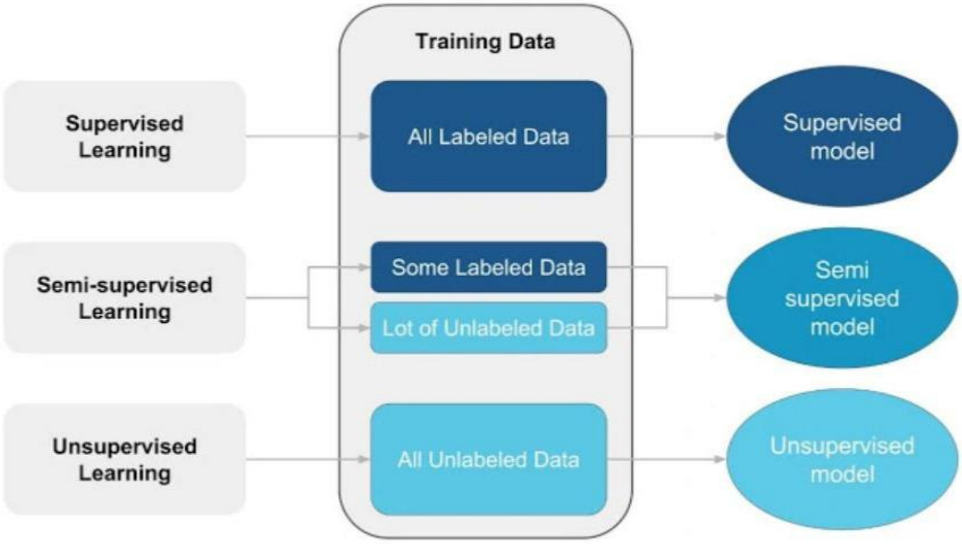

Figure 3.2: Classification of supervised, semi-supervised and unsupervised learning algorithms

3.1.1 Supervised MLTs

There are a variety of well-differentiated approaches within Machine Learning methodologies. The purpose of supervised algorithms that operate on labeled data sets is to develop predictive model, classification, or regression models. This type of learning involves the algorithm learning to assign labels to data input categories depending on the labels input by a human during training. A supervised learning algorithm analyzes the training data and generates an indirect function that can be used to map new illustrations. In a supervised learning technique, the input data is labeled to arrange the data. It provides a computer system with a training set of instances with appropriate objectives. The computer can analyze input-output examples and training the model to accurately fit the data. Categories of Supervised Learning include classification and regression.

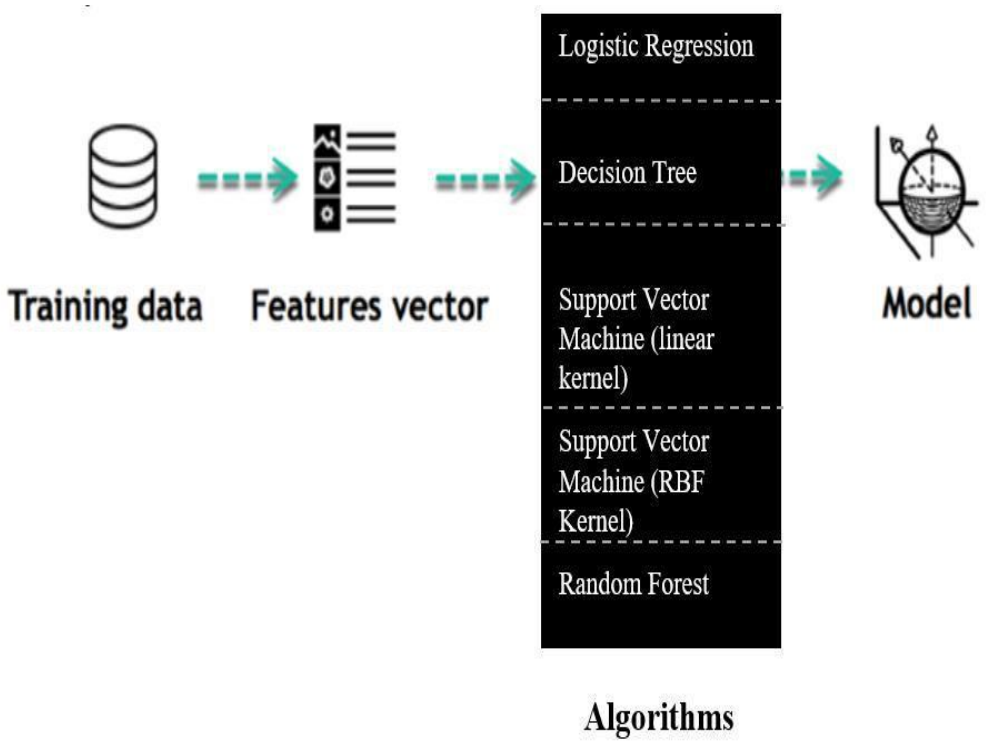

Figure 3.3: Workflow of Supervised Machine Learning Technique

• Using classification methods, inputs are divided into multiple classes, and the trained system must produce actions that assign hidden inputs to these classifications. This is known as the multiple labeling process. The purification of spam is an example of classification in which emails are categorized as "spam" or "not spam."

• Regression is a supervised method with continuous rather than discrete outcomes. In contrast to classification predictions, which use accuracy as a performance indicator, regression predictions are evaluated using root mean squared error (RMSE).

Some of the supervised learning algorithms that we will apply in our thesis are:

1. logistic regression
2. decision tree
3. support vector machine (linear kernel)
4. support vector machine (RBF Kernel)
5. Random Forest

- **Logistic regression:**

Logistic regression is one of the basic and popular algorithms to solve a classification problem. 'Logistic regression' is called logistic regression because its underlying technique is quite the same as Linear Regression. From Logit function the term 'Logistic' is being taken.

It predicts a binary outcome (yes/no, true/false, 1/0) from a given set of independent variables. It's a special case of linear regression when the outcome variable is categorical, here we use log of odds as independent variables.

Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model. In Logistic Regression, we are only concerned about the probability of outcome dependent variable (success or failure). The dependent variable is also called target variable. It predicts the outcome variable by using log function. Generally, we use sigmoid function to predict the outcome.

This function is established using two things: Probability of Success (p) and Probability of Failure (1-p). P should meet following criteria:

1. It must always be positive (since p >= 0)
2. It must always be less than equals to 1 (since p <= 1)

Sigmoid Function:

$$P = \frac{1}{1 + e^{\{-y\}}}$$

equation (1)

Where,

$$\mathbf{Y} = \mathbf{w^T x} + \mathbf{b}$$

equation (2)

y (or sometimes written as z) is the linear combination of the input features

- w^T is the vector of weights,
- x is the vector of input features,
- b is the bias (or intercept) term.

The threshold value of sigmoid function decides the outcome.

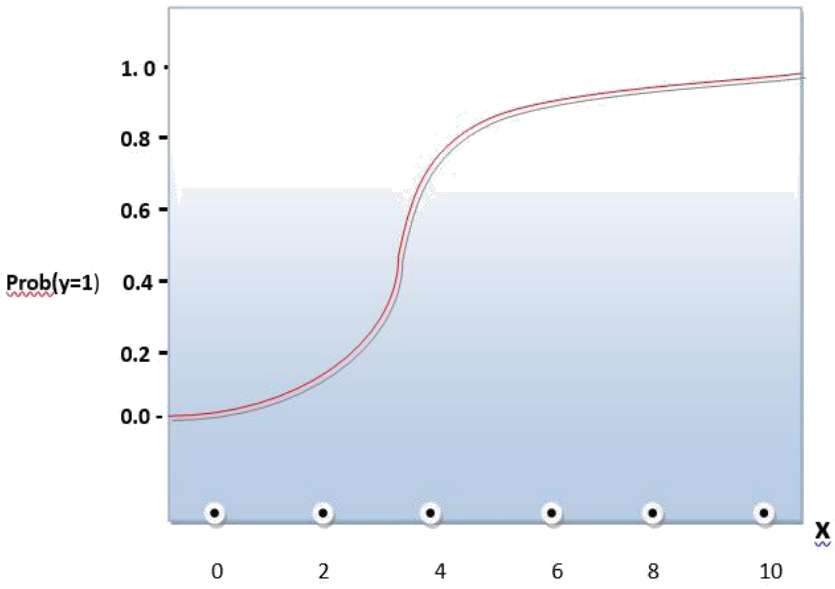

Figure 3.4: Logistic Regression Function

Here in the above figure, we can see, the result isn't going below 0 & above 1.

- **Decision Tree Algorithm:**

Decision Tree algorithm belongs to the family of supervised learning algorithms. Decision tree algorithm can be used for solving regression and classification problems unlike other supervised learning algorithms. We use Decision Tree to create a training model that can use to predict the class or value of the target variable. The classes of the test data are estimated based on the set of decision rules.

Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. When it creates sub-nodes, the homogeneity of resultant sub-nodes increases.

Working process:

The algorithm works on some stopping criteria. The criteria are:

1. If all the leaf nodes are labeled then it will stop.
2. If Maximal node is attained.
3. If there is no information to gain on splitting of any node then the algorithm will stop functioning.

Now, let's see how the algorithm works:

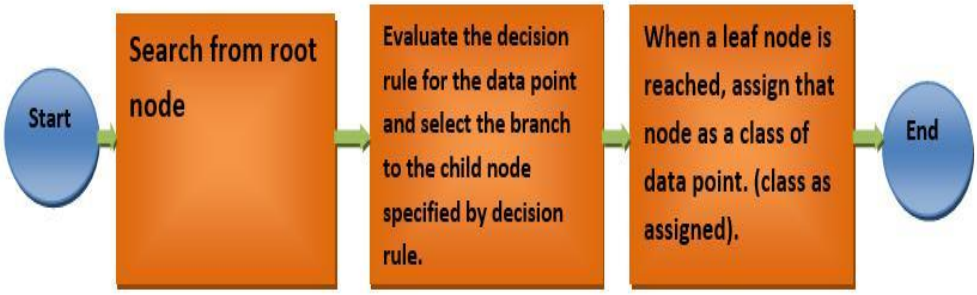

Figure 3.5: Work process of decision tree

First we need to input dataset for each feature. If leaf nodes don't satisfy the early stopping criteria then it selects attribute that gives the best split and assigned it as the root node. Now, repeat the process. It begins from the root node as the parent node. Then split parent node at some feature to maximize the information gain. Assigning training samples to new child nodes and it remains until for each new child node.

Output: Classified samples with labels allocate

- **Random Forest Algorithm:**

Random forest algorithm is a very popular ML algorithm which belongs to the supervised learning technique. This algorithm is used for both Classification and Regression problems. This algorithm is based on 'ensemble learning' concept.

Ensemble learning is a process where it processes multiple classifiers for solving a complex problem and it also enables the performance of the model. The larger number of trees in the forest belongs to higher accuracy and prevents the over fitting problem. Random forest takes the prediction from each tree instead of one decision tree and from all the predictions the majority vote comes, and it predicts the final output. From the graph given below we can see the works of this algorithm.

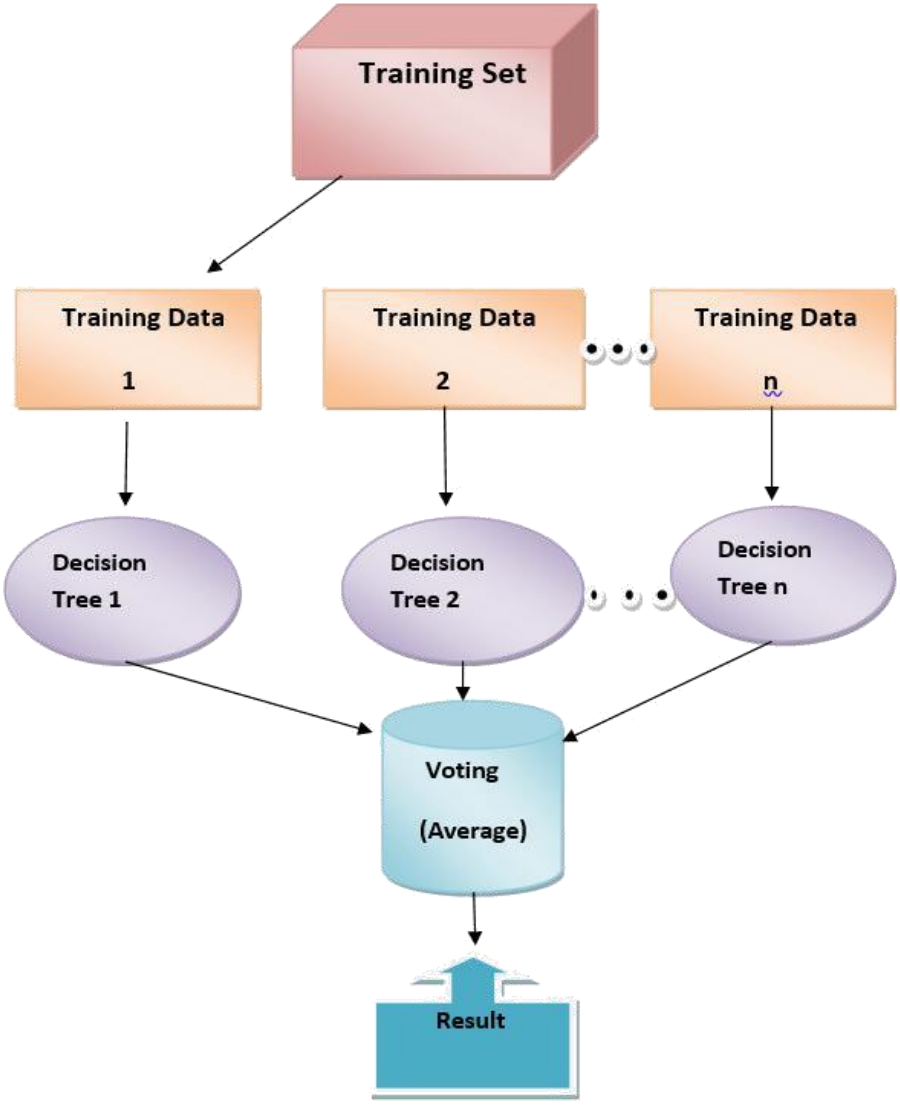

Figure 3.6: Working flow of Random Forest

Working Procedure:

Random Forest follows two phases to predict the result. Firstly, a random tree has to be created with combining n decision tree and second or final face make prediction for each tree created in the first phase.
The steps to be followed on Random Forest are given below:

- Selection of k random points from training set.
- Then building the trees associated with selected data points.
- Then building the trees associated with selected data points.
- Then repeat 1 & 2.
- Finding the predictions of each decision tree and assigning the new data points, then the winner is the category that wins the majority of votes.

- **Support vector machine (linear kernel):**

Support vector machine are regarded as a diverse system algorithm. When it comes to doing in-depth research and accurate analysis, support vector machines are widely regarded as a go-to tool.

Linear Kernel is utilized when the data is linearly separable or can be separated by a single Line. It is one of the most commonly utilized kernels. It is typically utilized when a Data Set contains a large number of Features. If a dataset can be divided into two classes using a single straight line, it is referred to as linearly separable data, and the linear classifier is called SVM classifier is that is implemented. This model was based on by the well-known statistical learning theory.

Support Vector Machines SVMs can perform both problems of classification and regression. The objective of SVM is to correctly classify objects using examples from the training data set. SVM has the ability to handle both semi-structured and structured data, as well as complex functions if the corresponding kernel function can be generated. As generalization is implemented in SVM, the probability of overfitting decreases. It is scalable with respect to high-dimensional data. It does not become deeply involved in local optima.

SVM has the following disadvantages: its performance decreases with larger data sets as training time increases. It will be challenging to locate an acceptable kernel function. SVM does not operate well when dataset is noisy. When it comes to probability estimates, SVM is useless, it cannot provide probability estimates. Capable of understanding the final SVM model is challenging.

Support Vector Machine has applications in cancer diagnosis, credit card fraud detection, handwriting recognition, face identification, and text categorization, among others. So, among the three approaches of Logistic Regression, Decision Tree, and SVM, logistic regression will be attempted first, followed by decision trees (Random Forests) to determine if there is a substantial improvement.

SVM can be utilized when the number of observations and characteristics is large. Powered by a proper kernel, SVM is enabled to deal with not only linearly separable problems but also linearly non-separable problems. In comparison, Nonlinear SVM is a black-box classifier for which the mapping function is not explicitly known.

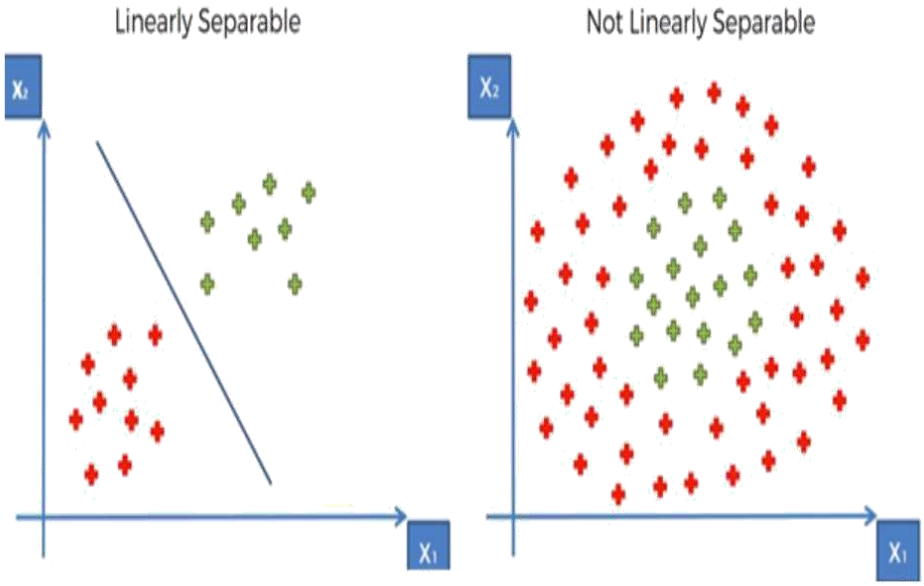

Figure 3.7: Linear SVM & NON-Linear SVM

- **Support vector machine (RBF Kernel):**

Linear kernel Support Vector Machine Radial Basis Function (RBF) kernel is a well-known technique for feature selection. Kernel function parameter selection is a key component in support vector machine (SVM) modeling. Although there are several other kernel functions to use with an SVM classifier, the RBF kernel is the standard and best option. It is default and recommended kernel function.

Suppose U ∈ Rn, V ∈ Rn, g ∈ R+, where g is an RBF kernel hyper-parameter.

RBF kernel is denoted by the formula:

$$\mathbf{K(U, V) = \exp(-g||U - V||2)}$$

equation (3)

RBF is the default kernel used by the SVM classification algorithm in sklearn and can be stated as follows:

Where, gamma can be set manually and must be greater than 0(gamma>0)

Support vectors are all that RBF Kernel Support Vector Machines need to keep during training, hence this method is ideal for this task not for full dataset.

3.1.2 Unsupervised MLT

Unsupervised learning is also known as unsupervised machine learning which uses the machine learning algorithms to analyze and cluster unlabeled datasets. Like supervised learning there is no correct answer and there is no teacher in unsupervised learning.

In unsupervised technique algorithms are left to their own devices and welcomed to discover, present the interesting structure in the data. As a result, when analyzing these reviews a clustering strategy is employed where groups elements of comparable sorts into a cluster. To reach the effectiveness a variety of evaluation parameters are being evaluated.

Unsupervised learning algorithms learn a few things from the data. To address the class of data it uses the previously learned features, when new data is introduced.

In the classification of text, clustering is the most often used unsupervised learning technique. Clustering can be done in a variety of ways, with K-means clustering being one of the most prominent.

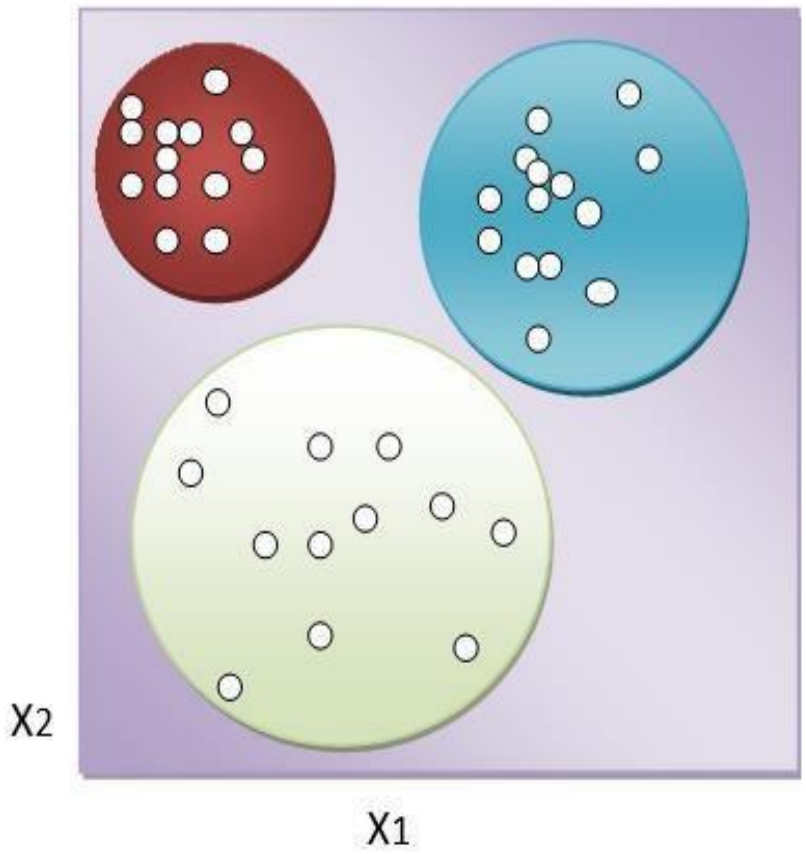

Figure 3.8: Unsupervised Learning

3.1.3 Semi-supervised MLT

This hybrid method of machine learning uses both supervised and unsupervised techniques, semi-supervised is a combination of supervised and unsupervised. It uses a small number of labeled data and a large number of unlabeled data, which provides the advantages of both unsupervised and supervised learning while avoiding the difficulties associated with locating a large amount of labeled data. This implies that you can train a model to classify data with less labeled training data. Labeled data is not fundamental issue for model.

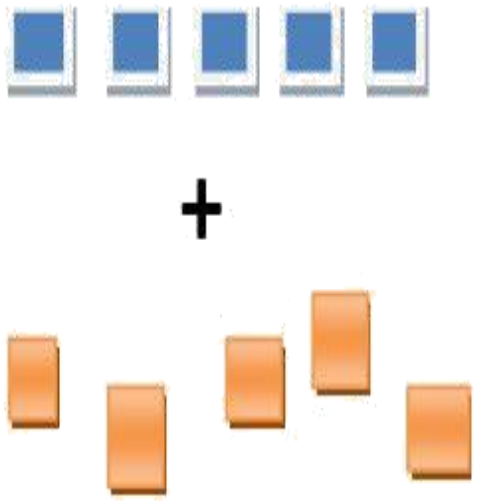

Figure 3.9: Labeled data and Unlabeled data

Chapter 4 Deep Learning

4.1 Deep learning

Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans: learn by example. Deep learning is a key technology behind driverless cars, enabling them to recognize a stop sign, or to distinguish a pedestrian from a lamppost. It is the key to voice control in consumer devices like phones, tablets, TVs, and hands-free speakers. Deep learning is getting lots of attention lately and for good reason. It’s achieving results that were not possible before.

In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers.. Deep learning has aided image classification, language translation, speech recognition. It can be used to solve any pattern recognition problem and without human intervention. [33]

When we talk about deep learning a word that is "Neural Network" comes to our mind first because it is an important part of deep learning. Neural networks mimic the human brain via a fixed of algorithms. Originally inspired by medical science, deep neural network models have become a powerful tool of machine learning and artificial intelligence.

4.2 Deep Neural Networks

A deep neural network (DNN) is a multiple hidden layers between the input and output layers. Deep neural networks (DNNs) are currently widely used for many artificial intelligence (AI) applications including computer vision, speech recognition, and robotics.

The main purpose of a neural network is to receive a set of inputs, perform progressively complex calculations on them, and give output to solve real world problems like classification. We restrict ourselves to feed forward neural networks. [34]

Neural networks are widely used in supervised learning and reinforcement learning problems. These networks are based on a set of layers connected to each other.

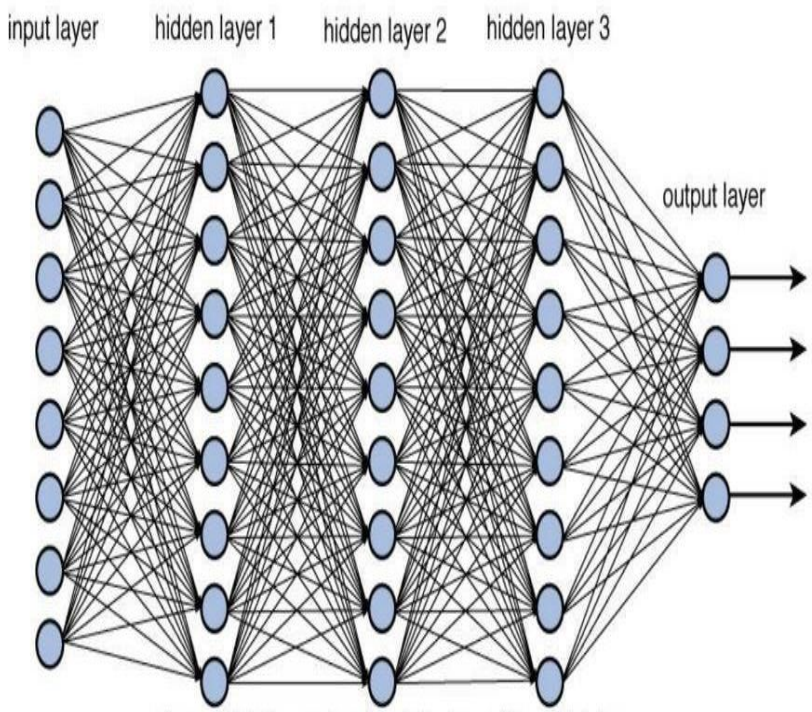

Figure 4.1: Deep Neural Network

4.3 Rise of Deep Learning

Machine language is said to have occurred in the 1950s when Alan Turing, a British mathematician, proposed his artificially intelligent "learning machine." Arthur Samuel wrote the first computer learning program. His program made an IBM computer improve at the game of checkers the longer it played. In the decades that followed, various machine learning techniques came in and out of fashion.[33]

Neural networks were mostly ignored by machine learning researchers, as they were plagued by the 'local minima' problem in which weightings incorrectly appeared to give the fewest errors. However, some machine learning techniques like computer

vision and facial recognition moved forward. In 2001, a machine learning algorithm called Ad boost was developed to detect faces within an image in real-time.

Neural networks did not return to favor for several more years when powerful graphics processing units finally entered the market. The new hardware-enabled researchers to use desktop computers instead of supercomputers to run, manipulate, and process images. The most significant leap forward for neural networks happened because of the introduction of substantial amounts of labeled data with ImageNet, a database of millions of labeled images from the Internet. The cumbersome task of manually labeling images was replaced by crowdsourcing, giving networks a virtually unlimited source of training materials. In the years since technology companies have made their deep learning libraries open source.[33]

Examples: Google TensorFlow

4.4 RNN - Recurrent Neural Networks

Recurrent neural networks (RNN) are the state-of-the-art algorithm for sequential data. It is the first algorithm that remembers its input, due to an internal memory, which makes it perfectly suited for machine learning problems that involve sequential data. It is one of the algorithms behind the scenes of the amazing achievements seen in deep learning over the past few years. In this post, we'll cover the basic concepts of how recurrent neural networks work, what the biggest issues are and how to solve them.[36]

RNNs are a powerful and robust type of neural network, and belong to the most promising algorithms in use because it is the only one with an internal memory. Because of their internal memory, RNNs can remember important things about the input they received, which allows them to be very precise in predicting what's coming next.

This is why they're the preferred algorithm for sequential data like time series, speech, text, financial data, audio, video, weather and much more. Recurrent neural networks can form a much deeper understanding of a sequence and its context compared to other algorithms.

4.4.1 When we need to use RNN

Whenever there is a sequence of data and that temporal dynamics that connects the data is more important than the spatial content of each individual frame. RNN can be used to create a deep learning model that can translate a text from the source language into the target language without human intervention.[37]

4.4.2 Types of RNNs

- One to One
- One to Many
- Many to One
- Many to Many

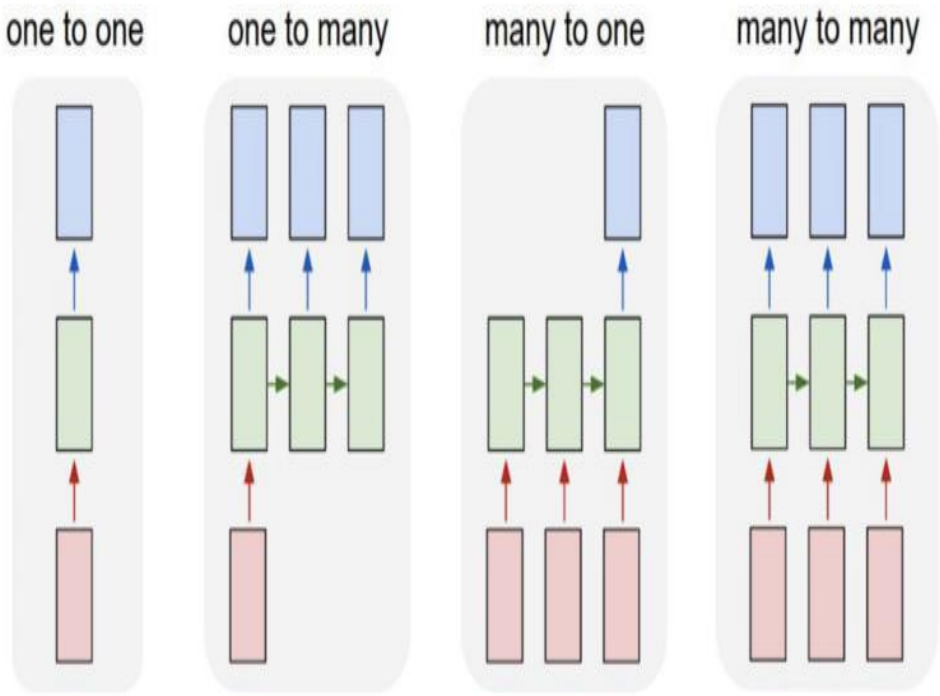

Figure 4.2: Types of RNN's

4.5 Long Short-Term Memory (LSTM)

Long short-term memory networks (LSTMs) are an extension for recurrent neural networks, which basically extends the memory. LSTMs assign data "weights" which

helps RNNs to either let new information in, forget information or give it importance enough to impact the output. The units of an LSTM are used as building units for the layers of a RNN, often called an LSTM network.
LSTMs enable RNNs to remember inputs over a long period of time. This is because LSTMs contain information in a memory, much like the memory of a computer. The LSTM can read, write and delete information from its memory.

4.6 How deep learning works

To understand RNNs properly, we'll need a working knowledge of "normal" feed-forward neural networks and sequential data. Sequential data is basically just ordered data in which related things follow each other. Examples are financial data or the DNA sequence. The most popular type of sequential data is perhaps time series data, which is just a series of data points that are listed in time order.

RNNs and feed-forward neural networks get their names from the way they channel information. In a feed-forward neural network, the information only moves in one direction — from the input layer, through the hidden layers, to the output layer. The information moves straight through the network.

Feed-forward neural networks have no memory of the input they receive and are bad at predicting what's coming next. Because a feed-forward network only considers the current input, it has no notion of order in time. It simply can't remember anything about what happened in the past except its training.

In a RNN the information cycles through a loop. When it makes a decision, it considers the current input and also what it has learned from the inputs it received previously.[36]

4.6.1 Input Layer

The input layer of a neural network is composed of artificial input neurons, and brings the initial data into the system for further processing by subsequent layers of artificial neurons. The input layer is the very beginning of the workflow for the artificial neural network

4.6.2 Hidden Layer

In neural networks, a hidden layer is located between the input and output of the algorithm, in which the function applies weights to the inputs and directs them through an activation function as the output.

In short, the hidden layers perform nonlinear transformations of the inputs entered into the network. In artificial neural networks, hidden layers are required if and only if the data must be separated non-linearly.

4.6.3 Output Layer

The output layer is the final layer in the neural network where desired predictions are obtained. There is one output layer in a neural network that produces the desired final prediction. It has its own set of weights and biases that are applied before the final output is derived.

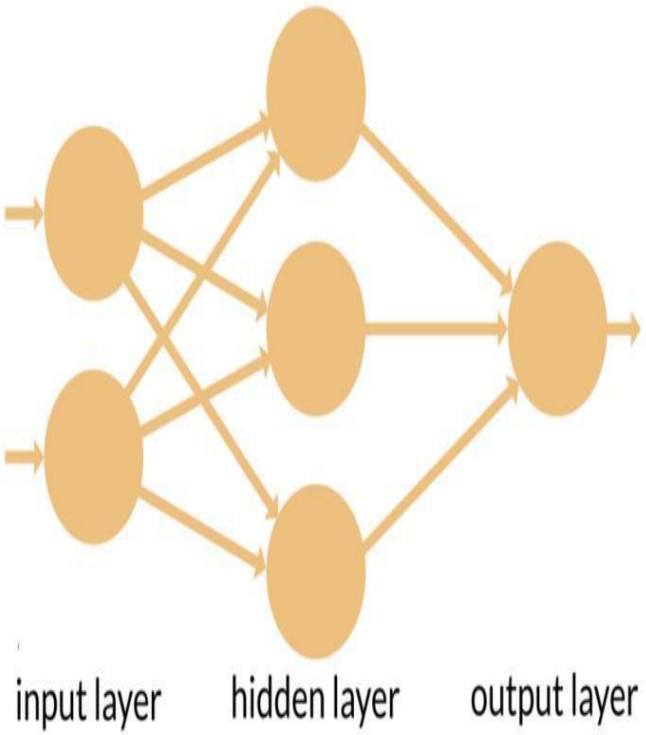

Figure 4.3: Deep learning architecture

Chapter 5 Machine learning vs Deep Learning

5.1 Advantages of using Machine learning

Learning is a natural human characteristic that has been introduced into robots as well. Due to the vast quantity of structured, unstructured, and semi-structured data, ML has grown in popularity over the past few years. For cost savings, faster processing of vast amounts of data, and faster identification of new vulnerabilities, ML is gaining significant appreciation.

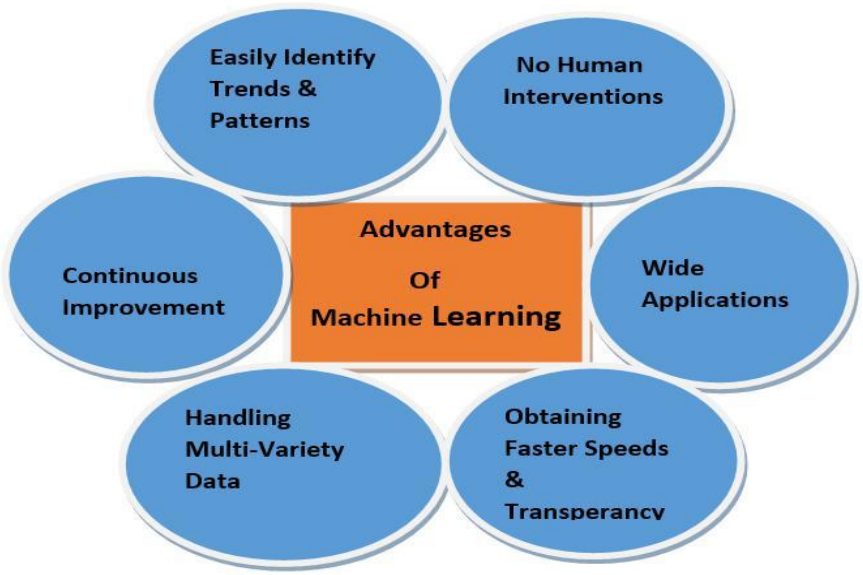

Figure 5.1: Advantages of Machine learning

Some of the benefits and advantages to gain through the usage of ML are:

- **Continuous Improvement:** Humans are susceptible to inaccuracy, especially when dealing with huge amounts of data. But in this case. If machines are programmed correctly, they are less prone to make mistakes. The primary advantage of using ML is that the resulting errors and earlier computations may be used to generate more trustworthy, accurate, and predictable data. The effectiveness and precision of ML algorithms continually rises as they acquire more and more training data. This allows them to make better choices. As the

amount of available data increases, your algorithms learn to generate more accurate predictions more quickly.

- **Obtaining faster speeds and transparency:** In a world where businesses must adapt to meet the expectations of customers who are constantly shifting their preferences, timing is everything. The increased transparency made possible by machines and algorithms enables for better future decision making. Therefore, the faster a company obtains customer insights, the more beneficial it will be for them.

- **Handling multi-dimensional and multi-variety data:** Machine Learning algorithms are capable of handling multidimensional and multi-type data, and they can do so in contexts that are dynamic or uncertain.

- **Widespread Applications:** As a healthcare professional, you may put machine learning to use for your patients. In order for the program to function successfully, it may also needs dependable and trustworthy resources. Analytics and machines are dependable and predictable. Users are able to acquire accurate and consistent insight by connecting analytics tools to a wide variety of structured, semi-structured, and unstructured data sources. Where it is applicable, it has the potential to help in delivering a far more customized experience to customers while also focusing the appropriate clients.

- **ML's industrial applications:** The industrial sector has been a revolutionary change by machine learning. Machine Learning is being utilized to create driverless cars, estimate emergency room wait times, and provide intelligent movie suggestion on services like Netflix. ML-based systems are the solution to employment tasks including stock market forecasting, medical diagnostics, employee selection, and energy demand predicting future. Some of the most popular cloud-based machine learning platforms based on the primary features provided by the largest suppliers. Ml can help doctors in making highly accurate predictions on the treatment of various diseases.

Healthcare x medical diagnosis:

- Classifying DNA sequences
- Disease identification and risk evaluation
- Structural health monitoring
- Healthcare provider sentiment analysis
- Proactive health management
- Predict avoidable emergency room wait times
- Predict preventable strokes and seizures
- Prevent wasteful hospital readmissions

Travel and hospitality:

- Social media-consumer feedback and interaction analysis
- Customer complaint resolution
- Dynamic pricing

Traffic patterns and congestion management

5.2 Challenges of Machine learning Algorithms

Machine Learning has several flaws despite its effectiveness and widespread use. Similar to any technology, there are risks connected with the use of machine learning. For instance, hackers can trick ML models into classifying harmful training samples as genuine or vice versa, leading to deceptive results that humans may not be able to identify as unusable.

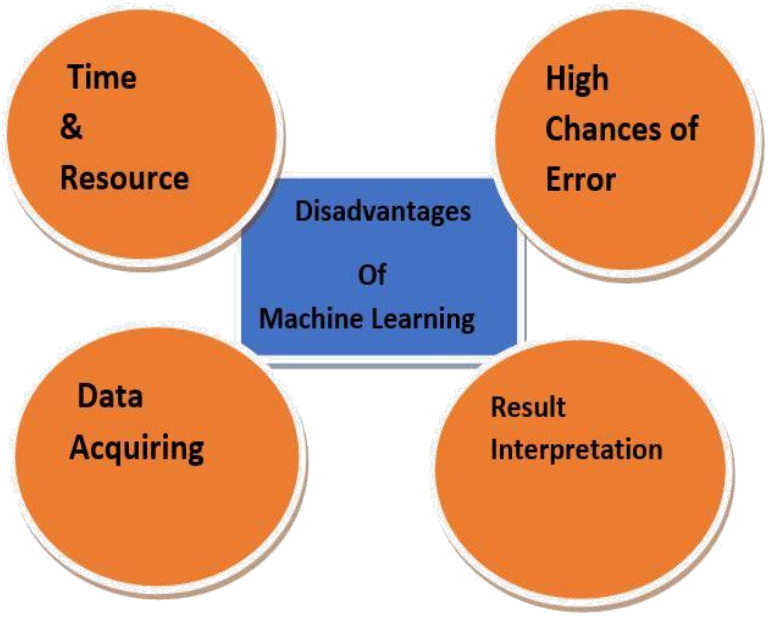

Figure 5.2: Disadvantages of Machine Learning

The following factors and challenges serve to limit:

Privacy protection: Concerning the explanation of machine learning algorithms, privacy protection is a crucial concern. Trust is a crucial issue when dealing with private and possibly sensitive data, particularly when the algorithms difficult or impossible to perceive. This can be a big risk for acceptance, not just among the end users, such as hospital patients, or in safety-critical decision making in general, but also among the experienced engineers who are required to train the models or, in the case of a skilled method, engage daily with the expert system. The challenge of fingerprinting/watermarking information is strongly related to the problem of security and privacy, yet it is different in nature. The introduction of so-called fingerprints or watermarks is a typical reactive solution to this problem.

Data Acquisition: Machine Learning involves training on huge amounts of data set which must be inclusive/impartial and of high quality. Often, they must also wait for newer data to be generated. The outcome or accuracy of experimental results is not always similar. On sometimes, the ML produce different result or variety of accuracy with same the data.

Time and Resources: ML requires sufficient time to enable the algorithms to learn and develop properly to achieve their function with a high level of accuracy and prediction. In addition, it requires huge resources to work. And it also takes a sufficient amount of time to calculate and compile all processes. This may usually require greater processor power on your computer.

High error-susceptibility: Machine Learning is independent yet extremely error prone. Consider training an algorithm with data sets that are too short to be exhaustive. The results of that set will give you a biased training set that are inaccurate predictions. At the end, this results in customers seeing advertisements that are inappropriate. In the case of ML, such inaccuracies can initiate a sequence of errors that can be undetected for longer durations. And when they are detected, it takes considerable time to identify the cause of the problem and even longer to fix it.

5.3 Advantages using Deep Learning

Deep learning is an artificial neural network-based machine learning concept. The goal of deep learning is to imitate how the human brain deals with information. The output of deep learning algorithms is accurate because they map inputs to previously learnt data. Deep learning is a technique that uses the recent advance technology in computing power in parallel with novel neural network architectures to discover hidden patterns in massive datasets. Deep learning provides various advantages:

Deep Learning is capable of automating diagnosis without human involvement. In medical applications, DL algorithms are utilized for tissue lesion detection and characterization. Deep learning algorithms are utilized for the diagnosis of diabetic retinopathy, for the analysis of disease progression, and for the early and accurate detection of patient health issues by medical specialists. Prescriptions and patient health information can be checked by deep learning models to detect and correct potential diagnostic and prescription errors.

- **Cost effectiveness:** While training deep learning models can be costly, but when once trained they can assist organizations reduce wasteful spending. In sectors such as manufacturing, consultancy, and even retail, the cost of an incorrect prediction model or defective product is high. It frequently outweighs the training expenses of deep learning models.
- **Scalability:** Deep learning is highly scalable due to its capacity to analyze vast quantities of data and execute a large number of calculations in a cost- and time-efficient manner. The capacity of deep learning algorithms to eventually learn from their own errors. It is able to validate the accuracy of its predictions and outputs and make any necessary improvements. This has direct effects on productivity (faster deployments or rollouts) as well as modularity and portability (trained models can be used across a range of problems). It increases productivity by automatically adjusting the number of active nodes

based on request quantity. For example, Google Cloud's AI platform prediction allows you to operate a large-scale deep neural network in the cloud.

Deep learning certainly has advantages:

- good at pattern recognition problems
- data-driven, and performance is high in many problems
- end-to-end training: little or no domain knowledge is needed in system construction
- learn of representations: cross-modal processing is possible
- gradient-based learning: learning algorithm is simple
- mainly supervised learning methods

5.4 Challenges of Deep learning Algorithm

There are challenges of deep learning that are more common:

- Data-hungry and thus not suitable for small data sets - Deep learning demands a huge amount of data because it trains the model itself. A deep learning algorithm serves two purposes. It must first acquire domain knowledge before attempting to solve the problem. To get knowledge about a certain area, the algorithm must optimize a huge number of parameters.

- The model is usually a black box and challenging to understand - The black box nature of DL models such as random forest and SVMs makes it practically challenging to predict how they will perform in a certain environment.

- The computational cost of learning is high - Deep learning requires equipment with a high configuration, in order to do such complicated computational analyses. Due to this reason, building and maintaining a deep learning model is expensive.
  .
- Unsupervised learning methods must be developed**:** In deep learning, it will be essential to train the model about deep structures using supervised, semi-supervised, and unsupervised methods. If this

component of the structure is built, then deep learning can easily be performed in billions of hours of processing time on highly parallel computer architecture.

- Ineffectiveness at inference and decision making**:** Having a conversation with multiple people at once is a time-consuming and complicated task. It includes language understanding, language production, dialogue management, access to a knowledge source, and inference.

- Inability to directly control symbols**:** Language data is fundamentally different because it is symbol data when compared to the vector data (real-valued vectors) typically used by deep learning. Currently, symbol data in a language are converted to vector data before being placed into neural networks; the output of neural networks is then converted back to symbol data. A significant amount of information required for natural language processing is actually represented by symbols, including linguistic knowledge, lexical knowledge, and world knowledge. Currently, deep learning techniques do not effectively utilize the knowledge. Symbol representations are straightforward to understand and manipulate, whereas vector representations are resistant to confusing and noise. In natural language processing, the challenge of how to integrate symbol data and vector data and take advantage of the benefits of both data formats remains unanswered.

- Challenging to manage long tail phenomena**:** Natural language data always follow a power law distribution. There are also particular problems to natural language processing, such as the difficulty of dealing with lengthy tails. As a result, for instance, the size of the vocabulary grows as the amount of data grows. This means that whatever the amount of training data available, there will always be circumstances that the training data cannot cover. How to tackle the problem of the long tail is a critical barrier for deep learning.

- Still lacks a theoretical foundation**:** Deep learning lacks a theoretical foundation. In addition to the lack of interpretability of the model and the need for a massive quantity of data and sophisticated computational processing capacity.

In summary, there are still a number of open challenges with regard to deep learning for natural language processing. Deep learning, when combined with other technologies (reinforcement learning, inference, knowledge), may further push the frontier of the field.

5.5 Comparison between Machine learning and Deep learning

Algorithms from the fields of machine learning and deep learning play a significant part in preparing a computer to act as an expert in a variety of domains, including prediction and decision making. Machine learning is the study of providing computers with the capability to learn without being explicitly programmed.

Deep learning is a sort of machine learning that enables systems to acquire knowledge and interpret the world in terms of a hierarchy of concepts. Deep learning methods are essentially an advanced phase of machine learning algorithms that use neural networks to classify data and make more accurate predictions.

In modern medical disciplines, disease diagnostic evaluation is a difficult task. Understanding the appropriate diagnosis of patients through medical examination and evaluation is essential. The healthcare industry generates a vast amount of data regarding medical examination, patient statements, therapy, supplements, etc. Initially, techniques for machine learning were proposed and utilized to identify medical data sets.

Machine learning provides several tools for the analysis of medical data in a structured manner. Second, the algorithms for deep learning serve in categorizing, classifying, and enumerating disease patterns. It also provides the highest degree of precision, allowing the broadening of analytical objectives, and develops therapy prediction models for patients.

For in our thesis, we employed both machine learning and deep learning. We used five prominent algorithms in machine learning. Among the algorithms we have got quite good accuracy from Decision Tree at 98.44% and Random Forest at 98.25%. From the deep learning algorithm, we applied Recurrent Neural Network, which gave us a satisfactory 97.7% accuracy.

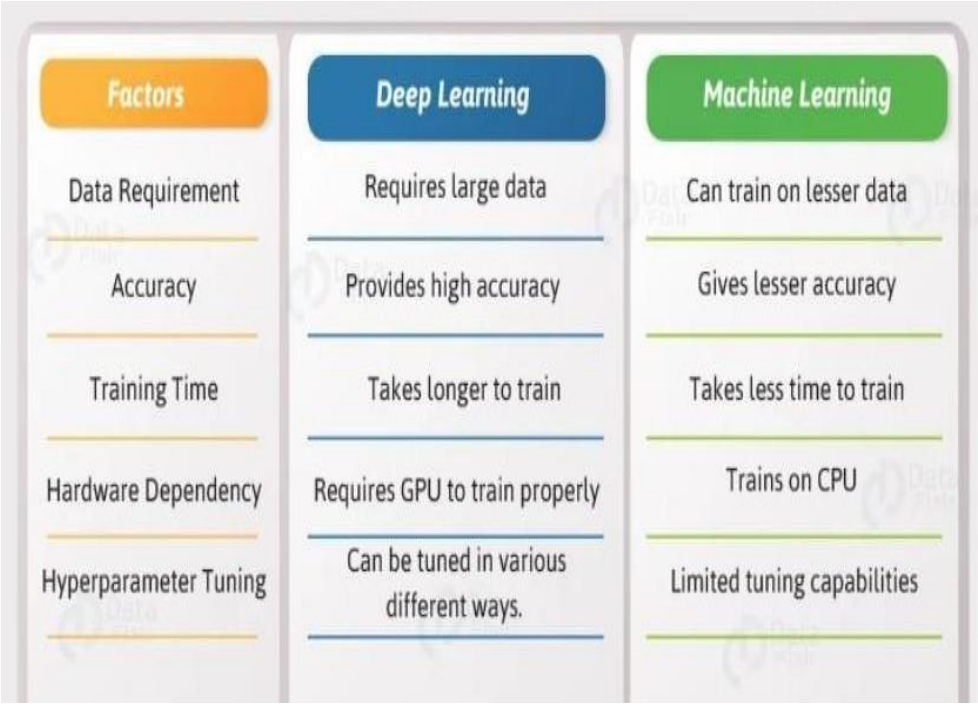

Figure 5.3: Comparison of ML and DL

- Machine Learning (ML) provides computers with the ability to learn without being explicitly programmed and offers a variety of data-learning and prediction methodologies. On the other hand, deep learning (DL) has become advantageous over the past few years, and there have been a large number of advancements that have made it a suitable choice on a large enough scale, such as the increased amount of available data, new storage technologies that can store large amounts of data, and the increased computing power that can process this vast amount of information data.

- Deep learning employs numerous layers and does not necessitate human intervention in feature engineering. Human and manual feature

engineering is impossible due to the big amount of unstructured data and the lack of understanding among security analysts and engineers regarding which features are necessary to detect risks due to the vast number of probable attacker behaviors.

Every Deep neural network includes three types of layers:

1. The Input Layer**:** the layers are represented by the input layer, which accepts input data.

2. The Hidden Layer **:** which is binary in the preceding illustration. It's important to keep in mind that a neural network may contain multiple number of hidden layers.

3. The Output Layer **:** These neural networks are utilized to predict output and classify data.

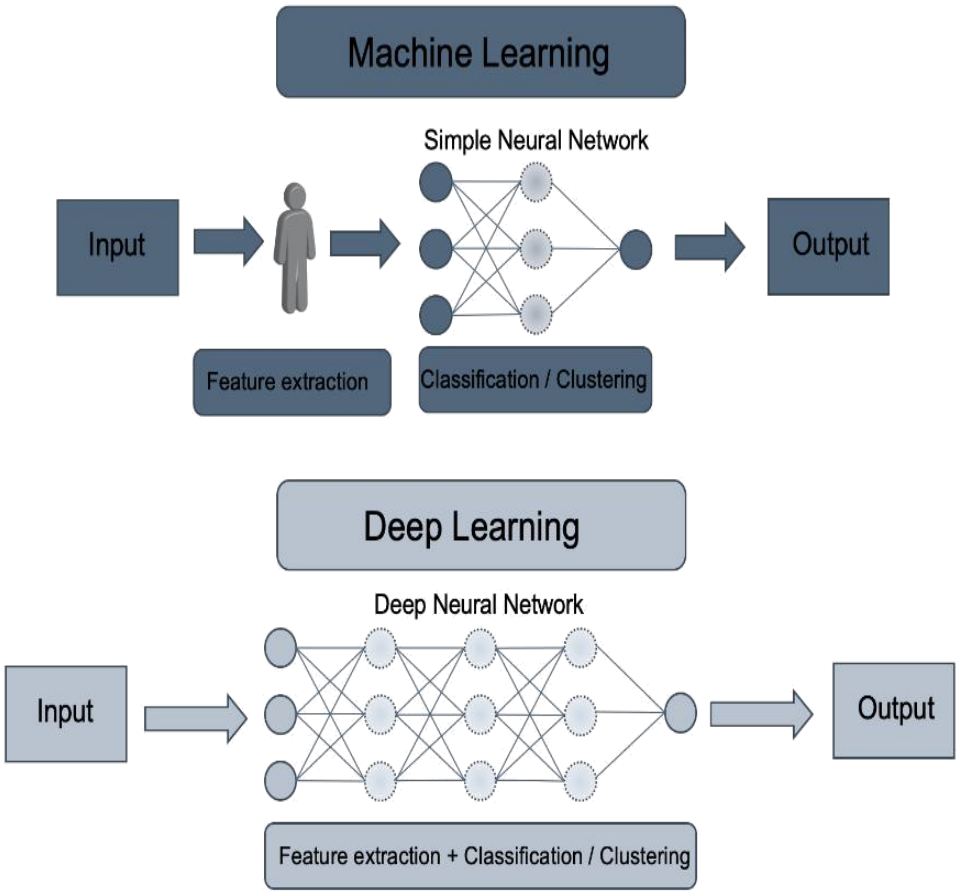

Figure 5.4: Different layer and work process ML and DL

The figure above demonstrates that DL consists of one input, one output, and multiple fully connected hidden layers in between. Neurons represent each layer of the network, which extracts successively more high-level feature of the data until the output is determined by the final layer. The more network layers there are, the more advanced features will be taught. The different layer enables the analysis of datasets in a different manner. Each layer converts incoming data into abstractions.

In other words, it provides many perspectives on various data and averages the results. These features are combined in the output layer to generate predictions. In the machine learning portion, there is one input and one output. In machine learning, humans gather features from a database, and then a simple neural network analyses the database and generates output.

In a direct comparison between DL and ML, DL demonstrates tremendous benefits in terms of performance and accuracy, as well as its capability to handle a far greater volume of data.

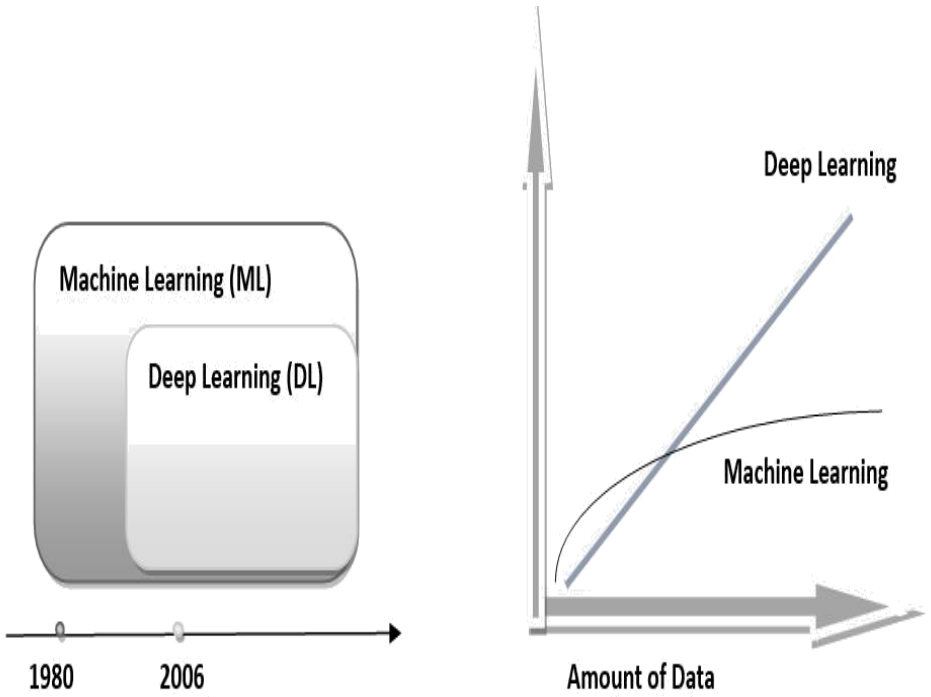

Figure 5.5: Performance of Machine learning and Deep learning

While it is true that deep learning algorithms have improved performance and are the current popular method for prediction across a wide range of industries, including medical, healthcare but it is also important to note that machine learning still has a high performance and accuracy rate. In our research, machine learning provided more precise, accurate results that highlighted an essential rule. In this study, both machine learning and deep learning produced high levels of

Chapter 6 Methodology

6.1 Methodology

This section of the thesis, entitled Methodology, discusses the approaches employed to justify or evaluate a required data set or script. This approach analyzes and predicts thyroid disease by gathering a set of data, processing that data, and implementing various machine learning and deep learning algorithms to predict outcomes in the data set. Models of efficiency acquire a set of data, analyze it, and then apply it. In this research thesis, we will examine an XYZ dataset utilizing some of the most used machine learning techniques and some well-established deep learning methodologies.

6.2 Architecture of the research work

The whole procedure of thyroid detection is shown below using a flow diagram:

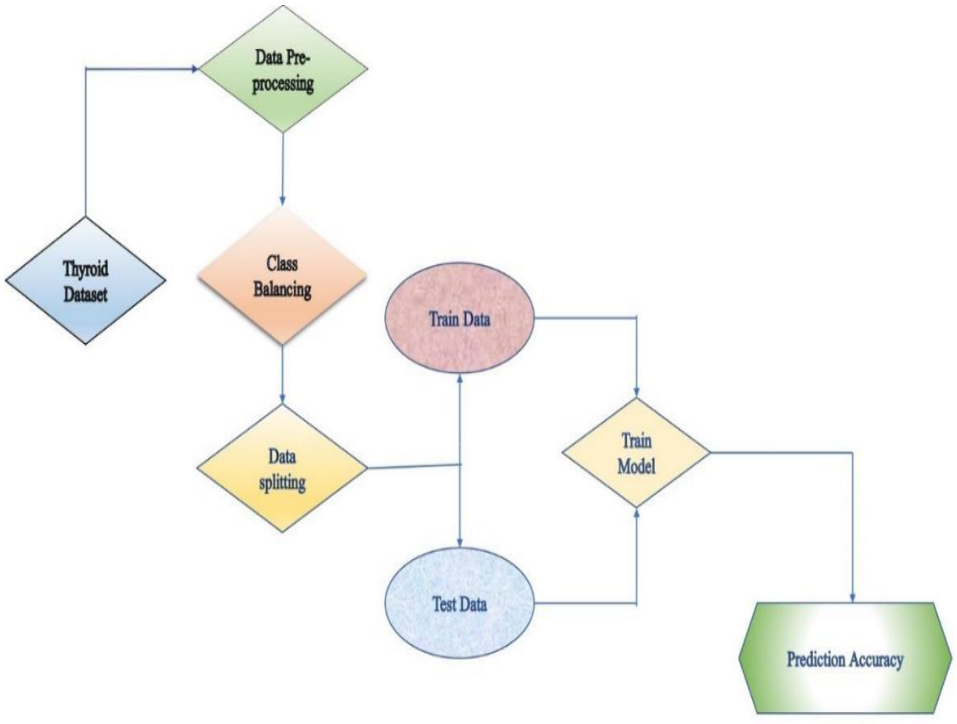

Figure 6.1: Architecture of Thyroid disease prediction

This architecture illustrates the methodical effort that has gone into completing this research. After the dataset has been preprocessed, it was observed. The issue of class imbalance is thus resolved prior to separating the dataset into train data and test data. The training dataset is delivered to classifiers to train the models. Finally, the performance of the trained models is tested through prediction using test instances.

6.2.1 Thyroid Dataset Selection

The thyroid illness data sets have been acquired via Kaggle [26]. Before we could begin diagnosing thyroid disease, we had to locate a relevant dataset for our research purposes. While searching for an appropriate set, we came across Kaggle. It seems pertinent given that it is a recent dataset on which few people have worked. The dataset contains 3,772 observation sets with 30 attributes. The target class is a binary type (sick, negative). The collection includes thyroid illness records compiled and supplied by the Garavan Institute and J. Ross Quinlan, Institute of New South Wales, Sydney, Australia, in 1987.

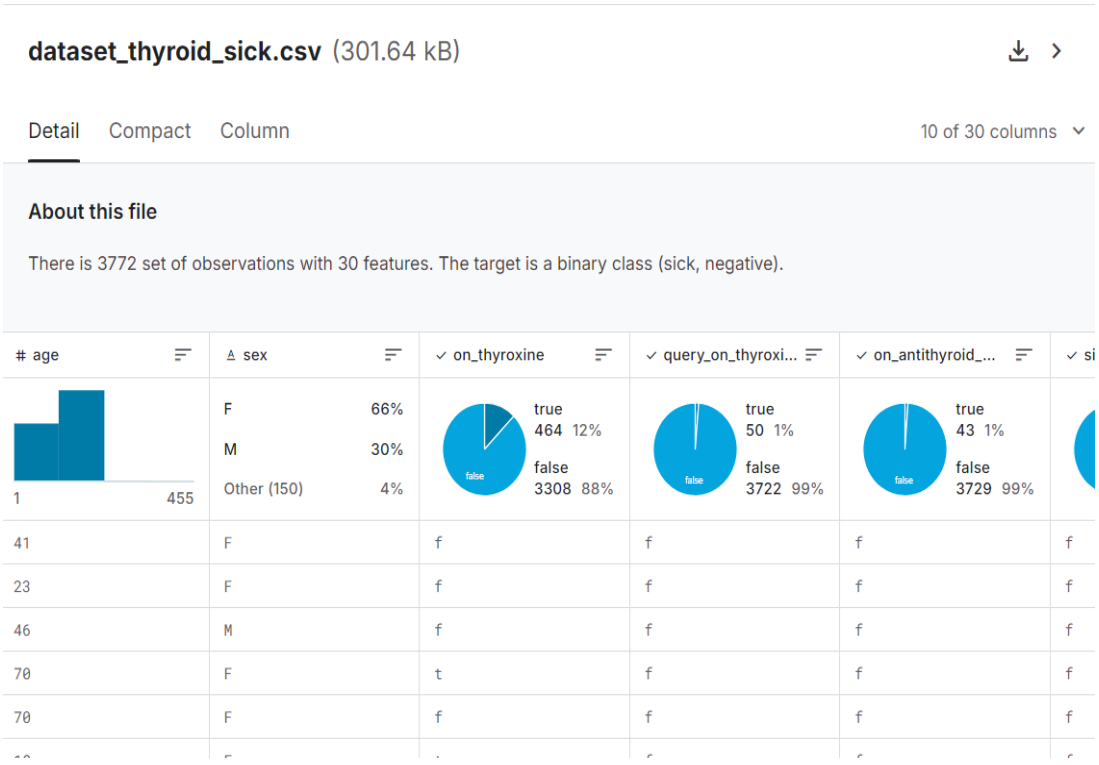

Figure 6.2: Dataset Sample

The dataset was created using the profile concept [27]. In this dataset, there are very few duplicate data records also the dataset is in CSV format, which stand for "Comma-Separated Values," are a file format that allows us to save tabular data, such as spreadsheets. It is useful for large datasets and can be used in programs., ready to use without further processing.

6.2.2 Used attributes in the thyroid dataset

During the period of the research, it was determined that Table No. 1 includes the medical features used to identify thyroid disease [28] through research:

Table 6.1: Attributes used to predict thyroid diseases

Attributes	Description	Data Type
Age	age of the patient	(int)
Sex	sex patient identifies	(str)
on_thyroxine	whether patient is on thyroxine	(bool)
query on thyroxine	whether patient is on thyroxine	(bool)
on antithyroid meds	whether patient is on antithyroid meds	(bool)
Sick	whether patient is sick	(bool)
Pregnant	whether patient is pregnant	(bool)
thyroid_surgery	whether patient has undergone thyroid surgery	(bool)
I131_treatment	whether patient is undergoing I131 treatment	(bool)
query_hypothyroid	whether patient believes they have hypothyroid	(bool)
query_hyperthyroid	whether patient believes they have hyperthyroid	(bool)
Lithium	whether patient * lithium	(bool)
Goiter	whether patient has goiter	(bool)
Tumor	whether patient has tumor	(bool)
Hypopituitary	whether patient * hyperpituitary gland	(float)
Psych	whether patient * psych	(bool)
TSH_measured	whether TSH was measured in the blood	(bool)
TSH	TSH level in blood from lab work	(float)
T3_measured	whether T3 was measured in the blood	(bool)

T3	T3 level in blood from lab work	(float)
TT4_measured	whether TT4 was measured in the blood	(bool)
TT4	TT4 level in blood from lab work	(float)
T4U_measured	whether T4U was measured in the blood	(bool)
T4U	T4U level in blood from lab work	(float)
FTI_measured	whether FTI was measured in the blood	(bool)
FTI	FTI level in blood from lab work	(float)
TBG_measured	whether TBG was measured in the blood	(bool)
TBG	TBG level in blood from lab work	(float)
referral_source		(str)
Class	Whether the patient is sick or negative	(str)

Using several machine learning and deep learning approaches, we have used the dataset to predict the accuracy percentage of thyroid disease. Logistic Regression, Decision Tree, Support Vector Machine (Linear Kernel), Support Vector Machine (RBF Kernel), Random Forest, and Recurrent neural networks (RNN) are the techniques used.

6.2.3 Data Preprocessing

Data preprocessing is the process of preparing raw data for use with a machine learning model. When developing a machine learning project, we do not always come across clean and formatted data. Real-world data typically contains noise, missing values, and may be in an unusable format that cannot be used directly for machine learning models. Data preprocessing is a necessary task for cleaning the data and preparing it for a machine learning model, which improves the accuracy and efficiency of the machine learning model [29]. Considering standardized numeric ranges for input for the algorithms seems effective.

The dataset underwent a comprehensive preprocessing routine to handle missing values, encode categorical variables, and transform the data into a suitable format for machine learning models. Initially, unnecessary columns were removed from the dataset, followed by filling missing values. The ‘age’ column had its missing values imputed using the mode, while continuous variables related to hormone levels (such as TSH, T4U, T3, and FTI) were filled using their mean values.

Categorical variables such as ‘sex’ were encoded into binary values (0 for female, 1 for male). Additionally, any remaining missing values in the dataset were dropped to ensure no incomplete rows existed. The target variable, which classified patients as either 'sick' or 'healthy,' was also encoded into binary values (0 for 'negative' and 1 for 'sick').

Furthermore, one-hot encoding was applied to categorical columns with multiple categories, such as ‘referral_source’, creating a separate binary column for each category. This ensured that all data was in a numerical format suitable for machine learning models.

Finally, the preprocessed dataset was split into features (`X`) and the target variable (`y`), which represented the class of the thyroid condition (sick or healthy), in preparation for model training.

In our dataset, we have included data that necessitates alteration and limitation. Since inconsistent data can produce in fatal errors, approaches are used to pre-process them. In this section, we will classify our data into attributes and labels, and then split the attributes and labels into training and test sets. This enables us to train our algorithm on one dataset and then test it on a different dataset that it has never previously encountered. This provides a more accurate interpretation of trained algorithm's actual performance.

6.2.4 Class Balancing

It is extremely important to train the model on a dataset with nearly the same number of samples when using a machine learning algorithm. This is referred to as class balancing. To train a model, we must have balanced classes; however, if the classes

are not balanced, we must use a class balancing technique before using a machine learning algorithm. Classification is a supervised learning concept in machine learning that divides a set of data into classes. The biggest concern with imbalanced dataset prediction is how accurately we predict both the majority and minority classes. It is critical to accurately identify minority groups. As a result, the model should not be biased toward detecting only the majority class but should also give equal weight or importance to the minority class.

To oversample the minority class in this thesis, we used the Synthetic Minority Oversampling Technique (SMOTE). Adding duplicate minority class records frequently does not add any new information to the model. SMOTE creates new instances based on existing data. To put it simply, SMOTE examines minority class instances and uses k nearest neighbors to select a random nearest neighbor [5].

6.2.5 Data Splitting

Data splitting is called the process when data is divided into two or more subgroups. A two-part split often involves testing or evaluating the data in one part and training the model in the other. Data splitting is a crucial component in data science, especially when building models from data. This is an important step in data preprocessing because it allows us to improve the performance of our machine learning model. If we train our machine learning model with one dataset and then test it with a completely different dataset, our model will struggle to understand the correlations between the models. If we train our model very well, and its training accuracy is also very high, but when we give it a new dataset, its performance suffers. As a result, we always strive to create a machine learning model that performs well with both the training and test datasets. Here, the data is split up into two datasets.

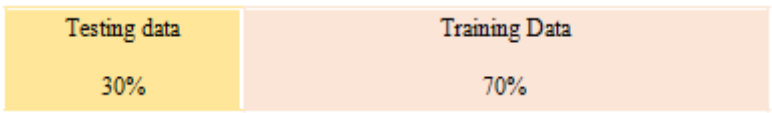

Figure 6.3: Data Splitting Percentage

Splitting Dataset into Train and Test Set:

The Data is divided into 2 sets of data:

- Training Data: The dataset upon which the model would be trained on. Contains 70% data. It is a subgroup of data to train the machine learning model, and we already know the output.
- Test Data: The dataset upon which the model would be tested against. Contains 20% data. It is a subgroup of data to test the machine learning model, and by using the test set, model predicts the output.

In this context, we used the train_test_split method from the sklearn.model_selection library to randomly divide the dataset into training and testing subsets. The independent variables, or features, were assigned to the variable X, while the target variable, labelled as "Class" in our dataset, was assigned to y.

- **X_train**: This variable holds 70% of the feature data, which will be used to train the machine learning model.
- **X_test**: This variable contains 30% of the feature data, which will be used to test the model's performance.
- **y_train**: This represents 70% of the target variable data that will be used during the training phase.
- **y_test**: This represents 30% of the target variable data that will be used during the testing phase.

The train_size parameter was set to 0.7, which specifies that 70% of the data will be used for training and 30% for testing. This split ratio ensures that the model has sufficient data to learn from during training while still having a significant portion of the data reserved for testing. By doing this, the model's ability to generalize to unseen data can be accurately assessed.

Additionally, the random_state parameter was set to 42 to ensure that the random split of the dataset is reproducible. By using the same seed (42), every time the code is run, the dataset will be split in the exact same way, allowing consistent comparisons across different runs or model

6.2.6 Train Model

A dataset used to train an algorithm is known as a training model. To compare the processed output to the sample output, the training model is utilized to run the input data through the algorithm. The first procedure in machine learning is model training, which produces an operational model that can subsequently be validated, tested, and implemented. The way a model performs during training ultimately determines how well it will perform when it is implemented in an application for end users.

The model training phase's two most important factors are the quality of the training data and the algorithm selection. Logistic Regression, Decision Tree, Support Vector Machine (Linear Kernel), Support Vector Machine (RBF Kernel), Random Forest and Recurrent neural network (RNN) are the models that we have implemented here. Specifically, to find the training and testing accuracy, we trained them.

6.3 Confusion Matrix:

A confusion matrix is a table that summarizes the number of true and false predictions made by a classifier. It is used to measure the performance of a classification model. It can be used to evaluate the performance of a classification model through the calculation of performance measures such as accuracy, precision, recall, and F1 score. The confusion matrix is a graph showing the effectiveness of a classification method. It is used to display and summarize the results of a classification algorithm.

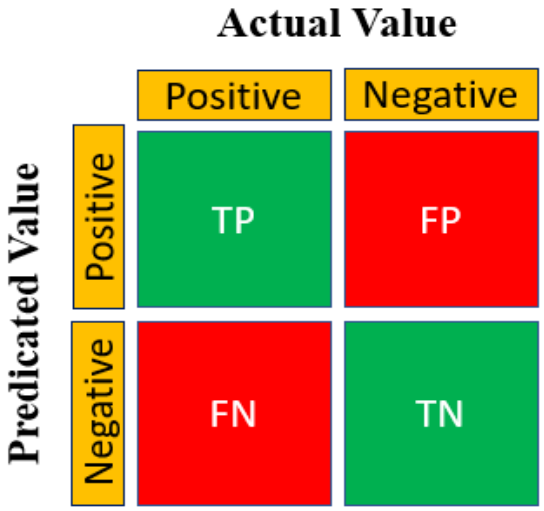

Figure 6.4: Confusion Matrix

The performance of a classification model is measured by an N x N matrix, where N is the number of the target classes. The matrix compares the actual target values with the predictions of the machine learning model. The confusion matrix is used to display important predictive data including recall, specificity, precision, and accuracy [32]. Confusion matrices are useful because they allow you to directly compare values such as true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The precision, recall, F-1 score, and accuracy are defined as below:

Precision: With resulting in positive observations, precision is the proportion of accurately predicted observations to all predicted positive observations. It determines how many of the classes we predicted would be positive are in actually positive are in actually positive. As much precision as possible should be used.

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP + FN}} \textbf{*100}$$

equation (4)

Recall: Recall is the proportion of accurately predicted positive observations to all of the actual positive observations.

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP + FP}} \textbf{*100}$$

equation (5)

F1-Score: Comparing two models that have a high recall, but a low precision is troublesome. So, we apply F1- Score to compare them. F1-score helps in evaluating both recall and precision simultaneously.

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

equation (6)

Accuracy: The percentage of accurate predictions is called accuracy. From all the classes (positive and negative), how many of them we have predicted correctly.

$$\text{Accuracy} = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100$$

equation (7)

Chapter 7 Result and Analysis

7.1 Result and Analysis

After processing the dataset, we train our dataset and use different models to get the output. The results of these applied algorithms will be described further. We have also used Confusion Matrix to visualize the performance of applied algorithms.

7.1.1 Result of Machine Learning Algorithms

Our tests include five different machine learning algorithms such as Logistic Regression (LR), Decision Tree, Support Vector Machine (Linear Kernel), Support Vector Machine (RBF Kernel), and Random Forest (RF) algorithm. We calculated the tests accuracy and the training accuracy of all these algorithms. Then we also calculated Precision, recall and f1 score. This is the result of machine learning algorithms shown in the table:

Algorithm	Logistic Regression	Decision Tree	Support Vector Machine (Linear Kernel)	Support Vector Machine (RBF Kernel)	Random Forest
Test Accuracy (%)	96.69%	98.44%	96.78%	97.15%	98.25%
Training Accuracy (%)	91.19%	99.73%	91.61%	99.33%	100%
Precision (%)	77.97%	88.24%	80.36%	82.76%	87.88%
Recall (%)	66.67%	86.96%	65.22%	69%	84.06%
F1-Score (%)	71.88%	87.59%	72%	75.59%	85.93%

Table 7.1: Result of Machine Learning Algorithms

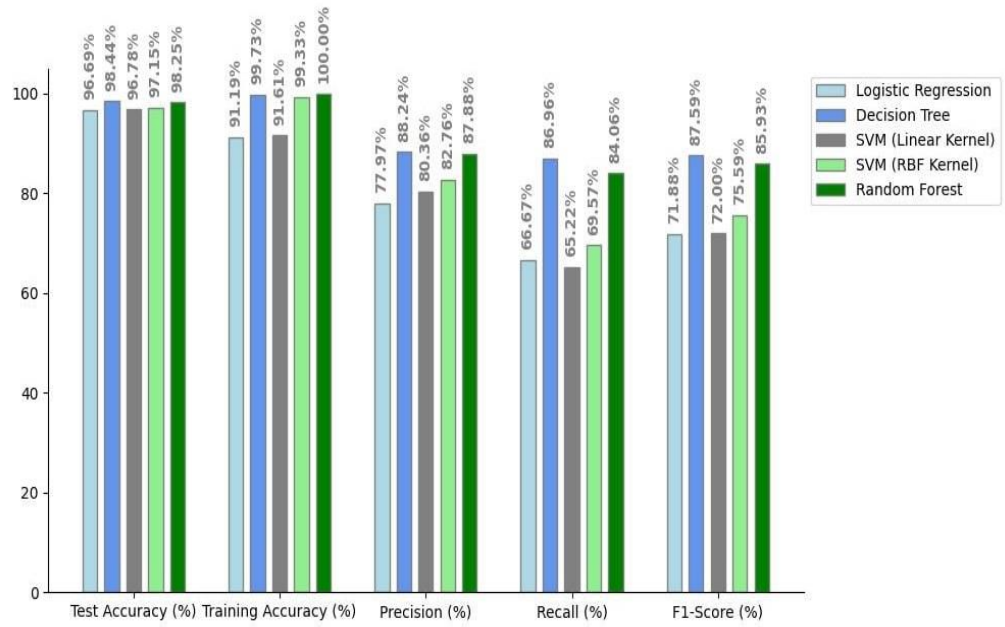

Figure 7.1: Plotting of Machine Learning Algorithms Results

The Decision tree model provides high performance for test data based on the experiment; however Random Forest provides the highest accuracy during training accuracy. Decision tree and Random Forest model provides almost equal precision results, whereas the Decision tree model has superior recall and F-1 scores.

By analyzing the precision, recall, and f1-score for each label, it is possible to figure out how well the model has learned to categorize each label. In comparison to the other two, the test and train accuracy of the Support Vector Machine (Linear Kernel), Support Vector Machine (RBF Kernel), and Logistic Regression are inadequate.

So, to get better accuracy result for training data, we can see that the Random Forest algorithm has the maximum accuracy level that is 100%. On the other hand, for testing accuracy, Decision tree model gives the highest accuracy of 98.44%. Overall decision tree gives the highest performance on all the classifier precision, recall, and f1 score. Even though the Decision tree classifier the data set has a score of 87.59% for the f1 score and training accuracy of 99.73%, which is still high compared to the other following models. It's interesting to compare the performance of the decision tree and Random Forest to the other algorithms. Here we get good result on both

decision tree and Random Forest classifier because is trained through bagging which improves the accuracy of machine learning algorithms, and it provides an effective way of handling missing data. But between the overall performance in this system, the decision tree algorithm represents the ideal model for our data. On the other hand, the Logistic Regression model, and Support Vector Machine (Linear Kernel) model gives the lowest training and test accuracy. So, we can see clearly that the decision tree classifier did the best in our model test. Decision tree also gives quite good result for the test accuracy. The overall performance in our system, the Decision tree algorithm represents the ideal model for our dataset.

7.1.2 Confusion Matrix applied on ML

Confusion matrix is widely used because they give a better idea of a model's performance than the classification accuracy. For our dataset we have used five machine learning algorithms. So, for every algorithm we have plotted the confusion matrix. The confusion matrix for all the algorithms is shown in below:

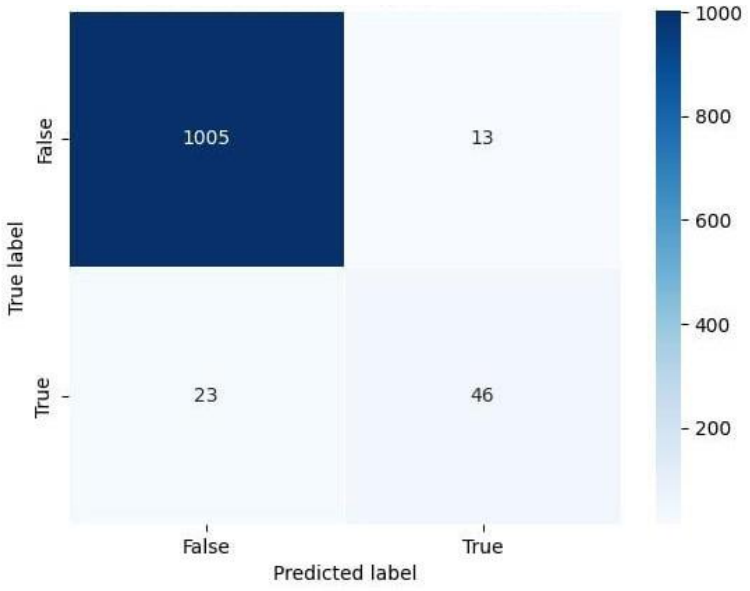

Figure 7.2: Confusion matrix on Logistic Regression

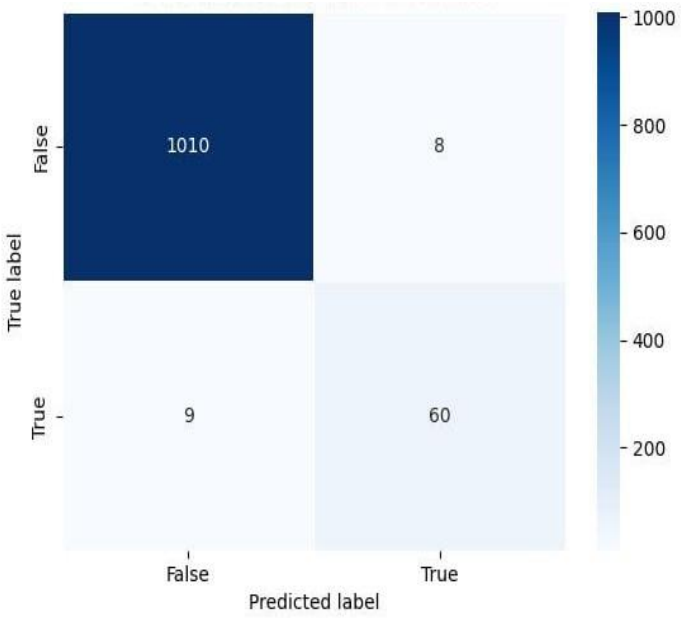

Figure 7.3: Confusion matrix on Decision tree

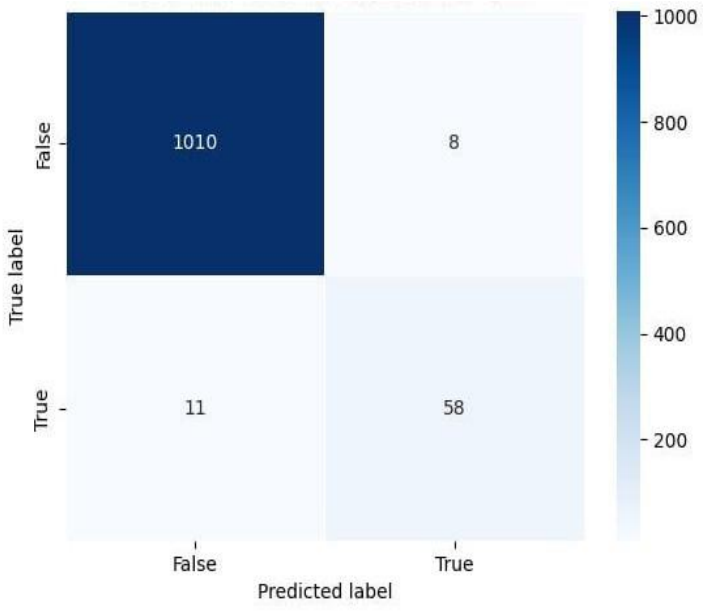

Figure 7.4: Confusion matrix on Random Forest

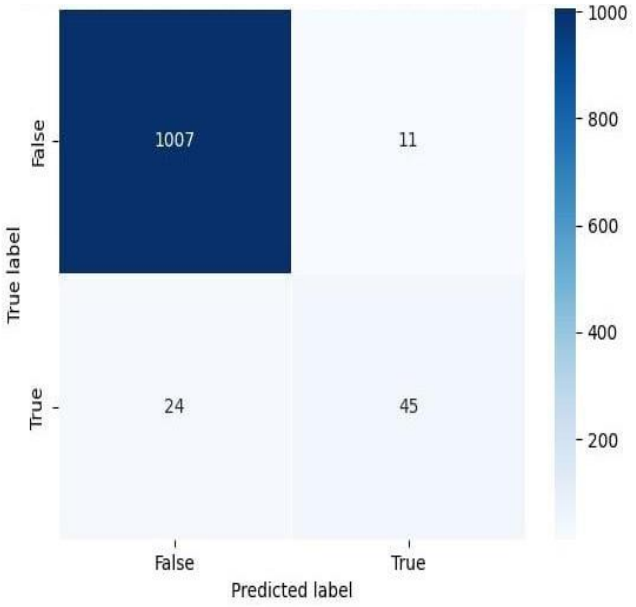

Figure 7.5: Confusion matrix on SVM(Linear Kernel)

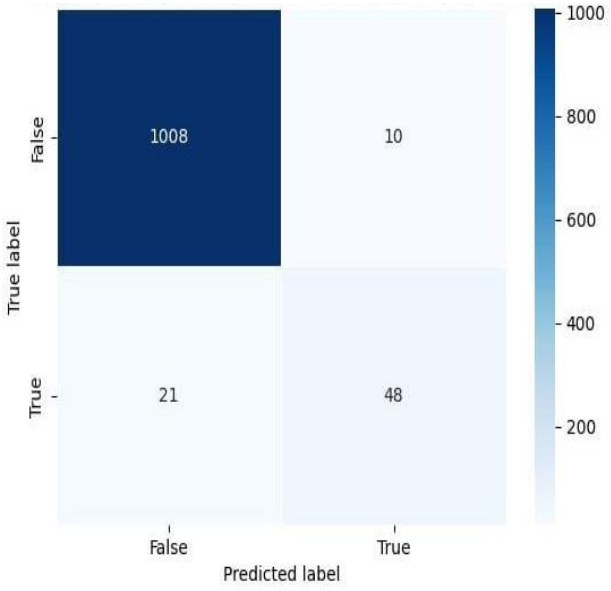

Figure 7.6: Confusion matrix on SVM(RBF Kernel)

7.1.3 Deep Learning algorithm
Using RNN - Long Short-Term Memory (LSTM)

The parameters used in deep learning is below-

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 256)               7424

dropout (Dropout)            (None, 256)               0

dense_1 (Dense)              (None, 128)               32896

dropout_1 (Dropout)          (None, 128)               0

dense_2 (Dense)              (None, 63)                8127

dropout_2 (Dropout)          (None, 63)                0

dense_3 (Dense)              (None, 1)                 64

=================================================================
Total params: 48,511
Trainable params: 48,511
Non-trainable params: 0
```

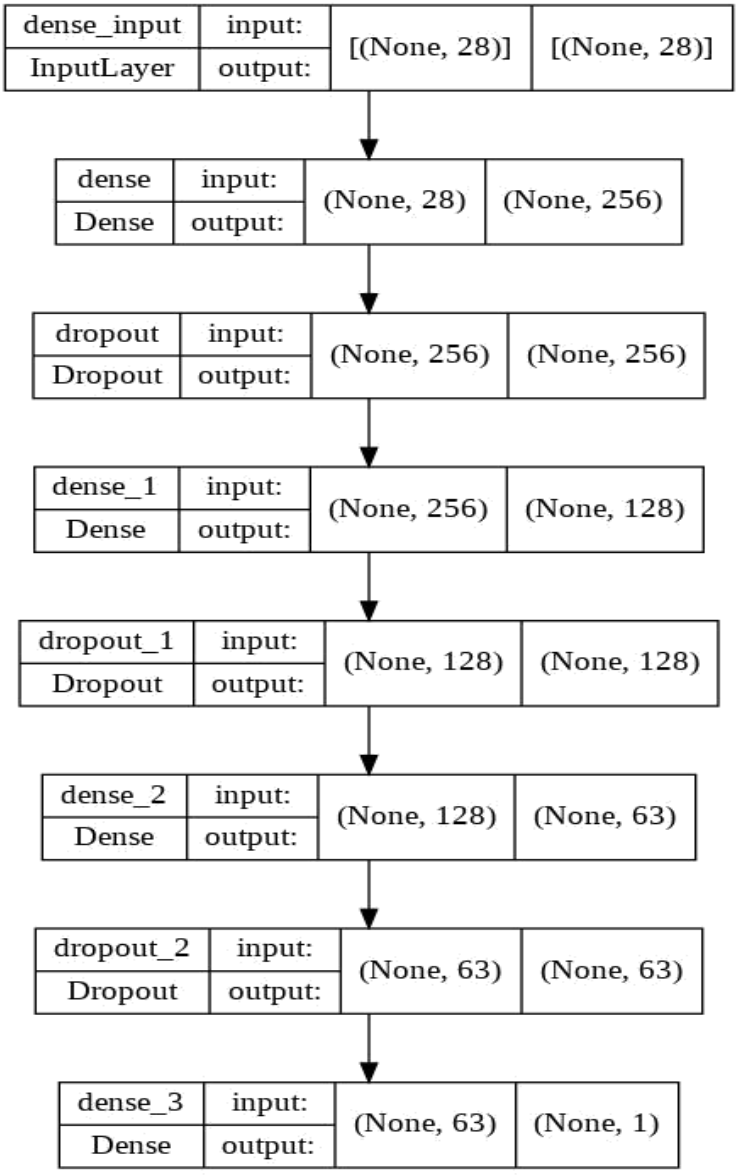

Figure 7.7: Parameters of Deep learning

To achieve good results in our dataset, we also used RNN. Our greatest accuracy in the test set utilizing RNN-Long Short-Term Memory (LSTM) is 97.70%, according to the data. However, it can be quite helpful to be able to see a summary of the model thus far, including the current output form, when building a Sequential model sequentially. Run the model for 40 epochs after fitting it.

7.1.4 Evaluation of Confusion Matrix

We have got accuracy, MSE and RMSE by applying five machine learning algorithms. Table 7.1 is represented graphically in Figure 7.1. However, the Figure shows the accuracy of the results in the Decision Tree is found to be the best accuracy level. Whereas the lowest accuracy percentage is found in the Support Vector Machine (Linear Kernel) and Support Vector Machine (RBF Kernel) algorithm which is quite similar to the accuracy percentage we got in the classification models.

Table 7.2: Result of Confusion Matrix accuracy

Algorithm	**Confusion Matrix Accuracy**
Logistic Regression	0.96688
Decision Tree	0.98436
Support Vector Machine (Linear Kernel)	0.96780
Support Vector Machine (RBF Kernel)	0.97148
Random Forest	0.98252

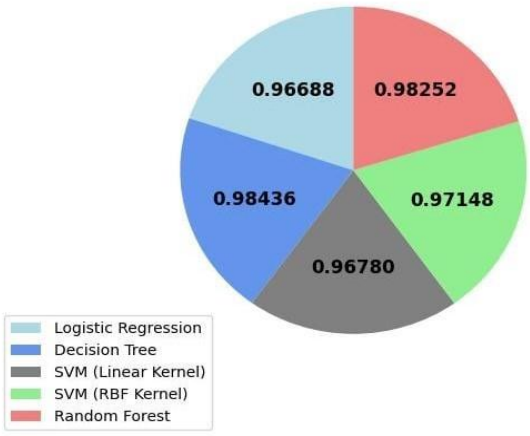

Figure 7.8: Plotting of Confusion Matrix accuracy Results

For MSE, there is no ideal value. Simply expressed, a lower value is preferable, and 0 indicates that the model is ideal. Here Decision tree has the nearest to zero MSE value making it better compared to other Machine learning algorithms.

Table 7.3: Result of Mean Squared Error (MSE)

Algorithm	**Mean Squared Error (MSE)**
Logistic Regression	0.03312
Decision Tree	0.01564
Support Vector Machine (Linear Kernel)	0.03220
Support Vector Machine (RBF Kernel)	0.02852
Random Forest	0.01748

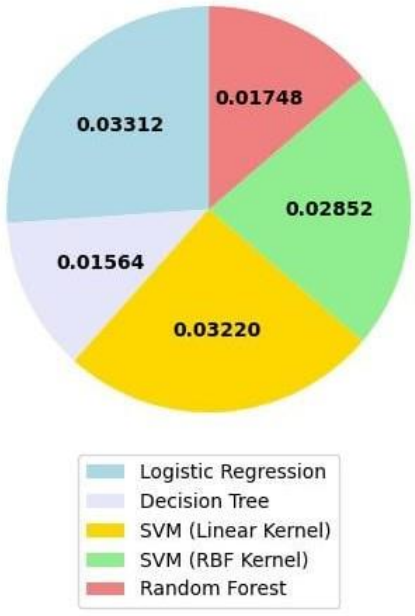

Figure 7.9: Plotting of Mean Squared Error (MSE)

RMSE (Root mean squared error) is the standard deviation of prediction errors. A given mode can "fit" a dataset more accurately the lower the RMSE. When the RMSE is low, the simulated and observed data are more accurate since they are close to one another. Therefore, model performance is improved by a decreased RMSE. Here also the decision tree has lower error.

Table 7.4: Result of Root Mean Squared Error (RMSE)

Algorithm	**Root Mean Squared Error**
Logistic Regression	0.18199
Decision Tree	0.12506
Support Vector Machine (Linear Kernel)	0.17944
Support Vector Machine (RBF Kernel)	0.16888
Random Forest	0.13221

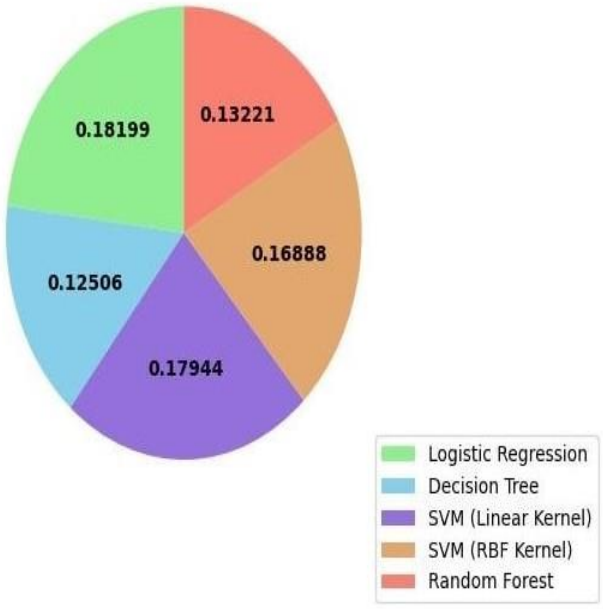

Figure 7.10: Plotting of Mean Squared Error (RMSE)

3.1.6 Comparison of Machine Learning and Deep Learning

In this section of the research, we will compare the predicted accuracy values of all the algorithms of machine learning and deep learning.

Table 7.5: Comparison of Machine Learning and Deep Learning

Algorithm	**Accuracy**
Logistic Regression	96.69%
Decision Tree	98.44%
Support Vector Machine (Linear Kernel)	96.78%
Support Vector Machine (RBF Kernel)	97.15%
Random Forest	98.25%
Recurrent neural network (RNN)	97.7%

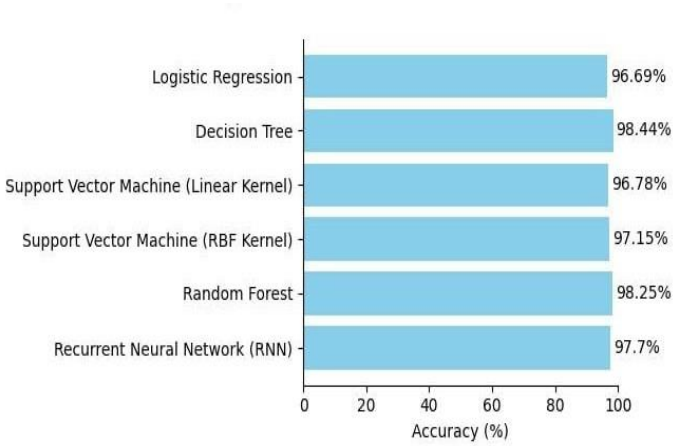

Figure 7.11: Plotting of Comparison of Machine Learning and Deep Learning

Five machine learning algorithms and one deep learning algorithm are seen in the plot above. RNN is the deep learning algorithm, and the machine learning algorithms are Logistic Regression, Decision Tree, Support Vector Machine (Linear Kernel), Support Vector Machine (RBF Kernel), and Random Forest. From all the methods, it is clear that Decision Tree, a machine learning technique, provided the best accuracy; While the deep learning model, Recurrent Neural Networks (RNN), also performed well, with a competitive accuracy of 97.7%, it did not surpass the Decision Tree in overall effectiveness. Conversely, Logistic Regression, other machine learning approach, yielded the lowest accuracy.

These results demonstrate that, while deep learning methods can provide robust outcomes, certain machine learning techniques, such as Decision Tree, can still outperform in specific contexts like thyroid disease prediction. This highlights the importance of selecting the most appropriate algorithm based on the nature of the data and the problem being addressed.

Chapter 8 Conclusion and Implications

8.1 Research Challenges

A few libraries were in more advanced forms when we were building various machine learning models, making it impossible to build models using those libraries. Additionally, this problem was not specifically addressed after the error, thus we had a number of problems to illustrate the error's cause. The construct showcase also becomes slower than it was previously, thus the show runs for about 7-8 minutes when we add more information to adapt the dataset from awkwardness. On the other hand, RNN preparation has its own unique set of issues. We discovered that getting them ready is difficult in our implementation. A basic show does require some time and resources to prepare. But that's a reasonable restriction of the equipment.

8.2 Limitations

There are some limitations in our research. This research will not cover all the detection techniques. We use several algorithms in our detection techniques, but we will not try all the algorithm for detection techniques. No model will give us the 100% accuracy.

However, in our research we will try to focus on the prediction of disease using different approaches as disease prediction structure is based on the patient's symptoms.

8.3 Future Scope

In the future, it is necessary to further expand the set of data and attributes considered to better generalize our findings. With more data, the training process can produce more effective classifiers that allow for more reliable estimates of exhibit performance. Finally, another aspect that can be investigated concerns the presence of any secondary thyroid disease associated with the patient, to understand if there is any specific additional thyroid disease affects thyroids. In fact, it often happens that patients suffer from multiple thyroid diseases at the same time

8.4 Conclusion

Medical data is vital for treatments alike vaccination development and drags design. The dataset is collected in the medical application because of testing the response of the patient to a particular drug or by collecting the medical tests for diagnosing a particular disease. This project is a connection between medical science and computer science.

Our main purpose is to help the medical science as well as the patients. This project has only one outcome and that is to know if a patient has thyroid or not. Thyroid disease is often difficult diagnose because it can easily confuse symptoms with other symptoms of the disease dysfunction. In our thesis, we analyzed the methodologies of different researchers for identifying various thyroid illnesses using machine learning and deep learning models. We applied machine learning algorithms, such the Logistic Regression, Decision Tree, Support Vector Machine (linear kernel), Support Vector Machine (RBF Kernel), Random Forest.. Our program is developed to justify these results properly. We have used machine learning algorithms to predict the disease based on those following symptoms. We have a datasheet and we have used the algorithms based on those.

This project is developed by python. Talking about the specification of thyroid we are predicting thyroid and thyroid by using our following program.

References

[1] Obermeyer Z, Emanuel EJ. Predicting the future— big data, machine learning, and clinical medicine. N Engl J Med. 2016; 375:12161219.

[2] Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature. 2015;521: 452-459.

[3] G. Chaubey, D. Bisen, S. Arjaria, and V. Yadav, “Thyroid Disease Prediction Using Machine Learning Approaches”, Natl. Acad. Sci. Lett., (2020).

[4] E. Dogantekin, A. Dogantekin, and D. Avci, “An automatic diagnosis system based on thyroid gland: ADSTG”, Expert Syst. Appl., 37:9, 6368-6372, (2010).

[5] Menger, V., Scheepers, F., & Spruit, M. (2018). Comparing deep learning and classical machine learning approaches, for predicting inpatient violence incidents from clinical text. Applied Sciences, 8, 981.

[6] Prasad, V., Srinivasa Rao, T., & Surendra Prasad Babu, M. (2016). Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms. Soft Computing, 20(3), 1179–1189.

[7] Raghuraman, M. T., Sailatha, E., Gunasekaran, S. (2019). Efficient thyroid disease prediction and comparative study using machine learning algorithms. International Journal of Information and Computing Science. 6(6), 617–624.

[8] Dharmarajan, K., Balasree, K., Arunachalam, A. S., & Abirmai, K. (2020). Thyroid disease classification using decision tree and SVM. Indian Journal of Public Health Research & Development, 11(03), 229–234.

[9] Yadav, D. C., & Pal, S. (2020). Prediction of thyroid disease using decision tree ensemble method. Human-Intelligent Systems Integration, 2, 89–95.

[10] Ghali, U. M., Usman, A. G., Degm, M. A. A., Alsharksi, A. N., Naibi, A. M.,&Abba, S. I. (2020). Applications of artificial intelligence-based models and

multi-linear regression for the prediction of thyroid stimulating hormone level in the human body. Int J Adv Sci Technol, 29(4), 3690-3699.

[11] Ioniţă, I., & Ioniţă, L. (2016). Prediction of thyroid disease using data mining techniques. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 7(3), 115-124.

[12] K. Geetha and C. S. S. Baboo, "An Empirical Model for Thyroid Disease Classification using Evolutionary Multivariate Bayseian Prediction Method", Glob. J. Comput. Sci. Technol. E Network, Web Secur., 16:1, 242-250, (2016).

[13] S. Shroff, S. Pise, P. Chalekar, and S. S. Panicker, "Thyroid disease diagnosis: A survey", Proc. IEEE 9th Int. Conf. Intell. Syst. Control. ISCO 2015, (2015).

[14] Children and adolescent patients with goiter and normal thyroid function: US findings related to underlying autoimmune thyroid diseases.Hwang SM, Hwang JY, Moon JH, Yang I, Woo JY, Lee HJ.

[15] Medicine (Baltimore). 2022 Sep 2;101(35):e30095.

[16] S. Razia and M. R. Narasinga Rao, "Machine Learning Techniques for Thyroid Disease Diagnosis - A Review", Indian J. Sci. Technol., 9:28, 1-9, (2016).

[17] A. K. Aswathi and A. Antony, "An Intelligent System for Thyroid Disease Classification and Diagnosis", Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT, 1261-1264, (2018).

[18] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms", 3rd Int. Conf. Comput. Sustain. Glob. Dev., INDIA, 1310–1315, (2016).

[19] A. K. Aswathi and A. Antony, "An Intelligent System for Thyroid Disease Classification and Diagnosis", Proc. Int. Conf. Inven. Commun. Comput. Technol.

[20] ICICCT, 1261-1264, (2018).

[21] Machine Learning & Security, Clarence Chio & David Freeman

[22] Kedar Potdar, Rishab Kinnerkar, “A Comparative Study of Machine Algorithms applied to Predictive Breast Cancer Data”, International Journal of Science & Research, Vol. 5, Issue 9, pp. 1550-1553, September 2016.

[23] S. B. Kotsiantis, I. Zaharakis, P. Pintelas et al., "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, Vols. 160, no. 1, p. 3–24, 2007.

[24] Y. Al Amrani, M. Lazaar, and K. E. El Kadiri, "Random forest and support vector machine based hybrid approach to sentiment analysis," *Procedia Computer Science,* vol. 127, p. 511–520, 2018.

[25] https://towardsdatascience.com/what-is-deep-learning-and-how-does-it-work2ce44bb692ac/

[26] https://www.kaggle.com/code/yashmehta648/thyroid-detection/notebook/

[27] https://www.kaggle.com/datasets/bidemiayinde/thyroid-sickness-determination/

[28] Interactive-Thyroid-Disease-Prediction-System-Using.pdf

[29] https://www.javatpoint.com/data-preprocessing-machine-learning/

[30] https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced- feed-forward-neural-network.png/

[31] https://blog.quantinsti.com/machine-learning-classification-strategy-python/

[32] https://www.labelf.ai/blog/what-is-accuracy-precision-recall-and-f1-score/

[33] Shukla, A. & Kaur, P. (2009). Diagnosis of thyroid disorders using artificial neural networks, IEEE International Advance computing Conference (IACC 2009)– Patiala, India, pp 1016-1020.

[34] S. Wang, C. Tang, J. Sun and Y. Zhang, "Thyroid disease detect connected neural network", *Frontiers Neurosci.*, vol. 13, May 2019.

[35] L. Ozilmaz, T. Yildirim, "Diagnosis of thyroid disease using artificial neural network method," ICONIP'029th international conference on neural Information, pp. 2033-2036, 2002.

[36] Ripley B. neural networks. Cambridge: Cambridge University Press; 1996.

[37] Khushboo Taneja, Parveen Sehgal, Prerana “Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network” International Journal of Research in Management, Science & Technology (E-ISSN: 2321- 3264) Vol. 3, No. 2, April 2016

[38]https://builtin.com/sites/www.builtin.com/files/styles/ckeditor_optimize/public/inli ne-images/national/feed-forward-neural-network.png/

[39] https://data-flair.training/blogs/deep-learning-vs-machine-learning/

[40]https://builtin.com/sites/www.builtin.com/files/styles/ckeditor_optimize/public/inline-images/national/Feed-Forward-Neural-Networks.png/

[41] https://towardsdatascience.com/training-deep-neural-networks-9fdb1964b964/