Application of Machine Learning for the Prediction and Management of Non-Communicable Diseases

Sepehr Jokanian
Artificial Intelligence Student
Asia Pacific University of Technology & Innovation
Kuala Lumpur, Malaysia
tp070463@mail.apu.edu.my

Abstract— **Non-communicable diseases, such as cancer, heart diseases, chronic lung disease, and diabetes, are some of the most challenging problems in modern society, causing the mortality of a significant number of people worldwide. Early detection of these diseases can reduce mortality rates, improve treatment processes, and decrease the risk of complications. On the other hand, considering the growth of technology and its effects, the usage of technology is inevitable in human life, and most people can benefit from it. One of the most effective aspects of technology is related to artificial intelligence, which has become widespread around the world. This paper focuses on the integration of machine learning, as a branch of artificial intelligence, in the detection of NCDs (non-communicable diseases) and how it can improve the health condition of society and control the mortality rate due to these diseases. The new system makes use of datasets that include records, imaging information, and patient backgrounds linked to non-communicable diseases. By using learning techniques like convolutional neural networks (CNNs) for image assessment and recurrent neural networks (RNNs) for analyzing sequential data, the system can detect early signs of diseases by recognizing patterns and markers. This holistic method aims to enhance the treatment of several patients, with these illnesses leading to overall public health results.**

***Keywords*—Non-communicable diseases (NCDs), Deep learning algorithms, Early detection, Artificial Intelligence (AI), Machine learning algorithm, convolutional neural network (CNN), SVM, Decision tree**

I. INTRODUCTION

In modern society, one of the most widespread and controversial topics relates to non-communicable diseases and their treatments for control. These diseases include high blood pressure, cancer, high cholesterol, diabetes, and so on. Statistics show that the number of cases for these diseases is growing worldwide, particularly in densely populated countries with stressful routines. For instance, in 2008 alone, 36 million global deaths were related to NCDs, and the number of annual deaths due to these diseases is increasing (Wagner & Brath, 2012). Several factors play a crucial role in causing these diseases, such as physical inactivity, obesity, diet, and genetics. Studies show that one of the best methods to control these diseases and significantly affect personal health is to detect these diseases in their early stages or use platforms for diagnosis (Budreviciute et al., 2020).

Non-communicable diseases, or chronic diseases, are not contagious and include a wide range of conditions such as cancer, heart problems, diabetes, and chronic respiratory diseases. In recent centuries, they account for 70% of global mortality, and this percentage is increasing. Delaying the detection of these diseases results in more expensive treatment (Samiratu Ntohsi, 2022).

To reduce this delay time, it is essential to understand why people do not refer to doctors for early detection. One of the most effective reasons is the high cost of clinics. Additionally, poor relationships between doctors and patients can deter individuals from seeking medical advice, leading them to self-medicate instead. Traditional methods of predicting diseases used statistical approaches, which were challenging and ineffective due to the vast number of influencing features and criteria. However, with significant technological advancements, it has become useful to predict NCDs. (Wellington Kanyongo & Ezugwu, 2023).

Machine learning is a subfield of artificial intelligence (AI) that consists of algorithms and structured data. It uses a human brain model known as an artificial neural network. (Brownlee, 2019). According to the massive progress of technology and artificial intelligence, their applications have increased in all aspects of people ‘s lives. In recent decades, artificial intelligence has been Widely used in many aspects of human life, and medicine is one of the best examples of its benefits. Profound uses of AI in healthcare are inevitable. AI can be applied in many aspects, such as assisting robots, detecting diseases, surgery, drug discovery, experimentation, and so on. The majority of doctors and surgeons use AI in their field due to benefits that it has and makes life easier not only for patients, but also for the medical industry. The studies show the role of AI in healthcare, such as the application of AI-assisted clinical procedures that are capable of handling large amounts of data and producing data with a high accuracy rate. It also eliminates the wasting of time for the data and can monitor data of patients with less time and more accurate results. (Mohammed Yousef Shaheen, 2021)

Considering the importance of early detection for NCDs to control them and the role of advanced AI in healthcare, as well as its applications in various aspects of medical treatment. This research focuses on application of machine learning as a branch of AI in early detection of NCDs. Machine learning, or in a more complicated method, deep learning, needs epochs or, in simple terms, datasets, to determine more data afterwards. The relationship between the number of datasets and the accuracy of machine learning algorithms is direct. This means that if more datasets are added to the system, the accuracy of the systems in result will be increased. According to the early detection system the correct interpretation of diagnostic results is used. These datasets help the system to have more accurate detection in the future. The statistics show that machine learning is one of the most effective approaches to detection. (Benning et al., 2022)

II. Problem Statement

As discussed earlier, however, non-communicable diseases have been widespread and have affected populations in both developing and developed countries. Still, early detection remains one of the most critical challenges. One of the most effective ways to address the prevalence of these diseases is through their early detection. However, current methods are not efficient and applicable due to limitations that affect their diagnosis. Machine learning approaches are the best ways to overcome these barriers by the power of technology to improve early detection systems. This approach offers several benefits, such as improving patient outcome, cost-effectiveness, and time-saving measures. The influence of this algorithm has been demonstrated in Bangladesh.

The results show that the prediction of the system for patients in these cases was correct 98% of the time, and the data also indicates that the data-driven approach for identifying individuals at risk of NCDs is feasible (Samiratu Ntohsi, 2022).

One of the most crucial challenges in implementing this algorithm is collecting and applying these data from resources, categorizing, and applying them as datasets. It starts with extracting the raw data from UNRWA servers. These data consist of different kinds, such as patient health records, laboratory results, blood tests, X-ray tests, and so on. After gathering raw data, they need to be transformed into usable data. For this step, we require experts who can establish relations between rows and columns. This stage is one of the most important due to a lack of proper expertise; the dataset will be inaccurate, and the entire system will lack functionality. The next barrier is related to measuring the data to achieve a desirable algorithm with minimal errors. For instance, there are some sensitivities of datasets for machine learning, such as overfitting data (too much perfect data) and underfitting data (too few meeting the requirements). Then, the training dataset must be constructed accurately. All in all, implementing machine learning with this large volume of data considering the lack of certain resources, such as funding, stability, scalability of machines and systems, and human resources, power, and so on. (Machine Learning Models to Predict Early Complications of NCD Patients - ProQuest, 2022).

III. Research Aim

The main aim of this research is to implement a high-accuracy deep learning system as a subset of artificial intelligence and apply it to the early detection of non-communicable diseases. It also focuses on enhancing the accuracy of deep learning algorithms.

IV. Research Objectives

A. Implementing a system with high accuracy outputs.
B. Collecting various datasets contain patient information from reliable resources.
C. Collaborating with healthcare experts who can retrieve useful and efficient data.
D. Updating and monitoring the system continuously after implementation to keep it up-to-date and efficient.

V. Research significance

As this research has explored earlier, non-communicable diseases, which consist of diseases such as cancer, diabetes, respiratory issues, and so on, are significant. The statistics show that around 41 million people annually are killed by these diseases, which comprise 71% of global deaths. This number is growing, and more people are engaging with them in all countries (Jean Joel Bigna & Jean Jacques Noubiap, 2019).

The crucial role of technology in all aspects of human life is inevitable in the contemporary century. This paper explores the application of technology, specifically the power of machine learning, in early detection of NCDs and, consequently, their control. Early detection is one of the most effective ways to manage these diseases, and this approach is not possible without the use of machine learning algorithms. Research shows that machine learning algorithms with high accuracy can achieve a success rate of 90% in diagnosing these diseases. This figure indicates that by relying on machine learning, we can detect these diseases and prevent global deaths (Skrede et al., 2020).

The most important aspect is achieving a high accuracy rate. Precise continuous updates and monitoring are necessary to maintain high accuracy. Furthermore, the paper will focus on how the algorithm is implemented in detail.

VI. Literature Review

A. Medical Image Analysis Using Deep Learning Algorithms

Deep learning is known as a branch of artificial intelligence that uses artificial neural networks consisting of multiple layers to diagnose patterns in images and predict image outputs. It has achieved impressive results in various aspects. These results are particularly vital for analyzing medical images in healthcare. Medical image analysis is a major field of study that consists of several steps such as processing, interpretation, and analysis of medical images. Deep learning speeds up image processing and analysis with powerful algorithms. It has the capability to determine and categorize medical images such as X-rays, MRIs, CT scans, and ultrasound images quickly and accurately. The application of deep learning in healthcare has revolutionized the field. There are real-world applications of this technology, such as skin cancer detection, tumor segmentation in MRI, and automated bone fracture detection. Considering that deep learning requires large amount of annotated data, and it is time-consuming to collect and requires deep understanding. Consequently, this is one of the barriers faced by deep learning. On the other hand, some patient’s records are confidential and cannot be used as a dataset (Li et al., 2023).

B. Machine and Deep Learning Approaches in Genome

The study of structure, operation, and transmission of genes is known as genomics. Machine learning has been applied in some research fields that are related to biology and healthcare, such as detection and understanding specific genes, binding and splicing effects on cell procedures, and the interaction of genes and environments. One of the most effective methodologies which applied in distinguishing between coding and noncoding districts is decision trees. By applying some other algorithms, such as logistic regression, random forest, and logistic model trees, we can detect the differentially expressed genes. By implementing a deep learning system, it can handle the RNA splicing, which is known as an approach for machine learning in the system. Another application of deep learning in genomics is its ability to obtain a precisely rational sequence motif. A pattern of sequence that occurs repeatedly in a set of consecutives is known as a motif. One obstacle that the system will face is unknown viruses. The machine learning algorithm must achieve high accuracy to detect viral genomes. (Mostafa et al., 2020).

C. Deep Learning in Human Activity Recognition with Wearable Sensors

In recent decades, application of deep learning in human activity recognition (HAR) via mobiles and wearable devices has increased. This growth has two main reasons: first, deep learning methods can transfer raw data to useful information and learning from them to apply to a specific aim. Second, a deep neural network can be used for most applications if it has big data. Due to this, deep learning has become popular in HAR- based applications. One of the effective approaches for deep learning algorithms is related to convolutional neural networks. It can retrieve features and categorize input data with high accuracy. It can be used when data from HAR is transferred in two dimensions from sensors. The pre-trained models can be started with a large dataset. The approach also has one disadvantage, which is that is requires fixed-sized input data. As a consequence, we must apply other algorithms, such as RNN, which accepts a flexible amount of dataset (Zhang et al., 2022).

D. Plant Disease Detection using AI Models

One of the greatest of impacts of automation and artificial intelligence is on agriculture and farming. Crop yield is a vital aim for every farmer, achievable through maintaining healthy plant and their environments. An effective way to achieve this is by applying artificial intelligence and machine learning to detect plants diseases in their early stages. This implementation improves detection efficiency and empowers farmers to monitor their health and control their environments. Early disease detection allows to prevent measurements or intervention at a treatable stage. Studies have shown that machine learning algorithms, powered by artificial intelligence, can achieve this detection with high accuracy rate and sensitivity. This is done by analysing plants leaves to predict future input data. In this methodology, a convolutional neural network VGG-19 model, has been applied to predict diseases. To train this algorithm, a dataset of 15,915 plants leaf images were used. This dataset includes a combination of healthy and diseased leaves, representing 19 different types of diseases. The data was collected from the Plant Village database for training purposes. The implemented algorithm achieved an accuracy of 65.2% with a testing loss of 0.4418. This high accuracy rate, powered by machine learning, allows algorithms to detect a large scale of plant diseases in the future (Anwar Abdullah Alatawi et al., 2022).

E. Al-based Tool for Early Detection of Alzheimer's Disease

Early detection is one of the most crucial ways to manage Alzheimer's diseases, as there is currently no cure. Medications can only slow the progression of the disease. The study proposes a solution that utilizes the hippocampus and the VGG16 model with transfer learning to detect Alzheimer's disease in its early stages. In this implementation, the hippocampus plays a crucial role in classifying patients into three main groups. Cognitive normal (CN); these individuals do not have cognitive problems. Mild cognitive impairment (MCI); these individuals have mild cognitive problems, and Alzheimer disease (AD); these individuals have obvious cognitive problems. The Alzheimer 's Disease Neuroimaging Initiative (ADNI) a dataset was used to train this model. The system was not only trained by dataset but also enriched by advanced image processing techniques. The allows model to achieve high accuracy statistics, including testing accuracy of 98.17%, validation accuracy of 97.52%, and training accuracy of 99.62%. In this system, efficient methods have been used to improve its ability to learn from different and complex examples, leading to better accuracy and easier integration with new data. The system prioritizes user-friendliness. It empowers radiologists to predict probabilities of classes and visualize patient images in 2D and 3D, which are effective for early detection of Alzheimer's disease. (Shafiq Ul Rehman et al., 2024)

F. A Machine Learning Model for Early Detection of Diabetic Foot Using Thermogram Images

Diabetes mellitus (DM) can cause some complications, including heart disease, stroke, blindness, and diabetic foot ulceration (DFU) with potential lower limb amputation. Treatment for DFU might be challenging, and delays increase the risk of infection and amputation. Therefore, early detection is crucial to prevent foot amputation. Research suggests that temperature monitoring with using thermogram images can effectively predict DFU in up to 97% of patients. This study focuses on the use of machine learning techniques for early detection of diabetic foot using thermogram images taken with Infta-Red((IR) cameras integrated with smartphones. Convolutional Neural Network (CNN) and K-Nearest Neighbour (KNN) algorithms were applied to analyse foot thermogram images and identify diabetic foot conditions. The results revealed that the network of the MobileNetV2CNN model achieved a high accuracy rate of 95% in classifying thermogram images of two feet. On the other hand, the AdaBoost classifier, utilizing only 10 features, surpassed the MobileNetV2CNN performance, reaching an accuracy rate of 97%. (Amith Khandakar et al., 2021)

G. Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning

Parkinson's disease (PD) is a neurovegetative disease that affects the nervous system. Early detection of PD is one of the most effective actions for slowing down its progression and providing effective treatments for Parkinson's patients. In the premotor stage of PD, individuals may experience a decreasing sense of smell, depression, night-time sleep problems, and mood disturbances. This study explores an innovative deep learning technique that introduces an early detection algorithm based on premotor features to diagnose Parkinson's disease. The approach combines the proposed deep learning technique with 12 machine learning models and leverages transfer learning from a small dataset of 183 healthy individuals and 401 PD cases. By combining these datasets and using them for training purposes, the system achieved a high accuracy rate of 96.45% in correctly identifying PD cases in the early stages. This high accuracy is possible due to the deep learning model's ability to learn both linear and not-linear features within PD's dataset. (Wang et al., 2020)

H. Similar Systems

These days, machine learning is widely used to analyze medical reports in the healthcare system. It is used to predict diseases in different stages. In this system, the random forest algorithm was used to forecast the number of diseases relevant to patients. The task works based on different categories of number of diseases: one, two, or more than two. The percentages of accuracy based on the different algorithms were 68.3% using random forest and 69.7% using the XGBoost method. Age, waist-hip, and weight were the most effective on numbers of diseases in random forest. One of the research limitations was related to different conditions of weather and quality of life in rural and city areas. They have a profound effect on the health of people who live in certain areas. For instance, the options that are available for treatment in cities are more than rural areas, or the cleaner air is more prevalent in rural areas compared to crowded cities, which can have a significant effect on people's health. Additionally, some patient records are confidential and are not usable as dataset. (Roy et al., 2023)

TABLE I. Comparisons Systems

Feature	Similar system	My system
Disease focused	Multi disease prediction	Early detection of NCDs
Algorithm	Random forest, XGBoost	Decision tree, SVM,CNN
Data type	Medical reports	X-rays, Laboratory records, test results, health record
Accuracy	68.3%(Random forest) 69.7%(XGBoost)	-
Strengths	Explores disease prediction using machine learning	Focuses on early detection NCDs with deep learning and diverse datasets

VII. Methodology

This research focuses on the quantitative approach through a retrospective cohort study to achieve the research objectives, which consist of early detection of non-communicable diseases with empowerment of machine learning as a branch of artificial intelligence. This design allows us to analyze large datasets of patient records, including laboratory results, medical history, and X-rays, to recognize patterns and risk factors with NCDs. This system uses large datasets, previously utilized by specialists.

These data are obtained through approved access procedures from reliable sources related to healthcare, such as research institutions (existing datasets), corresponding databases including electronic health records (EHRs) of patient and clinical trial databases, and patients records from medical data. By applying existing data from these sources, this research used the archival research method for collecting data. These data include X-rays, laboratory results, and patient medical records. Expertise in healthcare and machine learning is necessary to approach the algorithm of a system, since not all datasets are usable, and they need processing to become useful information from raw data. Consequently, expertise in both aspects of healthcare and machine learning algorithms is required to achieve the best result. On the other hand, to achieve the goal, the system needs computational resources, software related to that, and human resources to collect data from different resources. To implement this methodology, some approaches have been applied, such as using deep learning, which is applied to X-rays, laboratory records, and patient medical records.

Decision trees and SVM are used to classify tasks and features of different records, and CNN is utilized for the purpose of early detection of NCDs. Considering we use large datasets for our methodology, deep learning is one of the best systems for this purpose. It also provides high accuracy and is more efficient for detecting images and patterns by using trained datasets and making validation to prevent overfitting and underfitting. Therefore, by applying decision trees, SVM, and CNN, the detection system can handle different aspects of research problems. The performance of the implemented system is evaluated by statistical methods such as precision, recall, and accuracy. These metrics allow us to measure the effectiveness of deep learning models in identifying NCDs. Accuracy helps to indicate the accuracy of predictions. Recall evaluates the model's capability to identify actual cases of NCDs. Precision indicates the ratio of true positives to all positive predictions of NCDs.

By analyzing this data, the strengths and weaknesses of the implemented system will reveal. This research acknowledged some potential limitations. These limitations included data availability and completeness within specific healthcare sources. To implement this system, we tried to obtain the access to the most available and reliable datasets possible to address possible missing data points. On the other hand, deep learning models can be susceptible to overfitting, underfitting, and bias. To improve these risks, we applied suitable and valid techniques during model training to optimize performance and mitigate potential biases in the dataset. To implement this system, some ethical and safety problems were concerned as well, including the privacy of patients in their data, avoiding any possible bias in the algorithm, and maintaining the confidentiality of patient information. Finally, a crucial aspect of collecting a patient dataset is rating informed consent. For this approach, clear and comprehensive documentation related to informed consent for study purposes was provided. By applying these steps, the system covers ethical considerations. For safety considerations, the system was tested and will be updated and monitored to assure its accuracy and reliability to prevent misdiagnosis and ensure it is safe for patient usage.

VIII. Dataset

A. Direct Users

Doctors and Clinicians: These medical professionals interact directly with the system. The system assists them in making decisions about early detection of non-communicable disease and suggests its use to patients.

Patients: As users, patients can benefit from the system for early detection of non-communicable disease.

B. Indirect Users

Healthcare Researcher: Healthcare researchers can leverage the system and deep learning datasets to identify new risk factors and develop more accurate methods. They provide guidance on how the data can apply effectively in clinical practice as well.

Developers: Developers built the system using datasets. They will also monitor the system to maintain its accuracy.

Medical Educators and Data Collector: Medical educators integrate medical training and data filtering from sources into the system. The data collectors engage with the system by collecting datasets from reliable sources such as databases, clinics, and patient records.

C. Data Collection Methods

Archival data refers to the analyzing historical data, documents generated in the past. The extensive number of methods are applied in the archival method to make it convenient for studying documents. (Mohr & Ventresca, 2002). In the developed system for NCDs algorithm, the datasets were related to patients records. As a consequence, the best method for collecting data was archival data collection, which was applied for the system.

D. Data Analysis and Visualization

Variables	n	Data Type	Mean	Minimum	Maximum	Standard Deviation	50th Percentile
				Gender			
Male	86	Categorical	–	–	–	–	–
Female	60	Categorical	–	–	–	–	–
				Age Group			
<=35	15	Categorical	–	–	–	–	–
35<>65	67	Categorical	–	–	–	–	–
>=65	64	Categorical	–	–	–	–	–
Age	–	Numeric	59	18	95	16.43	60
Weight	–	Numeric	62.2	90	33	12.73	60
Height	–	Numeric	162.26	124.46	175.26	9.49	165.1
Waist Hip Ratio	–	Numeric	0.76	0.47	0.99	0.12	0.76
DM	80	Numeric	10.56	3.4	25.5	4.39	9.90
CKD	42	Numeric	2.54	1.20	5.40	1.01	2.35
Heart Issue	122	Categorical	–	–	–	–	–
IBS	2	Categorical	–	–	–	–	–
RTI	51	Categorical	–	–	–	–	–
Thyroid	12	Categorical	–	–	–	–	–

Fig. 1. The information of patients based on their age, height, gender, waist hip ration, and weight were collected. *(Roy et al., 2023)*

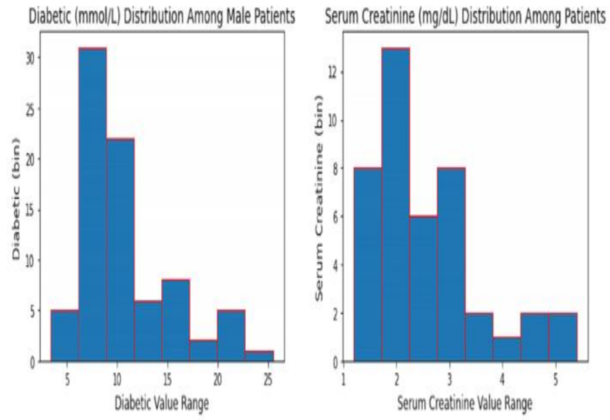

Fig. 2. DM Distribution and Serum Creatinine Distribution Among Patients

54.79% of patients had blood sugar from 61 to 459.45mg/dl and among them the male portion were 52.5% and rest for female. *(Roy et al., 2023)*

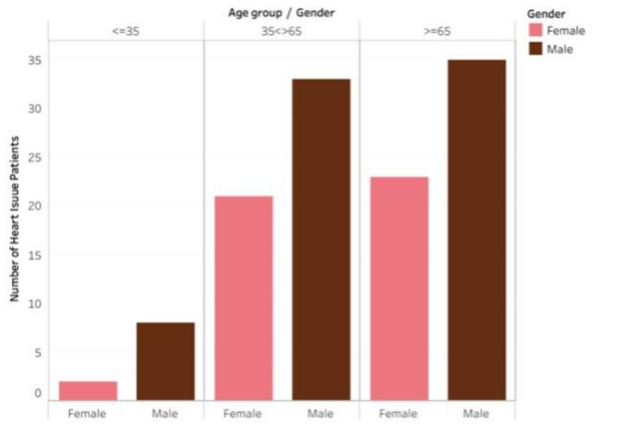

Fig. 3. Age and Gender based DM Distribution. *(Roy et al., 2023)*

The statistics show that the middle age (35 years old or above) people consist 90% of these population and the female portion is lower than male. (Roy et al., 2023)

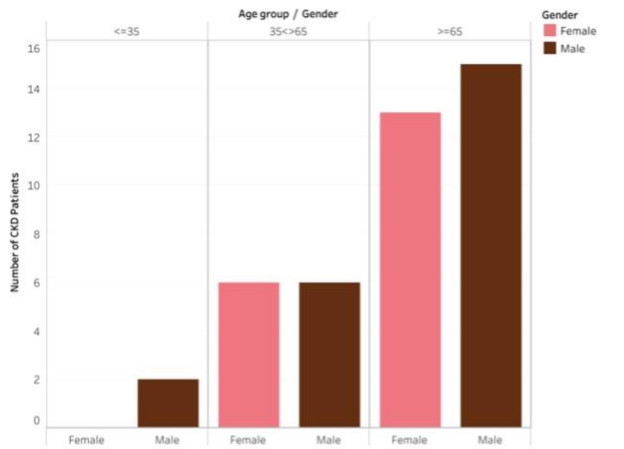

Fig. 4. Age and Gender based CKD Distribution *(Roy et al., 2023)*

Chronic kidney disease (CKD) has a prevalence among men and 2.5% of people are identified with that CKD. (Roy et al., 2023)

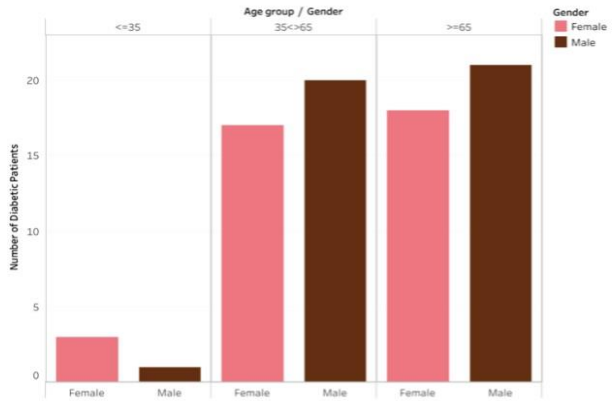

Fig 5. Age and Gender based Diabetes Distribution. *(Roy et al., 2023)*

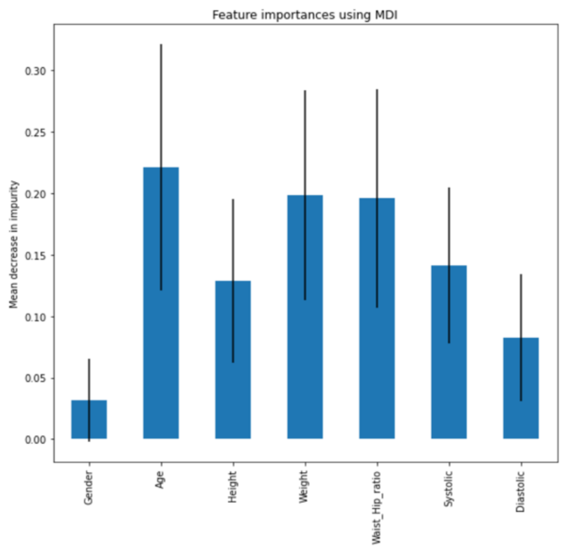

Fig. 6. Random Forest Feature Importance in Number of Disease Prediction. *(Roy et al., 2023)*

IX. Proposed System Overview

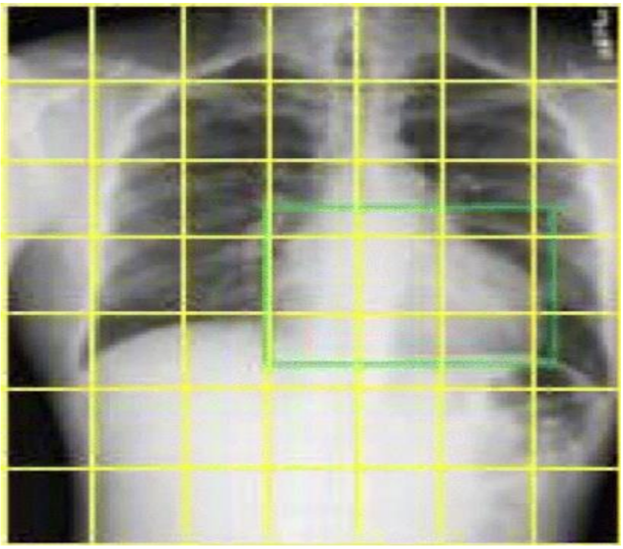

Fig. 7. Xray dataset. *(Al-antari et al., 2020)*

The proposed system is poised to revolutionize the early detection of non-communicable diseases such as cancer, diabetes, heart disease, and blood pressure. The system operates with a high accuracy rate of output diagnosis, leveraging the power of artificial intelligence and utilizing accurate dataset. Furthermore, it undergoes continuous updating and monitoring after implementation. The implementation of the system comprises several steps, which will be explained in detail in the following paragraphs.

Checkup Item (X)	NCD Risk Category(Y)*	Number of subjects of Y in X (%) within Group B (n=15,710)	Number of subjects of Y in X (%) within Group C (n=10,342)	Number of subjects of Y in X (%) within Group D (n=2,110)
BP	Green*	11,711 (74.55%)	8,325 (80.50%)	1,728 (81.90%)
	Yellow, Orange, Red*	3,999 (25.45%)	2,017(19.50%)	382 (18.10%)
BS	Green*	14,395 (91.63%)	9,551(92.35%)	1,988 (94.22%)
	Yellow, Orange, Red*	1,315 (8.37%)	791(7.65%)	122 (5.78%)
BMI	Green*	12,467 (79.36%)	8,499 (82.18%)	1,675 (79.38%)
	Yellow, Orange, Red*	3,243 (20.64%)	1,843(17.83%)	435 (20.62%)
Sex	Male	8,477 (53.96%)	6,257 (60.50%)	1,370 (64.93%)
	Female	7,233 (46.04%)	4,085 (39.50%)	740 (35.07%)
Age	10-19	32 (0.20%)	28 (0.27%)	5 (0.24%)
	20-29	6,827 (43.46%)	5,931 (57.35%)	1,026 (48.63%)
	30-39	3,936 (25.05%)	2,534 (24.5%)	638 (30.24%)
	40-49	2,354 (14.98%)	986 (9.53%)	256 (12.13%)
	50-59	1,334 (8.49%)	451 (4.36%)	106 (5.02%)
	60-69	792 (5.04%)	259 (2.50%)	44 (2.09%)

Fig. 8. Patient records. *(Hu et al., 2018)*

NCD	Relative risk (95% Confidence interval)	Priority population
Cardiovascular disease	1.61 (1.43–1.81)	People living with HIV [68]
Type 2 diabetes	7.33 (4.79-11.51)	Women with gestational diabetes [69]
Cancers – head and neck[2]	2.64 (2.00–3.48)	People who have had Tuberculosis
Cancer – cervical	6.07 (4.40–8.37)	Women living with HIV [70]

Fig. 9. Patient tests. (*NCD Alliance*, 2021)

A. Algorithm development

The algorithm is designed to identify patterns within datasets for early detection of non-communicable diseases. In this step, a training dataset is utilized to apply the algorithm with labelled data. enabling the recognition of relationships between various factors within the datasets and disease results. Techniques such as decision trees, support vector machine, and convolutional neural network (CNNs) will be applied to enhance the accuracy of the algorithm.

B. Decision tree

This is a flowchart-based structure that consists of several nodes, each determining a new path to flow. It is particularly useful for categorizing different datasets. Diseases such as kidney diseases, and diabetes can be detected based on this structure. (Zhuravel, 2023).

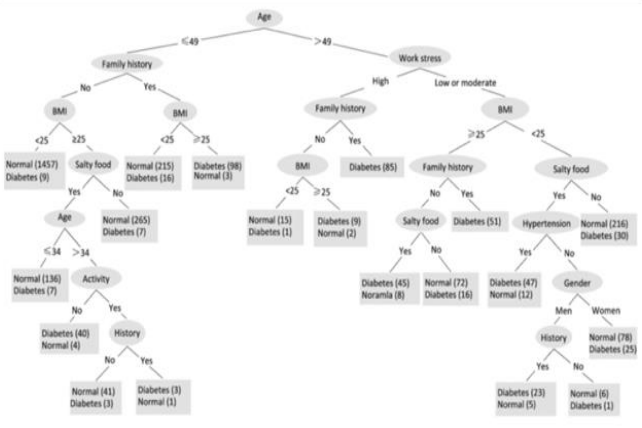

Fig. 10. Decision Tree Structure. *(Pei, D., 2019)*

C. Support vector machine (SVM)

This algorithm is utilized for data classification purposes. It divides the dataset into different parts based on hyperplanes to find the most suitable one. This algorithm is commonly applied to medical image records, where it effectively classifies them into different groups. (Zhuravel, 2023)

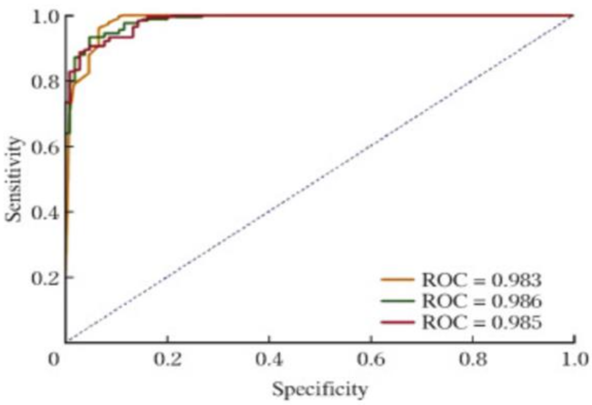

Fig. 11. SVM Algorithm (Receiver Operating Characteristics (ROC)). *(Akinnuwesi et al., 2023)*

D. Convolutional neural network (CNN)

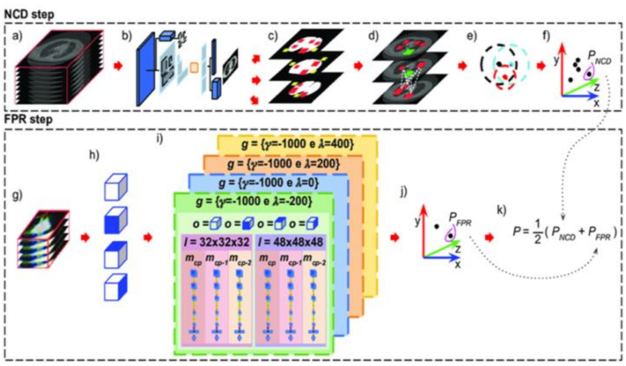

Fig. 12. CNN Algorithm. *(Pereira.F, J.2021)*

It is one of the most efficient algorithms to detect images. It consists of three layers: the convolutional layer, which processes images; the pooling layer, which summarizes and prioritizes images; and fully connected layers, which from the final layer and classify each image into its corresponding group. The algorithm is useful for detecting image patterns and is employed in diagnosing medical image patterns such as X-ray, MRI, and CT scan records. (Zhuravel, 2023)

X. Testing

To achieve the best results with high accuracy output, all the mentioned methods are applied to the system to recognize patterns within large datasets. After implementation, the system undergoes testing using different methods to ensure its high functionality rate and to debug any issues that may arise. Several testing methods are suitable for this system, including unit testing, integration testing, system testing, and security testing.

A. Integration testing

Focuses on the functionality of different components and how they interact. It involves testing different parts of system, such as the dataset, implemented algorithm, and output, to ensure an accurate result, a correct dataset, and adequate implementation.

B. System testing

This test evaluates scalability, stability, and tolerance to ensure its effectiveness in various scenarios.

C. *Security testing*

It is related to assessing the resistance of systems against security threats.

D. *Algorithm monitoring and updating*

After releasing the system and algorithm, it needs to be updated continuously and monitored for keeping a high accuracy rate of the system and the helping system to be optimized.

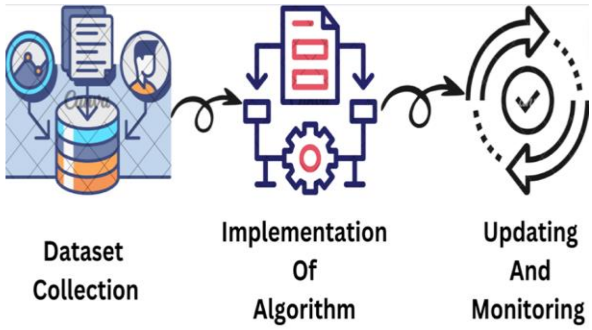

Fig. 13. workflow of system.

XI. Conclusion

This research has focused on the significant role of machine learning as a branch of artificial intelligence for the early detection of non-communicable diseases (NCDs). These diseases, such as diabetes, cancer, heart problems, and blood pressure, have been prevalent globally, and the advancements in early detection have a profound effect on patient outcomes and controlling mortality rates. The results of research show that leveraging large datasets of patient records, such as medical records, laboratory records, and X-rays, and applying different machine learning algorithms can achieve high accuracy results related to early detection of NCDs. By applying different methodologies of machine learning, such as convolutional neural network, support vector machine, and decision tree which are suitable for training datasets, one can achieve the early detection of NCDs. On the other hand, regular updates of the system and continuous monitoring help to maintain high accuracy results of the system. To implement the system, some limitations were faced, related to data collection, and its quality because they might be inconsistent or incomplete, privacy of patient records, limitation in data collection and computational resources, which are related to power requirements for training data. These are some limitations that were faced in implementing the system. One consideration of the system was related to the ethical aspects, including privacy of patient records and informed consent. All in all, based on the results, machine learning has a significant effect on the early detection of NCDs, improving treatment outcomes, and reducing the number of mortalities related to NCDs.

XX. References

1. Wagner, K.-H., & Brath, H. (2012). A global view on the development of non-communicable diseases. *Preventive Medicine, 54*(Suppl.), S38–S41. https://doi.org/10.1016/j.ypmed.2011.11.012
2. Shaheen, M. Y. (2021). Applications of artificial intelligence (AI) in healthcare: A review. *Humanities Commons CORE*.
3. Budreviciute, A., Damiati, S., Sabir, D. K., Onder, K., Schuller-Goetzburg, P., Plakys, G., Katileviciute, A., Khoja, S., & Kodzius, R. (2020). Management and prevention strategies for non-communicable diseases (NCDs) and their risk factors. *Frontiers in Public Health, 8*, Article 574111.
4. Benning, L., Peintner, A., & Peintner, L. (2022). Advances in and the applicability of machine learning-based screening and early detection approaches for cancer: A primer. *Cancers, 14*(3), Article 623.
5. Brownlee, J. (2019, August 15). What is deep learning? *MachineLearningMastery.com*.
6. Ghnaim, M. (2022). Machine learning models to predict early complications of NCD patients. *ProQuest*.
7. Ntohsi, S. (2022, April 10). Detection system for non-communicable diseases using machine learning. *ResearchGate*.
8. Bigna, J. J., & Noubiap, J. J. (2019). The rising burden of non-communicable diseases in sub-Saharan Africa. *The Lancet Global Health, 7*(10), e1295–e1296.
9. Skrede, O.-J., De Raedt, S., Kleppe, A., Hveem, T. S., Liestøl, K., Maddison, J., Askautrud, H. A., Pradhan, M., Nesheim, J. A., Albregtsen, F., Farstad, I. N., Domingo, E., Church, D. N., Nesbakken, A., Shepherd, N. A., Tomlinson, I., Kerr, R., Novelli, M., Kerr, D., & Danielsen, H. E. (2020). Deep learning for prediction of colorectal cancer outcome: A discovery and validation study. *Lancet, 395*(10221), 350–360.
10. Al-antari, M. A., Hua, C.-H., Bang, J., & Lee, S. (2020). Fast deep learning computer-aided diagnosis of COVID-19 based on digital chest x-ray images. *Applied Intelligence, 51*(5), 2890–2907.
11. Hu, M., Nohara, Y., Wakata, Y., & Nakamura, M. (2018). Machine learning-based prediction of non-communicable diseases to improving intervention program in Bangladesh. *ResearchGate*.
12. NCD Alliance. (2021, December 17). Integrating noncommunicable disease prevention and care into global health initiatives and universal health coverage. *NCD Alliance*.
13. Zhuravel, H. (2023, September 11). AI/ML algorithms for early disease detection and diagnosis. *Binariks.com*.
14. Pei, D., Gong, Y., & Kang, H. (2019). Fig. 3 Decision tree of diabetes classifiers. *ResearchGate*.
15. Akinnuwesi, B. A., Olayanju, K. A., Aribisala, B. S., Fashoto, S. G., Mbunge, E., Okpeku, M., & Owate, P. (2023). Application of support vector machine algorithm for early differential diagnosis of prostate cancer. *Data Science and Management, 6*(1), Article 101232.
16. Pereira, F., Mario, J., Luiz, D., & Ferrari, L. (2021). Figure 2. Flow chart of the NCD and FPR steps. *ResearchGate*.
17. Kanyongo, W., & Ezugwu, A. E. (2023). Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives. *Informatics in Medicine Unlocked, 38*, Article 101232.
18. Li, M., Jiang, Y., Zhang, Y., & Zhu, H. (2023). Medical image analysis using deep learning algorithms. *Frontiers in Public Health, 11*, Article 1273253.
19. Mostafa, B. M., El-Attar, N., Abd-Elhafeez, S., & Awad, W. (2020). Machine and deep learning approaches in genome: Review article. *Alfarama Journal of Basic & Applied Sciences*, 0(0).
20. Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., & Alshurafa, N. (2022). Deep learning in human activity recognition with wearable sensors: A review on advances. *Sensors, 22*(4), Article 1476.
21. Alatawi, A. A., Alomani, S. M., Alhawiti, N. I., & Ayaz, M. (2022). Plant disease detection using AI based VGG-16 model. *International Journal of Advanced Computer Science and Applications, 13*(4).
22. Rehman, S. U., Tarek, N., Magdy, C., Kamel, M., Abdelhalim, M., Melek, A., Mahmoud, L. N., & Sadek, I. (2024). AI-based tool for early detection of Alzheimer’s disease. *Heliyon*, Article e29375.
23. Khandakar, A., Mamun, M., Hasan, A., Kiranyaz, S., Rahman, T., Alfkey, R., Ashrif, A., & Malik, R. A. (2021). A machine learning model for early detection of diabetic foot using thermogram images. *Computers in Biology and Medicine, 137*, Article 104838.

24-Wang, W., Lee, J., Harrou, F., & Sun, Y. (2020). Early detection of Parkinson’s disease using deep learning and machine learning. *IEEE Access, 8*, 147635–147646.

25. Roy, M., Protity, A. T., Das, S., & Dhar, P. (2023). Prevalence and major risk factors of non-communicable diseases: A machine learning-based cross-sectional study. *Eureka, Health Sciences, 3*, 28–45.