A Comparative Analysis of Machine Learning Based Lightweight Disease Curator System

Hrishikesh Ghosh
Computer Science & Engineering
Calcutta Institute of Engineering & Management
Kolkata,India
hrishikeshghoshofficial@gmail.com

Ankit Mondal
Computer Science & Engineering
Calcutta Institute of Engineering & Management
Kolkata,India
mondalankit0508@gmail.com

Shemanti Pal
Computer Science & Engineering
JIS University
Kolkata,India
shemantipal.sun@gmail.com

Sandip Mandal
Computer Science & Engineering
JIS University
Kolkata,India
sandipmandal816708@gmail.com

Sayak Rudra
Computer Science & Engineering
JIS University
Kolkata,India
sayakrudra2244@gmail.com
Computer Science & Engineering

Paramita Sarkar
Assistant Professor, Computer Science & Engineering
JIS University
Kolkata,India
mailtoparo@gmail.com

ABSTRACT

Machine learning has become an important tool for the identification and management of different health problems, such as heart disease, renal failure or diabetic diseases. Machine Learning algorithms may detect patterns and signs of disease risk, based on an analysis of a lot of data from patients. In the case of heart failure, machine learning models can analyze diverse data sources such as medical records, electrocardiograms (ECGs), and imaging scans to predict the likelihood of heart failure development. By detecting early signs and risk factors, healthcare professionals can intervene proactively and provide appropriate treatment plans to improve patient outcomes. In the case of kidney failure, machine learning techniques can be used to analyse data from laboratory tests, imaging studies and demographics with a view to identifying signs of loss of function and predicting the risk for progression into End Stage Renal Disease. Early detection allows for timely interventions and the implementation of strategies to slow down disease progression. Machine learning also plays a crucial role in diabetes detection. By analyzing patient data, including glucose levels, lifestyle factors, and medical history, machine learning algorithms can predict the risk of developing diabetes and identify individuals who may be at a higher risk of complications. This enables healthcare providers to implement preventive measures, personalized treatment plans, and monitoring strategies to manage the disease effectively. We propose an efficient yet lightweight system for the detection of Diabetes, Heart Failure, and Chronic Kidney Disease. In addition to the detection capabilities, our system offers a doctor recommendation system and access to blogs where patients with similar conditions share their experiences.
Keywords: Early detection, deep learning, SVM, diabetes, heart failure, doctor recommendation system.

I. INTRODUCTION

High blood sugar levels caused by either inadequate insulin production or inefficient insulin usage characterise diabetes, a chronic metabolic condition. It has a serious impact on millions of individuals globally and is a major global health issue. Diabetes must be identified and managed early in order to reduce complications and enhance patient outcomes. In the identification and prediction of diabetes, machine learning approaches have become useful tools, enabling prompt interventions and individualised treatment strategies. Large datasets comprising patient information, including age, gender, body mass index (BMI), family history, and blood test results, can be analysed by machine learning algorithms to find patterns and signals linked to diabetes. These algorithms gain knowledge from the information and create models that can precisely forecast the likelihood. One commonly used technique in diabetes detection is logistic regression. This algorithm determines the probability of an individual having diabetes based on a set of input features. By training the model with labelled data [5], it can learn the relationship between the input variables and the presence of diabetes, allowing it to make predictions for new, unseen data. Deep learning techniques, particularly neural networks, have demonstrated promise in the identification of diabetes in recent years. Deep learning models can analyze complex relationships within the data and automatically extract relevant features for classification. These models can handle large scale datasets, allowing for more accurate predictions and capturing subtle patterns that may go unnoticed by traditional statistical techniques. To improve the accuracy of diabetes detection models, feature selection and dimensionality reduction techniques are often employed. By lowering the complexity of the model and improving its performance, these strategies aid in the identification of the most instructive and pertinent aspects for the classification assignment. Machine learning methods can also help with the early detection of issues linked to diabetes. By analyzing additional patient information, such as blood pressure, cholesterol levels, and kidney function, algorithms can assess the

risk of developing complications like diabetic retinopathy, neuropathy, and nephropathy. Early identification of these risks allows healthcare professionals to intervene with targeted preventive measures and treatment plans.

A critical medical illness known as heart failure occurs when the heart is unable to pump enough blood to meet the body's needs. Numerous things, including underlying cardiovascular conditions including coronary artery disease, high blood pressure, or prior heart attacks, can contribute to it. Early detection of heart failure is crucial for effective management and improved patient outcomes. In the identification and prediction of heart failure, machine learning approaches have proven to be useful tools, enabling prompt interventions and individualized treatment strategies. Machine learning algorithms can analyze large datasets containing patient information, such as medical records, electrocardiograms (ECGs), echocardiograms, and laboratory test results, to identify patterns and markers associated with heart failure. These algorithms learn from the data and develop models that can accurately predict the likelihood of heart failure development in individuals. The use of decision trees is one method frequently employed in the identification of heart failure. These models create a tree-like structure based on feature splits to classify individuals into different categories, such as heart failure or non-heart failure. Decision trees are interpretable, allowing healthcare professionals to understand the factors influencing the prediction. Another powerful machine learning technique for heart failure detection is logistic regression. This algorithm determines the probability of an individual having heart failure based on a set of input features. By training the model with labelled data, it can learn the relationship between the input variables and the presence of heart failure, allowing it to make predictions for new, unseen data. Support Vector Machines (SVMs) are also employed in heart failure detection. SVMs identify an ideal hyper plane that divides data points into distinct classes, such as people with heart failure and those without heart failure. By maximizing the margin between classes, SVMs can make accurate predictions based on new data. Deep learning techniques, particularly neural networks, have demonstrated promise in the identification of heart failure in recent years. Deep learning models can analyze complex relationships within the data and automatically extract relevant features for classification. These models can handle large-scale datasets, allowing for more accurate predictions and capturing subtle patterns that may go unnoticed by traditional statistical techniques. Additionally, the early detection of issues associated with heart failure can be facilitated by machine learning approaches. By analyzing additional patient information, such as blood pressure, cholesterol levels, and kidney function, algorithms can assess the risk of developing complications like arrhythmias or pulmonary edema. Early identification of these risks allows healthcare professionals to intervene with targeted preventive measures and treatment plans. Furthermore, machine learning algorithms can also assist in the monitoring of heart failure patients. By analyzing data from wearable devices, such as heart rate monitors or implantable devices, these algorithms can detect changes in heart rate variability, fluid retention, or other parameters that may indicate worsening heart failure. This information enables healthcare providers to make timely adjustments to treatment plans and provide personalized care.

Fig. 1. Diabetes Detection

The functioning of the kidneys is impacted by chronic kidney disease (CKD), a disorder that is progressive and incurable [1],[2]. It is characterised by a progressive decline in kidney function over time, which causes the body to swell with waste products and have fluid imbalances. Early detection of CKD is crucial for implementing interventions to slow its progression and manage associated complications. In the identification and prediction of CKD, machine learning techniques have proven to be useful tools [6], [7], enabling prompt interventions and individualised treatment programs. Large datasets comprising patient information, such as medical records, laboratory test results, and demographic data, can be analysed by machine learning algorithms to find patterns and markers connected to CKD. These algorithms take information from the data and create models that can precisely forecast a person’s risk of developing CKD. Deep learning models can handle largescale datasets, allowing for more accurate predictions and capturing subtle patterns that may go unnoticed by traditional statistical techniques. Additionally, machine learning methods can help in the early identification of CKD-related problems. By analyzing additional patient information, such as blood pressure, blood chemistry, and urine analysis, algorithms can assess the risk of developing complications like cardiovascular disease or end-stage renal disease. Early identification of these risks allows healthcare professionals to intervene with targeted preventive measures and treatment plans. Machine learning algorithms can also aid in the monitoring of CKD patients. By analyzing longitudinal data, including regular blood tests, vital signs, and medication adherence, algorithms can detect trends and changes in kidney function. This information enables healthcare providers to make timely adjustments to treatment plans, identify disease progression, and provide personalized care.

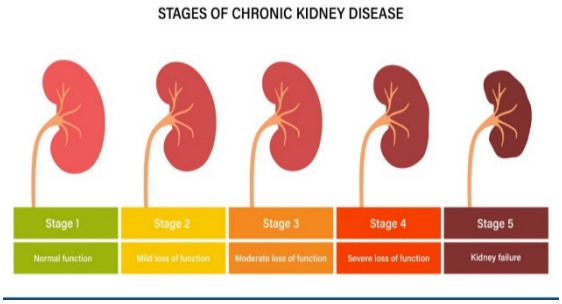

Fig. 2. Different Stages of Chronic Kidney Disease

In order to increase the precision and efficacy of heart failure, diabetes, and chronic kidney disease diagnosis, we conducted three thorough comparative studies that extensively analysed and compared various machine learning models on different datasets. These three conditions are significant health concerns that require early identification and management for optimal patient outcomes. In our studies, we assessed the performance of various machine learning models for detecting each individual condition, including logistic regression, decision trees, support vector machines (SVM), and deep learning techniques. The models were trained and tested on diverse datasets, consisting of patient records, clinical measurements, and relevant biomarkers. We meticulously assessed their accuracy, sensitivity, specificity, and other performance metrics to identify the models that exhibited the highest levels of accuracy and reliability. The selected models, based on their superior performance, were then utilized to develop an application programming interface (API) for our detection system. The API allows seamless integration of the chosen models into a user-friendly software platform, enabling healthcare professionals to input patient data and obtain prompt and accurate predictions regarding the presence or likelihood of heart failure, diabetes, or chronic kidney disease. This streamlined approach aids in early detection and facilitates timely interventions, ultimately improving patient outcomes and quality of care. Furthermore, in order to provide comprehensive assistance to healthcare providers and patients, we incorporated a recommendation system into the software platform. This recommendation system employs a simple yet effective weight-based mathematical model to suggest doctors with the highest ratings and expertise in their respective fields. By considering factors such as patient ratings, doctor qualifications, and experience, the system assists users in making informed decisions when selecting healthcare professionals for further consultation and treatment. By combining robust machine learning models for disease detection with a sophisticated recommendation system for doctor selection, our system aims to enhance the overall diagnostic process and patient experience. Early identification, accurate forecasting, and personalised suggestions are made possible by the integration of these elements, which improves patient satisfaction and results in lower costs and better healthcare outcomes. As the field of machine learning continues to evolve, we remain committed to exploring novel approaches and incorporating the latest advancements in technology to further refine and expand the capabilities of our detection system. Our ultimate mission is to improve the lives of people with chronic renal disease, heart failure, and diabetes by advancing medical diagnostics, advocating proactive healthcare management, and promoting proactive healthcare management.

II. Related Work

Using machine learning approaches for diabetes detection has made great strides in recent years. Researchers have employed various algorithms and data analysis methods to develop accurate models that can predict the likelihood of diabetes, aiding in early detection and prevention strategies. In order to create a system that is capable of reliably forecasting the onset of diabetes in patients, K.VijiyaKumar et al. [5] created the Random Forest algorithm for diabetes prediction. Their proposed model demonstrated excellent results in predicting diabetes and showcased the system’s effectiveness, efficiency, and real-time capabilities. Five commonly used classifiers and a meta-classifier were utilised in Nonso Nnamoko et al.'s [7] ensemble supervised learning strategy to predict the onset of diabetes. The results showed that the proposed strategy achieves superior accuracy in predicting the onset of diabetes when compared with research that used the same dataset. The prediction of diabetes has gained significant attention from researchers, as they aim to train programs to effectively identify diabetic patients using appropriate classifiers on the dataset.

Machine learning has demonstrated encouraging outcomes in the field of heart disease identification during the past few years. Large datasets and cutting-edge algorithms have been used by researchers to build models that can accurately predict the presence of cardiac disease, enabling early intervention and individualised treatment methods. To effectively predict cardiovascular illness, a variety of algorithms have been used, including Logistic Regression, KNN, and Random Forest Classifier. The outcomes illustrate each algorithm's advantages in reaching particular goals [3]. The decision boundary for the model containing IHDPS was determined using both old and modern machine learning and deep learning models. This made it easier to include important risk factors for heart disease, like family history. The IHDPS model's accuracy, however, was inferior to that of more recent models, such as those that identified coronary heart disease using artificial neural networks and other machine and deep learning techniques. Using neural network techniques and built-in implementation algorithms, McPherson et al.'s [4] research effectively predicted if a test patient had coronary heart disease or atherosclerosis by identifying the risk variables for the diseases. Neural networks were first used to diagnose and predict Heart Disease, Blood Pressure, and other characteristics by R. Subramanian et al. [8]. They created a deep neural network with 120 hidden layers that included pertinent disease-related features and assured correct findings when used on test datasets. For the purpose of diagnosing cardiac disease, the supervised network has been suggested [9]. The model's accuracy was determined when it successfully predicted the outcomes of a doctor's test utilising unfamiliar data using previously learnt data.

For the effective classification of chronic renal disease using patient data over the past few years, a variety of machine-learning algorithms have been employed. Using a dataset of Indians with Chronic Kidney Disease (CKD), Charleonnan et al. [2] carried out a comparative examination of several prediction models, including K-nearest neighbours (KNN), support vector machines (SVM), logistic regression (LR), and decision trees (DT). The objective was to identify the best classifier for correctly predicting the development of chronic renal disease. According to their research, SVM had the highest classification accuracy, reaching a whopping

98.3%, as well as the highest sensitivity rate, 0.99. These results highlight the superiority of SVM as a predictive model in identifying chronic kidney disease in the Indian population based on the examined dataset. In their study, Salekin and Stankovic [10] evaluated different classifiers on a dataset of 400 occurrences, including K-NN, RF, and ANN. The study utilized wrapper feature selection techniques to identify the most relevant features for model construction. As a result of their research, RF showed the highest classification accuracy of 98% and the lowest root mean square error (RMSE) value, 0.11, demonstrating its potency as a predictive model. Similar to this, S.Tekale et al. [11] concentrated on utilising machine learning algorithms to predict chronic renal disease. They utilized a dataset comprising 400 instances and initially consisting of 25 features. Through preprocessing steps, the number of features was reduced to 14. For their analysis, the authors used the decision tree and support vector machine (SVM) algorithms. The outcomes showed that SVM performed better than the decision tree model, averaging a 96.75% accuracy rate. These findings highlight how well machine learning algorithms are able to forecast chronic kidney disease. Salekin and Stankovic's research highlighted RF as the best classifier, exhibiting high accuracy and low RMSE. On the other hand, S. Tekale et al. emphasized the superiority of SVM over the decision tree model in terms of accuracy. These models and feature selection methods can help researchers better identify and forecast chronic kidney disease, allowing for early interventions and individualised treatment programs for those who are at risk.
In their study, Xiao et al. [12] proposed a prediction model for the course of chronic renal disease. They used a number of machine learning algorithms, such as neural networks, logistic regression, elastic net, lasso regression, ridge regression, support vector machines, random forests, XGBoost, and k-nearest neighbours. A dataset including the historical data of 551 patients with proteinuria and 18 associated features was used to compare the models' performances. The outcomes were classified into three categories: mild, moderate, and severe. After conducting their analysis, the researchers concluded that logistic regression yielded the best results among the tested models. With an area under the curve (AUC) value of 0.873, it successfully predicted the course of chronic renal disease. Additionally, logistic regression demonstrated a sensitivity rate of 0.83 and a specificity rate of 0.82, showing that it is capable of correctly categorizing patients into the proper illness severity groups. The findings of Xiao et al.'s study emphasize the potential of logistic regression as a valuable tool in predicting the progression of chronic kidney disease. By leveraging historical patient data and the selected features, this model can contribute to early detection and intervention, enabling healthcare professionals to tailor treatment plans and interventions based on the predicted disease severity. In their study, Mohammed and Beshah [6] focused on creating a self-learning knowledge based system utilizing machine learning for the diagnosis and treatment of the initial three stages of chronic kidney disease. The research involved a limited dataset, and the researchers developed a prototype that allows patients to query the knowledge-based system for receiving advice. The decision tree algorithm was utilized to generate the rules for the system. The prototype demonstrated an overall accuracy rate of 91%, indicating its effectiveness in assisting with the diagnosis and treatment process. Research was done by Almasoud and Ward [1] to determine how well machine learning algorithms can foretell chronic kidney disease. Their objective was to determine the predictive capabilities of these algorithms using a specific set of features. They used statistical tests including Pearson correlation, ANOVA, and Cramer's V to choose the most pertinent traits. For modelling objectives, they applied the machine learning algorithms LR, SVM, RF, and GB. In the end, they discovered that Gradient Boosting demonstrated the maximum accuracy, obtaining a remarkable F-measure of 99.1.

III. Datasets

In our work, we utilized three distinct datasets to train the various models. These datasets encompassed the Pima Indian Diabetes Dataset, the Heart Failure Prediction Dataset, and the Chronic Kidney Disease dataset.

I.Pima Indian Diabetes Dataset

A well-known and often used dataset in machine learning and data analysis is the Pima Indians Diabetes dataset. It consists of medical data from Pima Indian women, collected with the aim of predicting the onset of diabetes within five years. The dataset contains 9 columns

- glucose levels
- blood pressure
- body mass index
- number of pregnancies that occurred
- skin thickness,
- insulin level
- Diabetes Pedigree Function
- age
- the binary target variable indicating the presence or absence of diabetes.

II. Heart Failure Prediction Dataset

A frequently used dataset in the fields of machine learning and cardiovascular research is the Heart Failure Prediction dataset. It is made out of medical information from heart failure patients and aims to forecast the risk that a patient would pass away from heart failure. There are 13 columns in the dataset

• age

- anaemia
- Level of the CPK [creatinine phosphokinase] enzyme
- ejection fraction[percentage of blood leaving the heart with each contraction]
- platelet count
- serum creatinine
- serum sodium
- sex
- smoking status
- blood pressure
- diabetes
- time [follow-up time period]
- Death Event [If the patient deceased during the follow-up period].

Researchers and data scientists can create predictive models using this dataset to recognise high-risk patients and offer prompt solutions.

III. Chronic Kidney Disease dataset

The Chronic Kidney Disease (CKD) dataset is an important tool for machine learning and nephrology. It includes medical information from patients who may have renal disease and aims to determine if chronic kidney disease exists or not. There are 26 columns in the dataset-

- Age: age in years
- Blood Pressure
- Specific Gravity
- Albumin
- Sugar
- Red Blood Cells in urine • Pus Cell in urine
- Pus Cell clumps in urine
- Bacteria in urine
- Blood Glucose Random
- Blood Urea

- Serum Creatinine
- Sodium
- Potassium
- Hemoglobin
- Packed Cell Volume
- White Blood Cell Count
- Red Blood Cell Count
- Hypertension
- Diabetes Mellitus
- Coronary Artery Disease
- Appetite
- Pedal Edema
- Anemia
- Class

Researchers and data analysts utilize this dataset to develop predictive models that aid in early detection and management of CKD. The information covered in the introduction [4] should be expanded upon in this part, not repeated. A Calculation Section, on the other hand, is a theoretical advancement that has a practical application [5].

IV. Experimental Method/Procedure/Design

The information regarding our projected work is included in this part. This part contains information about your proposed models, approaches, algorithms, flowcharts, and other works [6, 7].

Machine Learning Algorithms

We have included a few traditional machine learning algorithms in our suggested methodology. Computers may learn from data and make predictions or take actions without being explicitly programmed thanks to machine learning algorithms. With the use of these algorithms, precise predictions or classifications can be made by analysing huge datasets, finding patterns, and generalizing data. Machine learning algorithms come in a variety of forms, such as supervised learning algorithms like decision trees and support vector machines, unsupervised learning algorithms like clustering and dimensionality reduction, and reinforcement learning algorithms that pick up information through trial-and-error interactions with the environment.

4.1 Decision Tree

An efficient machine learning approach that may be utilised for both classification and regression applications is the decision tree. It is a non-parametric algorithm that creates a model resembling a tree of decisions and potential outcomes. The root node of the decision tree algorithm is first chosen as the dataset's most important property. Following that, it divides the data into branches or child nodes based on various attribute values. Until a stopping requirement is satisfied, such as reaching a maximum depth or a minimum quantity of data points in a leaf node, this process is repeated iteratively for each child node. Decision trees' interpretability is its main benefit. The resulting tree structure is excellent for explaining the decision-making process since it is simple to comprehend and visualise. A decision rule based on the values of the attributes is represented by each path from the root to a leaf node. Decision trees are also capable of handling both categorical and numerical features, as well as missing data. They can handle outliers and are relatively robust to noise in the dataset. Moreover, decision trees can automatically select the most informative features and handle interactions between variables. Decision trees, however, are susceptible to overfitting, which happens when they memorise the training data too well and are unable to generalise to new data. This problem can be reduced using strategies like pruning, establishing a maximum

depth, or using regularisation. Decision trees are frequently enhanced and overfitting is minimised using ensemble approaches like random forests and gradient boosting.

Random Forest

The ensemble learning algorithm Random Forest combines the advantages of decision trees with the ideas of bootstrapping and random feature selection. It is a strong and flexible method. It is commonly used in machine learning for both classification and regression tasks. A large number of decision trees are built using Random Forest, and their forecasts are then combined to get the final prediction. A separate subset of the training data is used to train each decision tree in the forest, which is chosen by bootstrapping. Additionally, only a random subset of features are taken into account at each decision tree split, which lessens the association across different trees. Random Forest's capacity to handle high-dimensional datasets with a lot of features is its main benefit. It can effectively capture complex interactions between variables and handle outliers and missing data. Random Forest is also less prone to overfitting compared to a single decision tree, as the ensemble approach helps to improve generalization performance. Additionally, Random Forest offers measures of feature relevance that help us comprehend the relative contributions made by various features to the prediction process. The selection of features and comprehension of the underlying relationships in the data can both benefit from this knowledge. However, Random Forest comes with a trade-off between interpretability and complexity. While the individual decision trees in the forest can be easily interpreted, the ensemble as a whole may be more challenging to interpret due to its collective decision-making process. Random Forest has numerous applications in various fields, including finance, healthcare, bioinformatics, and remote sensing. It is a well-liked option for many practical machine learning problems due to its adaptability, robustness, and capacity for handling high-dimensional data.

Gradient Boosting

A potent machine learning approach called gradient boosting combines the benefits of decision trees with iterative optimisation. It is well known for its capacity to provide extremely accurate predictions and is commonly used for both classification and regression applications. Gradient Boosting builds decision trees onto the model one at a time, with each new tree being trained to fix the errors of the preceding ones. The algorithm calculates the loss function gradients with respect to the anticipated values at each iteration, and then fits a new tree to the negative gradient values. The method gradually reduces the overall loss, enhancing the performance of the model. Gradient Boosting's capacity to manage diverse data and capture intricate correlations between characteristics is one of its main advantages. It can effectively handle both numerical and categorical variables, as well as missing data. Additionally, Gradient Boosting allows for flexible loss functions, enabling customization based on the specific problem domain. Moreover, Gradient Boosting incorporates regularization techniques, such as shrinkage and feature subsampling, to control overfitting. This helps to improve the generalization performance and makes the algorithm more robust to noise and outliers in the data. However, Gradient Boosting can be computationally expensive and sensitive to hyper parameter tuning. Fine-tuning the learning rate, number of trees, and tree depth is crucial to achieve optimal performance. The ensemble nature and complexity of the process may also make it difficult to interpret the resultant model. Gradient boosting has gained popularity and been successfully used in a number of industries, including healthcare, financial forecasting, and click-through rate prediction. Its ability to generate accurate predictions and handle complex data makes it a valuable tool for advanced machine learning tasks.

4.2 Logistic Regression

The statistical and machine learning approach known as logistic regression is commonly utilised for binary classification applications. Modelling the link between a group of input variables (features) and the likelihood of a binary output is an extension of linear regression. Contrary to linear regression, which forecasts continuous values, logistic regression converts the linear combination of input data into a probability score between 0 and 1 using the logistic function (sometimes called the sigmoid function). The likelihood that the binary outcome will be either positive or negative is represented by this probability. By increasing or decreasing the log loss function, the logistic regression algorithm calculates the ideal coefficients for the input features. These coefficients establish how much each feature contributed to the prediction, allowing for the understanding of how it affected the result. The benefits of logistic regression are numerous. It is a rather straightforward approach with good computing performance that can handle both numerical and categorical input features. It provides interpretable results, as the coefficients can be used to understand the direction and strength of the relationship between the features and the outcome. Additionally, logistic regression can handle large datasets, is robust to outliers, and can handle multi collinearity among the features. Logistic regression, however, has several drawbacks. It presupposes that the attributes and the log-odds of the result have a linear relationship. The use of alternative algorithms or feature engineering may be necessary for non-linear interactions. However, by using methods like one-vs-rest or softmax regression, logistic regression can be expanded to handle multi-class problems. Logistic regression is also primarily intended for binary classification applications.

4.3 SVMs - Support Vector Machines

Powerful machine learning algorithms called Support Vector Machines (SVMs) excel in both classification and regression tasks. SVMs are renowned for their prowess in high-dimensional areas and their capacity to manage complex data. Finding the best hyper plane to divide the data into distinct classes is the main goal of SVM. The hyper plane is selected in a way that maximises the distance between it and the nearest data points for each class, known as the support vectors. Maximum margin categorization is the name given to this method. By utilising a method known as the kernel trick, SVMs are able to handle both linearly separable and non-linearly separable data. The input data are transformed into a higher dimensional space using the kernel function, where they might become

linearly separable. Intricate patterns in the data can be captured and complex decision limits can be learned by SVMs as a result. The capacity of SVMs to handle overfitting and generalise well to new data is one of its main advantages. The margin maximization approach provides a natural form of regularization, preventing the model from becoming overly sensitive to individual data points. SVMs also have a solid theoretical foundation and rely on convex optimization techniques, ensuring global optimality in finding the optimal hyper plane. SVMs, however, can be expensive to compute, particularly when working with huge datasets. Furthermore, selecting the proper kernel function and fine-tuning the hyperparameters might be difficult. The interpretability of SVMs may also be limited, particularly when using non-linear kernels. SVMs have proven to be effective in various applications, including image recognition, text classification, bioinformatics, and finance. Their versatility, robustness, and ability to handle complex data make them a valuable tool in the machine learning toolbox.

4.4 KNN Classifier

A straightforward yet effective machine learning technique used for classification and regression problems is the k-Nearest Neighbours (KNN) classifier. The fundamental principle of KNN is to categorize new instances according to the majority vote of their closest neighbors in the feature space. The algorithm for KNN classification determines the separation between a test instance and each instance in the training set. The majority class label among those neighbors is then applied to the test instance after choosing the test instance's closest k neighbours (where k is a user-defined option). The choice of distance metric, such as Euclidean or Manhattan distance, determines the proximity measure. One of the major advantages of KNN is its simplicity and ease of implementation. It does not require any explicit training or model building phase, making it a non-parametric algorithm. KNN is flexible in a variety of applications since it can handle both categorical and numerical input. KNN does have some restrictions, though. As it needs figuring out distances for each test instance, it can be computationally expensive, especially with big datasets or high-dimensional feature spaces. The bias-variance trade-off can be impacted by the choice of k. While a big k value can result in oversimplification, a little k value can cause overfitting. Additionally, KNN is sensitive to the scale and relevance of features. Pre-processing the data, such as scaling or feature selection, may be necessary for better performance. Interpretability is another concern, as KNN does not provide explicit rules or feature importance measures. Despite its limitations, KNN remains a popular algorithm for its simplicity, flexibility, and competitive performance. Applications for it include anomaly detection, recommendation systems, and pattern recognition.

ANNs - Artificial Neural Networks

Artificial neural networks (ANNs), also called neural networks or simply neural networks, are a potent class of machine learning algorithms that take their cues from the design and operation of biological neural networks seen in the human brain. ANNs are frequently employed to tackle challenging issues like pattern recognition, regression, and classification challenges. An input layer, one or more hidden layers, and an output layer are the three layers that make up an artificial neural network (ANN). Each neuron receives input values, computes them using activation functions and weights, and then sends the results to the following layer. The corresponding weights of the connections between neurons are modified during training. The ability of ANNs to learn from and generalise vast volumes of data is their key competitive advantage. Through a process called training, they may automatically discover intricate patterns and connections within the data. In order to reduce error and enhance the performance of the model, training entails repeatedly modifying the connection weights based on observed error or loss. Recurrent ANNs allow information to flow in cycles through feedback connections, allowing the network to have memory and handle sequential data. Feed forward ANNs allow information to flow from the input layer to the output layer in a single direction. The main goal of deep learning, a kind of machine learning, is to train neural networks with many hidden layers. Deep neural networks, also known as Deep ANNs, have shown to perform exceptionally well in a number of fields, including speech recognition, computer vision, and natural language processing. However, ANNs can be computationally intensive, requiring significant computational resources and large amounts of training data. They may also suffer from overfitting if the network is too complex or the training data is insufficient. Tuning the architecture, choosing appropriate activation functions, and regularization techniques are essential for achieving optimal performance. Despite the challenges, ANNs have revolutionized the field of machine learning and have become a cornerstone in many state-of-the-art models and applications. Their ability to learn complex patterns and extract meaningful representations from data makes them a powerful tool in solving a wide range of real-world problems.

V. Methodology

The datasets we worked with, as mentioned before, were structured as tabular data, which meant that models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) were not suitable since they are typically used for image data or sequential data. To ensure the reliability of our data, we performed pre-processing tasks that involved identifying and removing any rows that contained missing or corrupted data. This ensured that our dataset was clean and ready for analysis. We then proceeded to train and evaluate various machine learning models on the pre-processed data, including Decision Trees, Support Vector Machines (SVM), and Artificial Neural Networks (ANN), among others. All the models were trained on a T4 GPU environment and 16 GB RAM. During the training and testing phases, we closely monitored the accuracy of each model to gauge its performance on the given dataset.

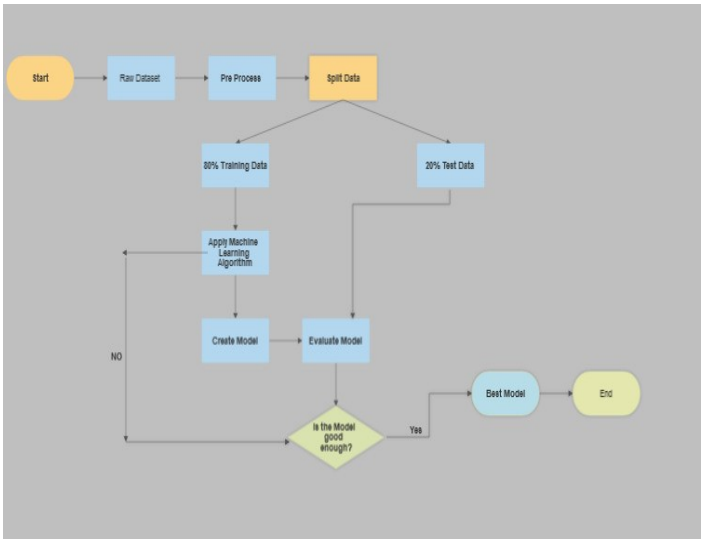

Fig. 3. Proposed workflow diagram

VI. Results and Discussion

Our main goal in our real-world disease detection study was to create machine learning models that could correctly identify diseases. Given the importance of accurate classification, we prioritized achieving high accuracy rather than focusing solely on creating lightweight models. To ensure that our models had both accuracy and efficiency, we conducted a comprehensive comparative study. We evaluated various lightweight machine learning and deep learning models, taking into consideration their performance in terms of accuracy. We aimed to strike a balance between model complexity and accuracy to meet the requirements of our project. One of the challenges we encountered was the limited availability of medical data. Obtaining a substantial amount of medical data can be a cumbersome process due to privacy concerns and data access limitations. Consequently, we needed to adapt our approach accordingly. Deep learning models, particularly deeper neural networks, typically require a large amount of data to achieve optimal results. However, due to our limited dataset, we opted for shallower artificial neural networks (ANNs) and machine learning algorithms. These models offered a good compromise between accuracy and data requirements, allowing us to make the most of the available dataset. By leveraging shallower ANNs and machine learning algorithms, we were able to develop models that provided accurate disease classification results while mitigating the challenges associated with limited data availability. This approach enabled us to effectively detect diseases in real-life scenarios, contributing to improved healthcare outcomes and decision-making processes.

VII. Result analysis

We performed training on each of the aforementioned datasets using a number of models, paying close attention to accuracy both during training and testing. As a result, we were able to evaluate how well each model performed using the particular dataset. By evaluating the accuracy at different stages, we were able to identify the models that demonstrated the highest accuracy and overall performance, shown in Table 1, 2 and 3. The process of selecting the best models was crucial for subsequent use in our project.

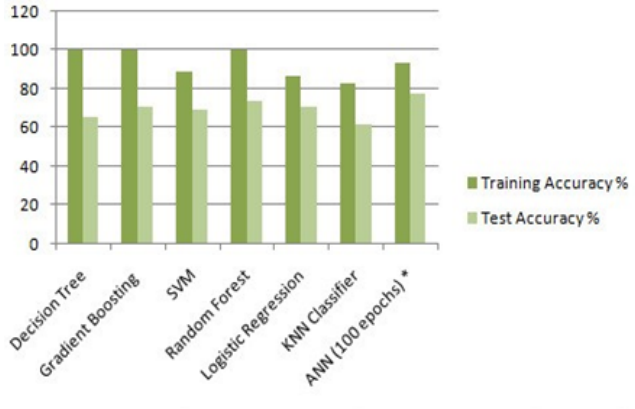

Fig. 4. Accuracy of Different Models on Heart Disease Detection

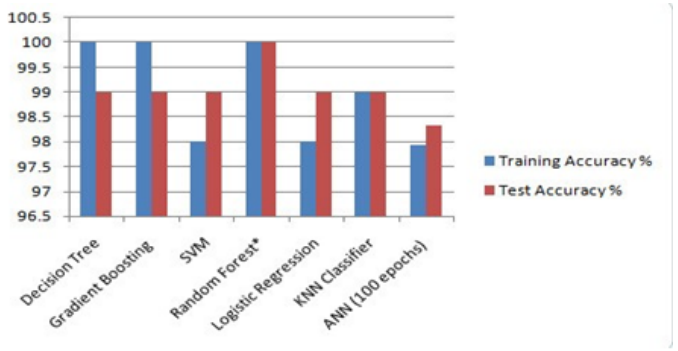

Fig. 5. Accuracy of Different Models on CKD Detection

By comparing the results obtained from different models, we were able to determine which models exhibited the most favourable performance characteristics. The selected best models were chosen based on their superior accuracy and ability to generalize well to new and unseen data. These models demonstrated robust performance during both training and testing, indicating their reliability and suitability for the specific dataset they were trained on. The identification and selection of the best models were instrumental in our project, as they provided the foundation for further analysis and implementation. These models contributed to the overall success and efficacy of our work by providing a useful resource for forecasting and deriving insights from the data.

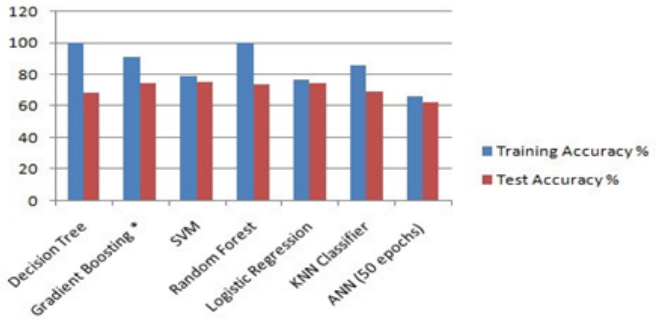

Fig. 6. : Accuracy of Different Models on Diabetes Detection

Table 1: Model Accuracy during Training and Test Phase for Diabetes

Model	Training Accuracy(%)	Test Accuracy(%)
Decision Tree	100	65.33
Gradient Boosting	100	70.66
SVM	88.83	69.33
Random Forest	100	73.33
Logistic Regression	86.60	70.66

Table 2: Model Accuracy during Training and Test Phase for Heart Disease

Model	Training Accuracy(%)	Test Accuracy(%)
Decision Tree	100	68.22
Gradient Boosting	90.79	74.47
SVM	78.99	75.52
Random Forest	100	73.95

Logistic Regression	76.56	74.47
KNN Classifier	86.11	68.75
ANN(50 epochs)	66.30	62.07

In the context of Diabetes Detection, while training the models, we observed that Decision Tree and Random Forest initially displayed higher accuracy. However, during the test phase, their performance significantly deteriorated. This led us to conclude that both models were overfitting the training data, resulting in poor generalization. To address this issue, we explored alternative models and found that Gradient Boosting showed promising results in both the training and test phases. Unlike the previously used models, Gradient Boosting demonstrated the ability to generalize well to unseen data, indicating its robustness and suitability for the task of Diabetes Detection. By opting for Gradient Boosting, we aimed to overcome the overfitting problem that plagued the Decision Tree and Random Forest models.

The improved performance of Gradient Boosting in both training and test phases instilled confidence in its ability to provide accurate and reliable predictions for diabetes detection. Our decision to transition to Gradient Boosting was based on the desire to select a model that not only demonstrated high accuracy during training but also maintained strong performance when faced with new and unseen data. This strategic shift in model selection helped us enhance the effectiveness and reliability of our Diabetes Detection system. For Heat Disease a basic ANN architecture with 4 hidden units were used, which achieved an accuracy of 93.30% . Random Forest Classifier outperformed all the other techniques by achieving 100% accuracy in both the training and test phase.

Table 3: Comparison Table with Different Algorithm

Model	Training Accuracy(%)	Test Accuracy(%)
Decision Tree	100	99
Gradient Boosting	100	99
SVM	98	99
Random Forest	100	100
Logistic Regression	98	99
KNN Classifier	99	99
ANN(100 epochs)	97.92	98.33

VIII. CONCLUSION

The Machine Learning Based Lightweight Disease Curator System aims to develop an advanced system capable of predicting three major health conditions: diabetes, heart failure, and chronic kidney disorder. To system capable of predicting three major health conditions: diabetes, heart failure, and chronic kidney disorder. To provide precise predictions and aid healthcare professionals in early detection and intervention, this system combines machine learning algorithms, data analysis methods, and medical knowledge. We have used the Pima Indian Dataset for Diabetes Prediction, the Heart Failure Prediction dataset for heart failure, and Chronic Kidney Disease dataset for chronic kidney diseases. The collected data then undergoes pre-processing to ensure its quality, completeness, and consistency. This step involves data cleaning, handling missing values, and standardizing the data to create a uniform and reliable dataset for analysis. Once the relevant features are identified, machine learning models are developed to predict the likelihood of each health condition. The labelled data is used to train sophisticated algorithms like Gradient Boosting, ANN, and Random Forest so they can understand patterns and make predictions based on fresh, unforeseen data. Utilising suitable evaluation criteria like accuracy and precision, the performance of the prediction models is assessed. We have done a comparative study and only gone with the most accurate ones. The effectiveness of the system in correctly forecasting the existence or risk of diabetes, heart failure, and chronic renal disease is evaluated as part of this review procedure. The project aims to provide healthcare professionals with a valuable tool for early detection and intervention, ultimately leading to improved patient outcomes. By leveraging machine learning and data analysis techniques, this system has the potential to enhance medical decision-making, facilitate timely intervention, and reduce the burden of these chronic diseases on individuals and healthcare systems. Apart from this, we are also incorporating a Doctor Appointment Portal. If the outcome comes positive, we will recommend certain doctors who are willing to provide diagnosis online and who may partner with our system. We will list all these online doctor chambers sorting them with our review algorithm systems. The user can select any doctor from the list. We shall not provide direct communication to the doctor keeping in mind that the doctor has a fixed schedule to provide online treatments. So we will ask our users to book a slot or appointment to do an online diagnosis, We may also provide an offline appointment of a doctor in the same slot booking algorithm. There may be a condition

when the outcome is negative i.e. the patient is non-diabetic but still feeling sick or unwell. At this part, we have our own GPT 3.5 which produces some fantastic health advice whenever you ask for it.

Conflict of Interest

The authors affirm that there is no conflict of interest.

Authors' Contributions

Author 1 and 2 conducted literature searches and developed the study.
Authors 1, 2, 3, 4,5, and 6 contributed to the creation of the protocol, as well as to patient recruiting, ethical review, and data analysis.
Author-3, 4, and 5 wrote the first draft of the manuscript
Author 6 led, mentored, and reviewed and edited the work of the other authors, and all authors have given their approval to the final product.

Acknowledgements

We sincerely appreciate the chance to work on the project "A Machine Learning Based Lightweight Disease Curator System" under the guidance of Mrs. Paramita Sarkar and Mr. Abhijit Mitra. Without their creative suggestions and his unwavering encouragement, we could never have brought this project to this point.

References

[1] Marwa Almasoud and Tomas E Ward. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. International Journal of Advanced Computer Science and Applications, 10(8), 2019.

[2] Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, and Nitat Ninchawee. Predictive analytics for chronic kidney disease using machine learning techniques. In 2016 Management and Innovation Technology International Conference (MITicon), pages MIT–80–MIT–83, Oct 2016.

[3] Andrea Ganna, Patrik K E Magnusson, Nancy L Pedersen, Ulf de Faire, Marie Reilly, Johan Arnlo¨v, Johan Sundstro¨m, Anders Hamsten, and Erik Ingelsson. Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9):2267—2272, September 2013.

[4] M. Akhil Jabbar, B. L Deekshatulu, and Priti Chandra. Heart disease prediction using lazy associative classification.

[5] Vijiyakumar Krishnan, B. Lavanya, I. Nirmala, and S. Caroline. Random forest algorithm for the prediction of diabetes. pages 1–5, 03 2019.

[6] Siraj Mohammed and Tibebe Beshah. Amharic based knowledge-based system for diagnosis and treatment of chronic kidney disease using machine learning. International Journal of Advanced Computer Science and Applications, 9(11), 2018.

[7] Nonso Nnamoko, Abir Hussain, and David England. Predicting diabetes onset: An ensemble supervised learning approach. In 2018 IEEE Congress on Evolutionary Computation (CEC), pages 1–7, 2018.

[8] Latha Parthiban and R Subramanian. Intelligent heart disease prediction system using canfis and genetic algorithm. International Journal of Biological, Biomedical and Medical Sciences, 3(3), 2008.

[9] M Raihan, Saikat Mondal, Arun More, Pritam Khan Boni, and Omar Faruqe Sagor. Smartphone Based Heart Attack Risk Prediction System with Statistical Analysis and Data Mining Approaches. Advances in Science, Technology and Engineering Systems Journal, 2(3):1815–1822, 2017.

[10] Asif Salekin and John Stankovic. Detection of chronic kidney disease and selecting important predictive attributes. In Wai-Tat Fu, Kai Zheng, Larry Hodges, Gregor Stiglic, and Ann Blandford, editors, Proceedings - 2016 IEEE International Conference on Healthcare Informatics, ICHI 2016, Proceedings - 2016 IEEE International Conference on Healthcare Informatics, ICHI 2016, pages 262–270. Institute of Electrical and Electronics Engineers Inc., December 2016. 2016 IEEE International Conference on Healthcare Informatics, ICHI 2016; Conference date: 04-10-2016 Through 07-10-2016.

[11] Siddheshwar Tekale, Pranjal Shingavi, and Sukanya Wandhekar. Prediction of chronic kidney disease using machine learning algorithm. IJARCCE, 7:92–96, 10 2018.

[12] Jing Xiao, Ruifeng Ding, Xiulin Xu, Haochen Guan, Xinhui Feng, Tao Sun, Sibo Zhu, and Zhibin Ye. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. Journal of Translational Medicine, 17:119, 04 2019.