AI-Based Mobile Platform for Predicting The Risk of Having NCDs

Ariane Shimirwa
Information & Communication Technology
Carnegie Mellon University
Kigali, Rwanda
ashimirw@andrew.cmu.edu

Claudine Mahoro
Information & Communication Technology
Carnegie Mellon University
Kigali, Rwanda
claudinm@andrew.cmu.edu

Tyson Muvunyi
Information & Communication Technology
Carnegie Mellon University
Kigali, Rwanda
tmuvunyi@andrew.cmu.edu

Supervised by: Dr. Ahmed Biyabani

***Abstract*— Non-communicable diseases (NCDs) are chronic medical conditions that are not caused by infectious agents and are primarily associated with lifestyle factors. This paper highlights the significant impact of NCDs on global public health, particularly in low- and middle-income countries where they account for most premature deaths. A privacy-preserving AI platform is proposed as a solution to predict the risks of having NCDs, such as cardiovascular diseases and type 2 diabetes, using clinical data and advanced machine learning algorithms such as ANN, SVM, Naive Bayes, and Random Forest. ANN proved to be the best performing for Diabetes and SVM for Heart disease with an accuracy of 92.15% and 92.59 respectively. Utilizing the top-performing models for predictions, the mobile application alerts users who are at high risk and also offers them the opportunity for further medical consultation.**

Keywords—Non-communicable diseases (NCDs), Cardiovascular disease, Type 2 Diabetes disease, AI-Platform

I. INTRODUCTION

Non-communicable diseases (NCDs), also known as chronic diseases, are medical conditions that typically have a long duration, progress slowly, and are not caused by infectious agents. They have an impact on long-term health and frequently necessitate long-term treatment and care [1]. These diseases include a wide range of conditions such as cardiovascular diseases, cancer, chronic respiratory diseases, diabetes, and mental health disorders [1]. NCDs are often associated with lifestyle factors, including unhealthy diets, physical inactivity, tobacco, excess alcohol use, and environmental factors, such as pollution [2]. NCDs have become a significant public health concern in recent years, particularly in low- and middle-income countries. These diseases account for the majority of deaths worldwide, every year, NCDs cause the premature death of 17 million individuals who are below the age of 70. Approximately 86% of these fatalities happen in low and middle-income countries because universal health coverage or access to healthcare services is often limited [3]. Furthermore, NCDs can cause significant disability and reduce the quality of life for those affected, as well as their families and caregivers [4]. Given the significant impact of non-communicable diseases on both individual and societal health, it is important to focus on prevention and management strategies to address this growing health concern.

AI-powered platforms offer the potential to analyze large amounts of clinical data and predict the risk of having NCDs. However, privacy concerns around sharing personal data have slowed the development and implementation of such platforms. This paper proposes the result of a privacy-preserving AI platform for predicting the risk of having non-communicable disease (NCD) such as cardiovascular and type 2 diabetes diseases. The platform is built on a Machine Learning framework and employs supervised learning algorithms to detect hidden patterns and relationships in the data. It uses clinical data and utilize advanced machine learning algorithms such as Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), and Artificial Neural Networks (ANN) to predict the risks of HAVING diabetes or heart disease. The proposed platform has the potential to significantly improve NCD risk prediction while also safeguarding privacy. It will share alerts to individuals with high risks(positive) of having NCDs and they will be provided with an option of scheduling appointments with medical health professionals for further medical consultation.

Assumptions made in this study

This study presumes that the platform will be utilized by individuals previously undiagnosed with heart disease or type 2 diabetes, and who are interested in evaluating their disease risk. We also posit that users are informed about their health metrics and familiar with certain medical terms, allowing effective use of our platform. Moreover, we assume that users are knowledgeable about and provide accurate health information. We further assume that the developed platform can supplement clinical screening and diagnosis, with high-risk individuals identified by the platform seeking subsequent medical consultations.

II. PROBLEM STATEMENT

The prevalence of NCDs continues to be a major global public health concern, particularly in low and middle-income countries. Among NCDs, type II diabetes and cardiovascular diseases, including heart attacks and strokes, are the leading cause of death globally affecting millions of people worldwide. According to the World Health Organization, an estimated 17.9 million deaths occurred due to cardiovascular diseases in 2016, representing 31% of all global deaths [5].

From 1980 to 2014, there was a notable increase in the global increase of type II diabetes, with the number of individuals diagnosed with it rising from 108 million to 422 million. Between 2000 and 2019, there was a 3% increase in the age-standardized mortality rates related to diabetes, with 1.5 million deaths directly attributed to diabetes in 2019 and 48% of these deaths occurred before the age of 70 years. Additionally, the mortality rate attributed to diabetes surged by 13% in lower middle-income countries during the same period [6]. These statistics highlight the urgent need for effective strategies to prevent and manage the risks associated with these diseases. Also, the increasing prevalence and mortality rates of type II diabetes and heart disease emphasize the critical need for early detection and effective management of this disease. If people can detect heart or diabetes disease at an early stage, the number of deaths caused by these conditions could significantly decrease. This study aims at developing an AI-powered platform to help in the early detection of such diseases. That will be an act of raising awareness for people by being able to predict the risks of having heart and diabetes diseases, based on their health and lifestyle information.

III. PROPOSED SOLUTION

The proposed solution presents an innovative approach of using an AI-based platform that will collect important health data from the user and train those data using a machine learning algorithm. The model will analyze those inputs and predict the risk of the user having heart or diabetes disease. The proposed platform is an application that is accessible through a smartphone. The user must enter their health information into the app in order to receive a prediction for the specific disease they are interested in.

If the prediction results present a high risk of having one of the diseases, the system prompts the user to book an appointment and they are given an option to select the desired doctors and hospitals based on their availability at that moment. Since the privacy of the user is important, the platform uses some privacy-preserving techniques. This includes 1) To log in, the user must insert his phone number and it should not be any other thing other than 10 numbers and all numerical. 2) the user is required to enter a strong password that includes both numerical and non-Alphanumeric characters and the password should be longer than 5 characters. 3) The user's password when registering is hashed before being saved in our database. 4) When the user is logging in, their password is obscured, and it can not be visually seen. Overall, this platform offers a convenient and secure solution for individuals looking to manage their health risks proactively.

IV. RELATED WORK

Several researchers have utilized machine learning techniques to predict the risks of getting diabetes and heart disease. Here is a literature review of relevant papers that have employed this technique.

Knam et al. [7] study highlighted the significance of early detection of diabetes and the potential of machine learning techniques in predicting the disease. They employed the Pima Indian Diabetes dataset and developed a support vector machine (SVM) model, which yielded an accuracy of 78.2857%. By utilizing a deep learning model, Artificial Neural Network (ANN) with multiple hidden layers, they were able to increase the accuracy to 88.6%. Pasha et al. [8] used Support Vector Machines (SVM), K-Nearest Neighbor (KNN), and Decision Trees (DT) to predict the accuracy of cardiovascular diseases with precise accuracy and reliability. However, their performance was not satisfactory when executed with large datasets. They improved the accuracy by using Artificial Neural Networks (ANN) and Tensor Flow Keras. Yang et al. [9] developed machine learning models that would support a health practitioner in decision-making to detect multiple chronic conditions. The study used seven machine learning models including k-nearest neighbors (kNN), decision tree (DT), random forest (RF), gradient boosting tree (GBT), logistic regression (LR), support vector machine (SVM), and Naive Bayes (NB) and the results showed that both the RF classifier and the GB classifier achieved high predictive accuracy using AUC score.

Tigga et al. [10] aimed to evaluate the likelihood of diabetes in patients by analyzing their lifestyle, daily habits, health issues, and family background. Different machine learning algorithms were used for predicting the risk of diabetes, and the Random Forest Classifier algorithm showed the most accurate performance on both the experimental dataset of 952 instances and the Pima Indian Diabetes database. Alfian et al. [11] propose a personalized healthcare monitoring system for diabetic patients using a BLE-based sensor device, real-time data processing, and machine learning algorithms. The system aims to help patients better self-manage their chronic condition by monitoring their vital signs data. The system collects important health data of the user such as blood pressure, heart rate, weight, and blood glucose from sensor nodes and transfers it to a smartphone where they will be passed through a Multilayer Perceptron model that will be used to classify diabetes. Therefore, the proposed diabetes classification and BG prediction could be combined with personalized diet and physical activity suggestions to improve patients' health quality and prevent critical conditions in the future. These findings demonstrate the effectiveness of machine learning techniques in diabetes and heart predictions and underscore the importance of early detection for the management of the disease.

V. METHODOLOGY

To develop the proposed AI-powered platform that will help users predict the risks of having heart or diabetes diseases, we used different qualitative methodologies. These include data collection through online available datasets, data pre-processing to prepare the dataset for training machine learning models, exploratory data analysis to explore insights in data, and building models that predict risk. After building different models, the best-performing ones were saved and used within the mobile application built for the users to use in accessing our platform.

I. Data collection

Data used in this study were got from online available datasets. For heart diseases, we used the "Heart_Disease_Prediction" dataset from data world [12].

This was a dataset with 270 entries of which 150 represents the presence and 120 represents the absence of the disease. It had 14 features described in Table 1.

Table1. Heart disease dataset

Feature	Values	Description
Age	29 - 77	Age of patient
Sex	1(Male), 0(female)	Patient's gender
Chest pain type	(1) typical angina, (2) atypical angina,(3)non-anginal pain,(4)asymptomatic	Chest pain caused by blockages in the blood vessels leading to the heart
BP	94 - 200	resting blood pressure (in mm Hg on admission to the hospital)
Cholesterol	126 - 564	serum cholestoral in mg/dl
FBS over 120	1(true),0(false)	fasting blood sugar > 120 mg/dl
EKG results	0(normal),1(having ST-T abnormality),2(showing probable or definiteleft ventricular hypertrophy)	resting electrocardiographic results
Max HR	71 - 202	maximum heart rate achieved
Exercise angina	1(Yes),0(No)	exercise induced angina
ST depression	0 - 6.2	ST depression induced by exercise relative to rest
Slope of ST	1(upsloping),2(flat),3(downsloping)	the slope of the peak exercise ST segment
Number of vessels fluro	0 - 3	number of major vessels (0-3) colored by flourosopy
Thallium	3(normal),6(fixed defect),7(reversable defect)	how well blood flows through the heart muscle while exercising or at rest
Heart disease	1(Yes), 0(No)	Presence or absence of heart disease

For diabetes as well, a dataset named "Early-Stage Diabetes Risk Prediction Dataset" got from Kaggle [13] was used. This dataset originally had 510 entries (which were later reduced to 269 after removing duplicates with the presence of 173 records and the absence of 78) and 17 features described in Table 2.

Table2.Type 2 Diabetes disease dataset

Feature	Values	Description
Age	16 - 90	Patient's age
Gender	1(Male), 0(female)	Patient's gender
Polyuria	1(Yes), 0(No)	Urinating often
Polydipsia	1(Yes), 0(No)	Feeling more thirsty than usual
Sudden weight loss	1(Yes), 0(No)	Losing weight without trying
Weakness	1(Yes), 0(No)	Feeling tired and weak
Polyphagia	1(Yes), 0(No)	Feeling more hunger than usual
Genital thrush	1(Yes), 0(No)	Getting a lot of genital infections
visual blurring	1(Yes), 0(No)	Having blurry vision
Itching	1(Yes), 0(No)	Having itchy skin
Irritability	1(Yes), 0(No)	Feeling irritable or having other mood changes
Delayed healing	1(Yes), 0(No)	Having slow-healing sores
Partial paresis	1(Yes), 0(No)	Numbness or tingling in the hands or feet
Muscle stiffness	1(Yes), 0(No)	A tight feeling in the muscles, accompanied by pain and difficulty moving or limited joint mobility
Alopecia	1(Yes), 0(No)	Loss of patches of hair on the head and other parts of the body

Obesity	1(Yes), 0(No)	Abnormal or excessive fat accumulation
Class	1(Yes), 0(No)	Presence or absence of diabetes

To confirm the adequacy of these datasets, we consulted other literature that considered them as adequate for these prediction studies. For Heart_Disease_Prediction dataset, the same dataset was used in the study entitled “An Augmented Artificial Intelligence Approach for Chronic Diseases Prediction” [14] and for Early-Stage Diabetes Risk Prediction Dataset, it was used in the study entitled “Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques” [15]. From those literatures, we considered these datasets to be adequate for our study as well.

Both datasets were clean, with no missing variables, except the diabetes dataset that contained some duplicates which we got to drop to avoid any effect they might have on our models like causing bias, increasing noise or reduce generalizability or cause overfitting of models.

II. Data pre-processing & Exploratory Data Analysis

After getting data from the datasets, and cleaning them, we explored the patterns behind them, like the statistical distribution of presence and absence of risks, how risks vary according to gender and age, how each feature variable correlate with the presence or absence of risks.

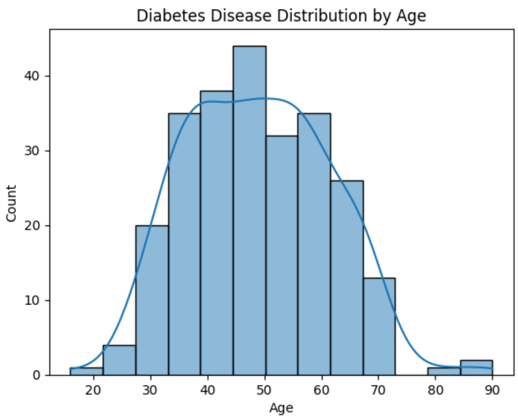

Figure 1: Diabetes distribution by age

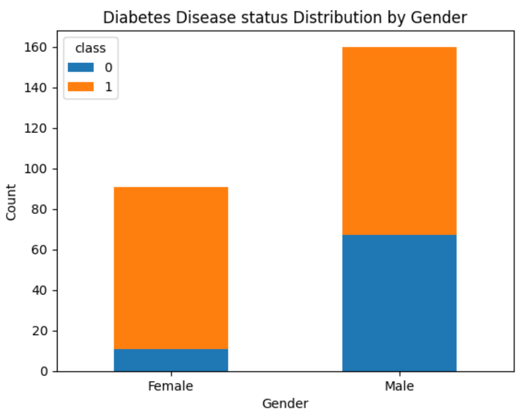

Figure 2: Diabetes distribution by gender

For diabetes, our data shows that males have more risks around the age of 50 compared to females. This complies with other medical documentation like the National library of Medicine [16] and the Centers for Disease Control and prevention [17].

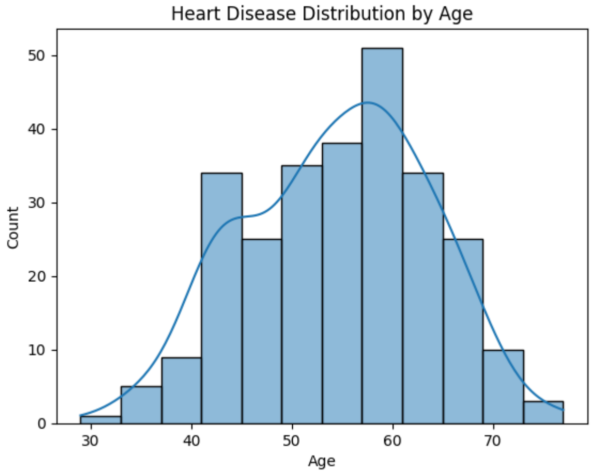

Figure 3: heart disease distribution by age

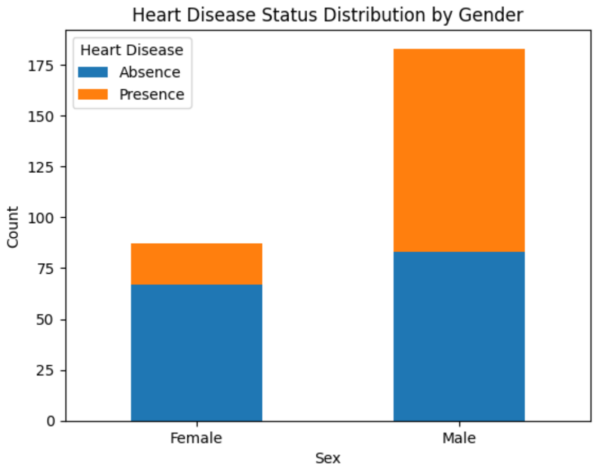

Figure 4: heart disease distribution by gender

For heart disease, our data shows more risks at the age around 60, and in males more than females. According to Medicine in Novel technology and Devices, heart diseases risks are more in men at an early age compared to females whose risks tend to increase at late age of menopause [18], we considered this dataset of ours to be in that same case.

In pre-processing, we also checked for outliers that might affect our modeling and found none. We then moved to encoding our features [19] and selecting which ones to use among the others for our model’s development. For feature selection, we used forward stepwise regression [20] which select significant features depending on their p-value less than the threshold set (0.05 in our case).

For Heart diseases, among the 18 initial features, only 8 were selected.

```
Add Thallium                with p-value 1.57235e-20
Add Number of vessels fluro with p-value 3.2213e-11
Add Exercise angina         with p-value 3.3546e-07
Add Chest pain type         with p-value 0.000104509
Add ST depression           with p-value 0.000261108
Add Sex                     with p-value 0.00799378
Add Max HR                  with p-value 0.0158747
Add EKG results             with p-value 0.025863
```

Figure 5: variables selected for heart disease

And for diabetes, among the 17 initial features, only 6 were selected:

```
Add Polyuria                 with p-value 3.69061e-28
Add Polydipsia               with p-value 3.21959e-11
Add Gender                   with p-value 0.000203876
Add Genital thrush           with p-value 0.000302736
Add Irritability             with p-value 0.00675417
Add Itching                  with p-value 0.00901926
```

Figure 6: variables selected for diabetes disease

III. Modelling

After successfully selecting features to use for modelling, we built 4 different supervised models for each dataset, including Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and Artificial Neural Network (ANN) among others. Before building any of them, we split the datasets into an 80% training set and 20% testing set.

Random forest is an ensemble classifier that combines bagging and random selection of features [21]. It is an ensemble of individual decision trees; each tree predicts a class, and the tree with the highest probability is selected [22]. One of its many advantages is that it does not suffer from the overfitting problem, and it takes the average of all the predictions from every tree, which cancels out the biases. To build the random forest, we first calculated the optimum values for the hyper-parameters we had to tune (n-estimators and max-depth). Support Vector Machine (SVM) is a very useful model in solving classification problems, it is applied in predicting various diseases. It has different kernels including linear, polynomial, sigmoid, and rbf (radial basis function), in this study, we used the linear one as the one that outperformed the others [23]. SVMs are based on the principle of structural risk minimization, which aims to minimize the true error rate [24]. Naïve Bayes (NB) classifier is a probabilistic classification based on applying Bayes' theorem. It assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature; it often works much better in many complex real-world situations than one might expect [25]. Finally, ANN is adopted from the biological complex system, the human brain, which consists of a huge number of highly connected elements called neurons. It tries to find the relationships between input-output data pairs [26].

Each of these models was trained on the training set and tested on the testing set portions of the data. To choose the best model to use, we evaluated each of them using different metrics. Accuracy which measures the proportion of correct predictions made by the model by finding the ration of correct predictions to the total predictions, Precision which measure the ration of true positive predictions to the sum of true positive and false positive predictions, Recall which measures the proportion of true positive predictions over all actual positive instances in the test set, and F1 Score which is the harmonic mean of precision and recall [27].

IV. Mobile application development

We built a mobile application through which our platform will be accessed. The main tools used were flutter for the mobile application and the Postgres database to keep patient records. On the other hand, we used Fast API to create APIs for our backend service. To test the functionality of our APIs, we used Postman as the main tool that allowed us to test the response of our requests before integrating them to the main application. For the functionality part of our system, the user's information that was collected at the frontend side using the mobile application were sent to a backend service as JSON objects and were passed on to the machine learning algorithm which then made predictions based on the information obtained from the user as shown in the architecture figure 7 below.

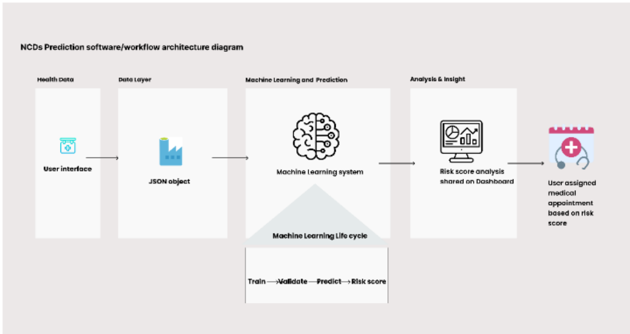

Figure 7: NCD prediction software/workflow architecture diagram

For the use cases of our system, the participating actors are the ones responsible for initiating the system by inserting their diabetes and heart measurements in case they wish to predict their score of acquiring non-communicable diseases. In this case, the user would then be allowed to provide vital information such as the chest pain rate, and their exercise level to mention a few. Additionally, the entire flow of the entire journey would start when the user enters their health metrics into the provided input fields. The system would then validate the provided metrics and in case the user did not provide all the required fields, the system will then deny their submission request until they enter all the required fields. The considered user story was built based on the user requirements which was basically to create a system that would allow them to predict their risk of having non-communicable diseases for them to grow awareness and take precautionary measures if necessary. On the other hand, a patient can immediately book an appointment in case their risk of having Type II diabetes or heart disease is high. The below figures are a sequence diagram for the system to return the NCD prediction scores to the user and scheduling the appointment in case there is a positive prediction.

Figure 8 illustrates how the user makes a request by inserting their metrics into the system, and then the backend service will then validate their inserted information before sending a request to the NCD prediction model. Once the prediction score is obtained, a response will then be sent back to the user in the form of text.

Figure 9 illustrates how the user makes an appointment by selecting the date available on the schedule, then his request is saved into the system's database. The backend service will then validate their request and the system will return the approval of the appointment scheduled to the user.

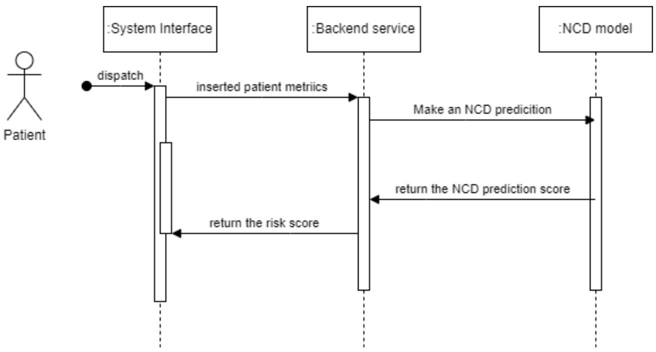

Figure 8: Sequence diagram for making an NCD prediction

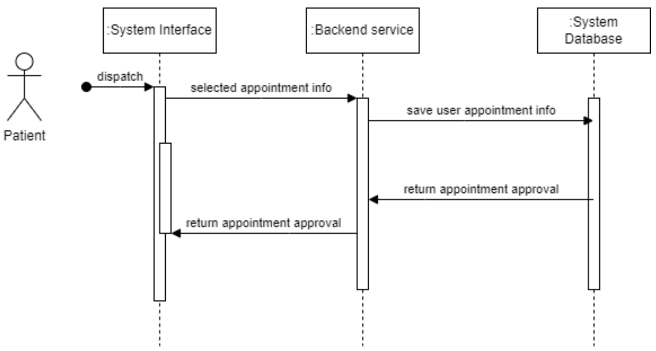

Figure 9: Booking a medical appointment sequence diagram.

We also built a class diagram that represents a high-level set of classes that were present in our implementation and the relationship they had with each other. This can be seen on figure 10.

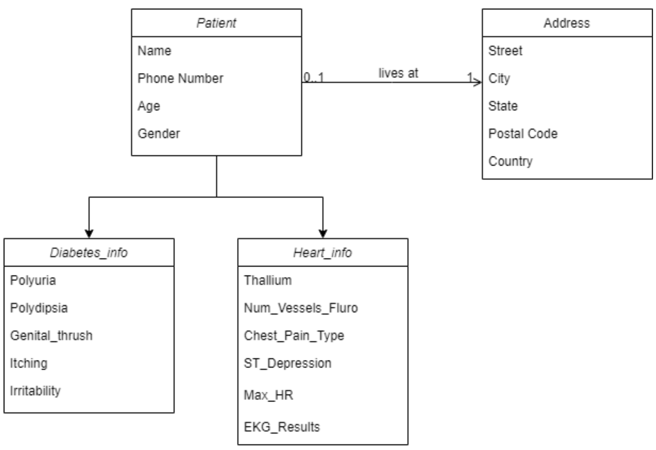

Figure 10: All the participating classes

Figure 10 illustrates the classes we had, there is the patient, address, diabetes info and cardiovascular info as participatory classes for our system. The user information was generally categorized based on their name, phone number, and gender. However, we also discovered a relationship between a user and their diabetes and heart disease records which are all considered while making NCD predictions.

V. Security and Privacy Preserving

For the security part, we have been able to ensure that the privacy of the user is guaranteed by implementing different techniques such as encryption and hashing the user credentials. This is a very crucial and important part because NCD (non-communicable disease) prediction using mobile applications typically collects sensitive personal health information, such as medical history, symptoms, and test results. It is crucial to ensure the security of this information to protect the privacy of users and prevent unauthorized access or disclosure of this information. To safeguard this information, we required users to provide correct authentication credentials to verify their identity. To prevent unauthorized access, we restricted the phone number field to numerical values and a maximum of 10 digits. We also obscured the user's password as they entered it into the password field and mandated the use of a strong password during account creation [28]. Furthermore, we employed a hashing algorithm to encrypt the user password during transmission and storage, adding an extra layer of protection to sensitive information [29].

To maintain privacy within an AI-based platform for predicting the risk of NCDs, several key principles were adhered to. Firstly, data anonymization was employed, removing all personally identifiable information from the collected and stored clinical data to protect individual identities. A robust consent management system was established, giving users full control over what data they shared and the ability to withdraw this consent at any time.

The principle of least privilege was implemented to ensure that users and administrators only had access levels necessary for their roles, minimizing potential data exposure. Transparency and accountability were upheld through the development of a comprehensive policy that detailed how the platform collected, used, and protected data, coupled with clear mechanisms for accountability in case of privacy breaches.

Additionally, detailed audit trails were maintained to track data access and changes, which facilitated the detection and investigation of unauthorized or suspicious activity. Regular security audits and updates were conducted to identify potential vulnerabilities and apply the latest security patches. Compliance with international and local data protection regulations was also ensured, providing an additional layer of security, ensuring privacy, and building user trust.

Overall, these measures demonstrate the commitment to ensuring the security and privacy of our users' personal information.

VI. RESULTS AND DISCUSSION

I. Model's performance

From each section of our study, these are the results we got:

From the exploratory data analysis, the patterns revealed by our data complies with other literatures and it revealed how these NCDs (heart and diabetes disease) affect males and

females differently on different ages. This insight is so significant as it shows which group to emphasize on in raising awareness.
From feature selection, among all the features, we tried different feature selection algorithms and forward stepwise regression turned out to select the most crucial ones among the others. Like for diabetes, it selected features that can help determining risks in anyone regardless to the age range; and since a person of any age can use this platform, excluding the age would make the model less bias. And since the registered users will have provided their age in their personal information, if it is to be used for exploratory analysis again, age would be associated to risks again.

From modeling, as provided in results tables (table 3 & table 4), support vector machine models outperformed others predicting cardiovascular disease with 92.59% accuracy and Artificial Neural Network performed other models in predicting type 2 diabetes with 92.15% accuracy. Therefore, SVM and ANN being the ones saved and used within our mobile application.

Model	Accuracy	Precision	Recall	F1 Score
RF	0.8703	0.9375	0.7142	0.8108
SVM	0.9259	1.0	0.8095	0.8947
NB	0.9074	0.9444	0.8095	0.8717
ANN	0.9074	1.0	0.7619	0.8648

Table 3: heart diseases models' performance

Model	Accuracy	Precision	Recall	F1 Score
RF	0.8823	0.8717	0.9714	0.9189
SVM	0.8823	0.8717	0.9714	0.9189
NB	0.8627	0.9375	0.8571	0.8955
ANN	0.9215	0.9189	0.9714	0.9444

Table 4: Diabetes models' performance

After evaluating all the models, we ended up saving the outperforming one for each disease, and that ended up being SVM model for heart disease and ANN model for type 2 diabetes disease. These two were deployed and used in the mobile application for prediction.

From application development, we successfully built the app with interfaces to take the features used by the models in prediction from the user and use the SVM deployed models in predicting the risk scores and display the results back to the user.

II. Application Interface

Figure 11 and Figure 12 present the interface of the application when the user inserts the records in the available fields for predicting heart disease and diabetes disease.

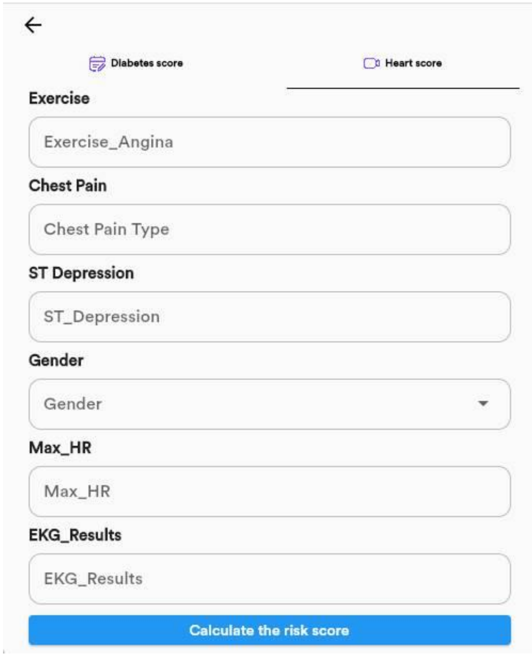

Figure 11: Interface for heart disease risks prediction

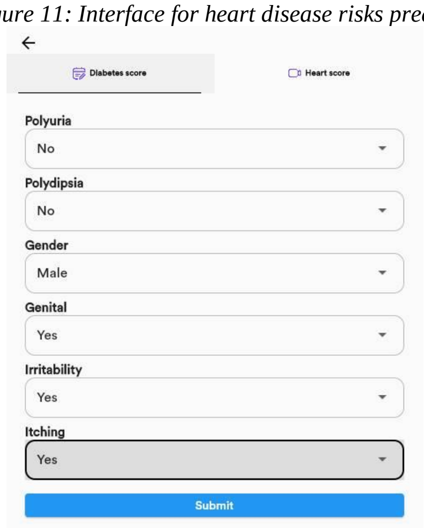

Figure 12: Interface for diabetes risks prediction

After the user has got the prediction results as negative, the system returns a message of low risk indicating negative. The system also encourages the user by message to keep exercising regularly, avoid smoking and eat healthily to prevent acquiring any of the NCDs. if the results come as positive, the systems ask the user to schedule an appointment using our booking system. The Booking system allows a patient to book a medical appointment at a hospital. To do this, a user needs to confirm to immediately make an appointment after their NCD prediction appeared as high risk with their desired hospital or reserve it for a later period. The user will be required to select a hospital, select a service, select their desired doctor, and pick a date and time before approving their appointment. The below figure shows all the fields required before making a medical appointment.

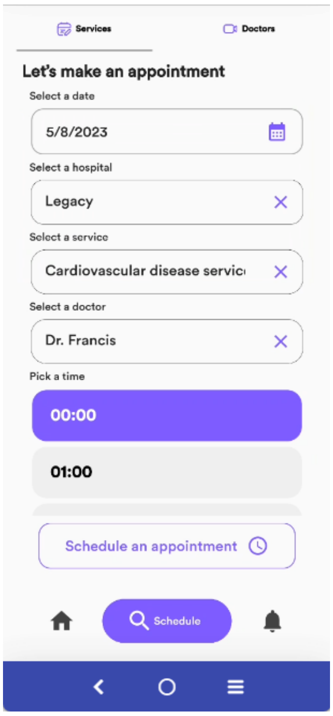

Figure 13: making an appointment by a user

Once the user has scheduled an appointment, he will be taken to another page to see the appointment details, if they are all correct, the user will then click on confirm to confirm the appointment. This is shown on figure 14.

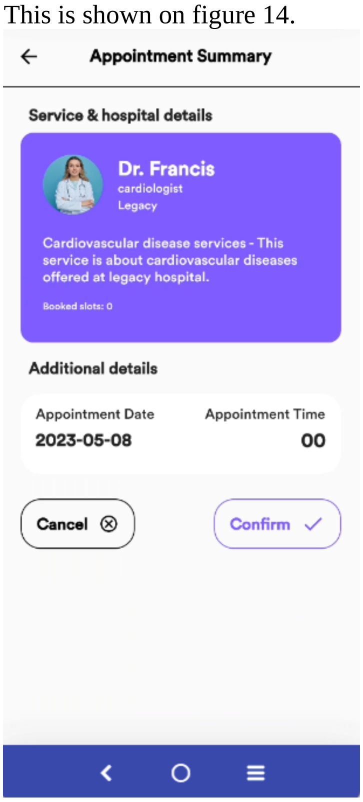

Figure 14: approving appointment info before confirmation

After confirming, the user will immediately receive and be redirected to a confirmation page confirming their medical appointment as seen in the below figure 15.

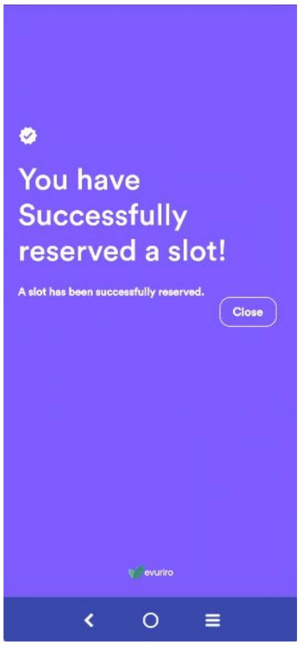

Figure 15: medical appointment confirmation message

III. System testing and security testing

System testing

Our system was tested by creating a user, log in his credentials which is a phone number and a password. After logging in, the user sees the home page. In the home page, there is a button for make predictions. When he clicks to that button it takes him to the home page of the application.

The homepage has many sections including the risk score to evaluate the risk of having NCDs and in our testing, we are interested in making a prediction of having either the heart or diabetes disease.

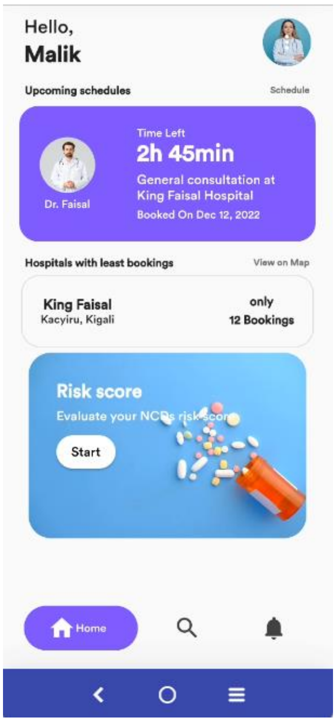

Figure 16: Application dashboard after login

After clicking on the button start shown on figure 16, the user is required to fill all the fields shown on either figure 11 or figure 12 to make predictions. Once he has filled all the required fields and clicked on submit. The system immediately predicts his risks of having either diabetes or heart disease. If the user's prediction indicates a high risk, it implies that they have tested positive for the disease, while a low-risk prediction indicates a negative result, indicating that the user does not have the disease.

Figure 19 and 20 shows the prediction results shown to the user incase of high risk(positive) and low risk(negative).

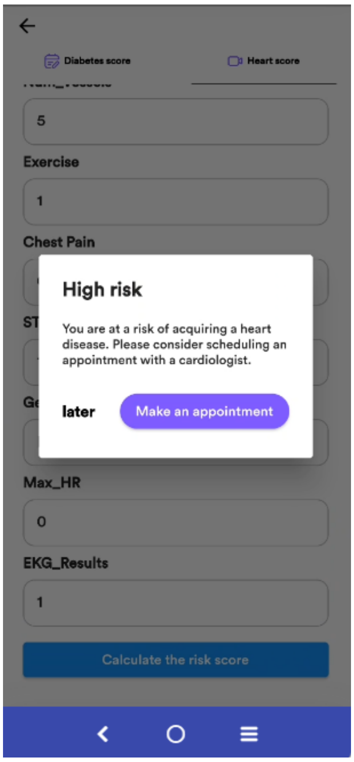

Figure 17: predictions of a user with high risk

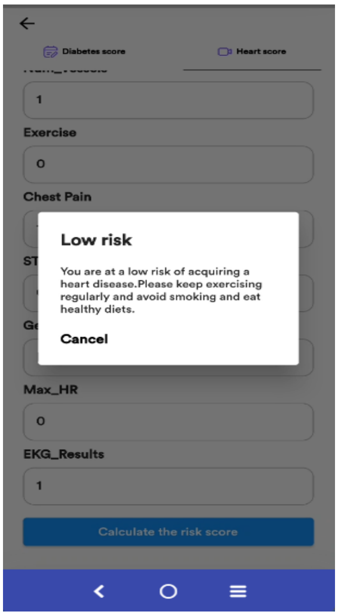

Figure 18: Message when having a low NCD risk

Security testing

After setting the privacy preserving techniques, we have conducted comprehensive testing of our system to verify the effectiveness of the privacy techniques that we have employed. Figure 19 depicts two security measures that have been implemented in our system: limiting the phone number field to accept only ten digits and hiding the password upon entry. These measures serve the purpose of preventing unauthorized access to the system by restricting input to only the required number of digits and safeguarding user passwords from being visible to anyone else. By implementing these measures, we have significantly increased the security of our system and ensured that sensitive information remains confidential.

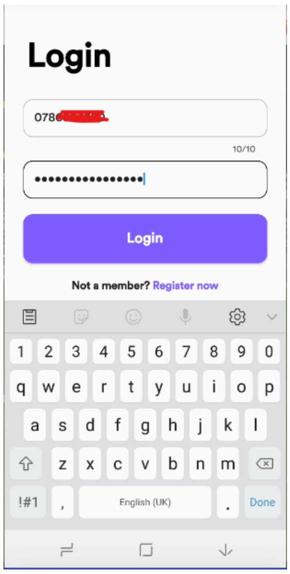

Figure 19: Limiting the phone number to only 10 digits.

Additionally, we have implemented an error message system that alerts users of incorrect login credentials, as demonstrated in Figure 20. This feature helps to safeguard our system by preventing any unauthorized access attempts, maintaining the confidentiality of our clients' data.

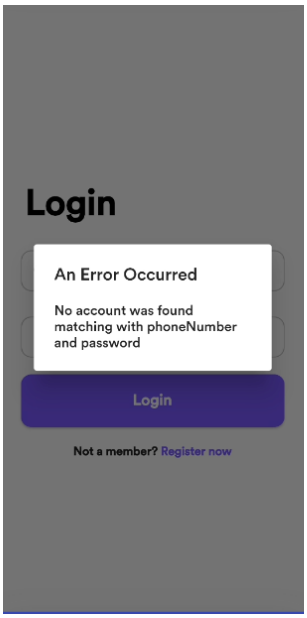

Figure 20: Error message upon incorrect user credentials

To conclude, the mobile app was tested among our colleagues for usability and performance, and this was done by observing how different people were interacting with the application to collect any feedback based on their experience. The feedback obtained was quickly implemented to ease the process of navigation by the user while they are using the application.

VII. CONCLUSION

Detecting diabetes and cardiovascular diseases at an early stage poses a significant challenge in the healthcare industry [30]. In our research, we designed a smartphone application, which can predict cardiovascular disease and diabetes with high accuracy. We used clinical data from an online source, preprocessed it, explored it, used feature selection to select significant variables for prediction, and trained different machine learning algorithms, Random Forest, Support Vector Machine, Naive Bayes, and Artificial Neural Network with the selected features on each dataset. The model's performance was evaluated using classification metrics; accuracy, precision, recall, and F1-score. All models show good results Support Vector Machine and Artificial Neural Network outperformed other models in predicting heart disease and diabetes with an accuracy of 92.59% and 92.15% respectively. The best models were then deployed into the application built using Flutter. The user can use his mobile phone to log into the application and fill in the fields required to predict whether he has diabetes or heart disease. When the predicted results are positive, the system prompts the user to book an appointment at a desired hospital for further medical consultation and when it is negative, he gets a message that encourages him to keep exercising regularly, avoid smoking and eat healthy to prevent acquiring any of the NCDs.

VIII. RECOMMENDATION

Based on the findings presented in the paper, generally, it is recommended that healthcare organizations consider utilizing AI-powered platforms for predicting the risk of non-communicable diseases. The proposed platform, which uses clinical data and advanced machine learning algorithms, offers a promising solution for early disease detection, specifically, for cardiovascular diseases and diabetes.

In this study, there have been limitations such as sampling bias, whereby the selected sample of the dataset might not represent the entire population in reality. Therefore, other researchers are recommended to utilize real clinical datasets, which are derived from data collected within healthcare organizations, to increase the accuracy of prediction. Such datasets can provide a more comprehensive and representative sample of the population being studied, allowing for more accurate and reliable predictions.

Moreover, we recommend investigating the use of wearable sensors to collect patient data, as there are numerous specialized sensors accessible in healthcare settings such as hospitals and pharmacies. For instance, Continuous Glucose Monitoring (CGM) devices are available for individuals with type 2 diabetes to measure their interstitial glucose levels [31]. For heart disease, there are wearable sensors that can be worn on the wrist, such as smartwatches, that can monitor heart activity, blood pressure, heart rate variability, and oxygen levels. Examples of these smartwatches include the Apple Watch, Fitbit Sense, and Garmin Venu 2 [32]. These wearable sensors provide an efficient and convenient way to collect real-time patient data, and this may improve the accuracy of predictions and make it easier for patients to participate in the screening process. Thirdly, we suggest exploring the use of other machine learning algorithms and techniques to improve the accuracy of predictions. Lastly, as future work, we suggest considering other non-communicable diseases such as cancer and other respiratory diseases in the development and implementation of AI-powered platforms and wearable sensor technologies. By expanding the scope of these technologies to include a wider range of diseases, healthcare organizations can further improve disease detection and prevention, ultimately leading to better health outcomes for patients.

REFERENCES

[1] "Noncommunicable Diseases," *PAHO/WHO | Pan American Health Organization*.2023

[2] "Non-communicable diseases," *UNICEF*. [Online]. Available: https://www.unicef.org/health/non-communicable-diseases.

[3] "Non-communicable diseases," *World Health Organization*, September 2022.

[4] J. E. Prynn and H. Kuper, "Perspectives on disability and non-communicable diseases in low- and middle-income countries, with a focus on stroke and dementia," *International*

[5] "Cardiovascular diseases," World Health Organization, September 2022.

[6] "Diabetes," World Health Organization, https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed May 8, 2023).

[7] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432–439, 2021.

[8] S. N. Pasha, D. Ramesh, S. Mohmmad, A. Harshavardhan, and Shabana, "Cardiovascular disease prediction using Deep Learning Techniques," *IOP Conference Series: Materials Science and Engineering*, vol. 981, no. 2, p. 022006, 2020.

[9] J. Yang et al., " Prediction for the Risk of Multiple Chronic Conditions Among Working Population in the United States With Machine Learning Models", IEEE Open Journal of Engineering in Medicine and Biology, 2: 291–298, 2021.

[10] N. P. Tigga and S. Garg, "Prediction of type 2 diabetes using machine learning classification methods," *Procedia Computer Science*, vol. 167, pp. 706–716, 2020.

[11] G. Alfian, M. Syafrudin, M. Ijaz, M. Syaekhoni, N. Fitriyani, and J. Rhee, "A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing," *Sensors*, vol. 18, no. 7, p. 2183, 2018.

[12] "Heart disease prediction - dataset by Informatics-Edu," data.world, https://data.world/informatics-edu/heart-disease-prediction

[13] I. Dutta, "Kaggle," [Online]. Available: https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset.

[14] J. Rashid *et al.*, "An augmented artificial intelligence approach for chronic diseases prediction," *Frontiers in Public Health*, vol. 10, 2022. doi:10.3389/fpubh.2022.860396

[15] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," SpringerLink, 2019.

[16] [1] M. A. Siddiqui, M. F. Khan, and T. E. Carline, "Gender differences in living with diabetes mellitus," Materia socio-medica, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3769156/ (accessed May 8, 2023).

[17] "Centers for Disease Control and Prevention," cdc, 30 09 2022.https://www.cdc.gov/diabetes/data/statistics-report/diagnosed-undiagnosed-diabetes.html

[18] Z. C. A. S. X. D Zujie Gao, "Gender differences in cardiovascular disease," Medicine in Novel Technology and Devices, vol. 4, no. ISSN 2590-0935, 2019.

[19] S.-l. developers, "Scikit," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html.

[20] S. Kaushik, “Introduction to feature selection methods with an example (or how to select the right variables?),” Analytics Vidhya, 2016.

[21] B. L. D. P. C. M. A. Jabbar, "Prediction of Heart Disease Using Random Forest and Feature Subset Selection," in *Innovations in Bio-Inspired Computing and Applications*, Springer Link, 2015, pp. 187 - 196.

[22] N. D. O. A. Peter D. Caie, "Precision medicine in digital pathology via image analysis and machine learning," in *Artificial Intelligence and Deep Learning in Pathology*, Elsevier, 2021, pp. 149-173.

[23] N. M. a. V. Jain, "Performance Analysis of Support Vector Machine in Diabetes Prediction," 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), India, 2020.

[24] N. &. B. A. Barakat, "Rule Extraction from Support Vector Machines: A Sequential Covering Approach," *Knowledge and Data Engineering*, vol. 19, no. 10.1109/TKDE.2007.19610, pp. 729-741, 2007.

[25] S. A. P. a. A. Parveen, "PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES," *International Journal of Advanced Computer and Mathematical Sciences*, vol. 3, no. 3,2012, pp. 290-294, 2012.

[26] H. Turabieh, "A Hybrid ANN-GWO Algorithm for Prediction of Heart Disease," *American Journal of Operations Research*, vol. 6, no. 2, 2016.

[27] [1] H. M and S. M.N, “A review on evaluation metrics for Data Classification Evaluations,” *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, 2015. doi:10.5121/ijdkp.2015.5201.

[28] Microsoft. (n.d.). What is Password Protection? [Online]. Available: https://www.microsoft.com/en-us/security/business/security-101/what-is-password-protection.

[29] Devglan. (n.d.). Bcrypt Hash Generator - Online Tool. [Online]. Available: https://www.devglan.com/online-tools/bcrypt-hash-generator.

[30] Health Care Quality Indicators - cardiovascular disease and diabetes - OECD, https://www.oecd.org/els/health-systems/hcqi-cardiovascular-disease-and-diabetes.html.

[31] A. J. Modaragamage, “Top wearable medical devices used in Healthcare | HealthNews,” *Health news*, Mar-2023.

[32] C. Vaughn, “10 different sensors you'll find in your smartwatch,” *MUO*, 08-Dec-2022.