Predictive Medical Diagnostic Tool for Electronic Health Records (EHR)

Robert A. Sowah[1*], Appah Bremang[1], Joshua Ocrah[1], Emmanuel Ashie[1], Faisal Alhassan[1] and Chrispine Songkyebe Domo [1]

[1] Department of Computer Engineering, University of Ghana, Legon, Ghana
***Corresponding author:** rasowah@ug.edu.gh

ABSTRACT

Healthcare management cannot be undertaken efficiently unless the necessary required data is available. It is essential to ensure that the most appropriate and accurate information for the management of patient health is continuously collected, processed, and provided in a required format and timely. Many avoidable shortcomings in the health sector that result in poor quality are due to the inaccessibility of data and mining necessary knowledge from it. The development and establishment of a robust health information system are essential for the sustainable management of patient healthcare records. This research work was carried out to develop a centralized database for the management of patient health records. This paper seeks to eliminate the traditional paper-based approach of keeping records in various hospitals of the health sector and also to create a central platform where patient health records can be accessed across multiple hospitals that run this system. The web application developed allows for the registration of patients, vital signs entry, doctor diagnoses, laboratory results upload, drug administration as well as service payment. The incorporation of seven (7) different machine learning algorithms in the web application assists the doctor in analyzing historical data on patients as well as their laboratory test results accurately. The mobile app also operates like the web application with an additional chat functionality that makes it possible for the patient to have remote access to medical help by initiating a chat session with a doctor without being present at the hospital.

Keywords: patient health records, machine learning, historical data, central platform, remote access.

I. Introduction

An Electronic Health Record (EHR) is a concept defined as an orderly way of collecting electronic health information about individual patients [1]. It is the systematized collection of patient and population electronically-stored health information in a digital format. These records can be shared across different health care settings. Records are shared through network-connected, enterprise-wide information systems, or other information networks and exchanges [1], [2]. An electronic health record (EHR) is a digital version of a patient's paper chart. EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users. While an EHR does contain the medical and treatment histories of patients, an EHR system is built to go beyond standard clinical data collected in a provider's office and can be inclusive of a broader view of a patient's care. EHRs are a vital part of health IT and can: (1) contain a patient's medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results (2) allow access to evidence-based tools that providers can use to make decisions about a patient's care and (3) automate and streamline provider workflow [2].

One of the key features of an EHR is that health information can be created and managed by authorized providers in a digital format capable of being shared with other providers across more than one health care organization. EHRs are built to share information with other health care providers and organizations – such as laboratories, specialists, medical imaging facilities, pharmacies, emergency facilities, and school and workplace clinics –

so they contain information from all clinicians involved in a patient's care. This shared information includes a wide range of data like medical history, medication and allergies, immunization status, laboratory test results, vital signs, personal statistics like name, age and weight, and others. In many cases, this sharing can occur by way of network-connected enterprise-wide information systems and other information networks or exchanges. The health care community generally agrees that improved use of accurate, current, and clearly understood health information is essential to deliver high-quality, cost-effective health care.

EHR thus makes it possible for paradigms like machine learning (ML) to be applied to the records to help health practitioners in medical diagnoses. That is essential because, in the United States, for example, medical error is the third leading cause of death [3]compiled by the Centers for Disease Control and Prevention (CDC. The world's leading cause of death over the past few years has been heart disease followed by cancer. Chronic diseases, in general, have assumed a more dominant role in the 20th century as the leading cause of death in the world [4]; hence, some chronic diseases were selected for the application.

With the implementation of EHR in health care management, large volumes of data are generated and stored. These stored data, if mined properly using sophisticated algorithms from machine learning, can provide essential information for patients' care and treatment and assist doctors and specialists in healthcare service delivery and management. What is the scope of this problem?

This paper presents a predictive diagnostic tool for EHR using machine learning algorithms such as decision trees, naïve Bayes classifiers, support vector machines, and others.

Electronic Health Records Management systems are not new concepts. There are many implementations of such systems, but the application of machine learning algorithms on stored medical data to gain extract, mine data, and gain insight and help in patients' diagnosis and care is new and very vital.

This paper is structured into five sections. Section 1 introduces the work and provides background knowledge for Electronic Health Records. Section 2 describes the existing works and systems related to the proposed system of designing, developing, and deploying an Electronic Health Records Management systems with machine learning capability as a diagnostic tool. Section 3 describes the entire system design methodology processes that were employed. The breakdown of the whole system into various modules and the methods to implement them are presented. Section 4 describes the implementation and testing results with discussions. The implementation of the design and development are validated for the software systems and their integration. Section 5 concludes the work, discusses the limitations, and gives an outlook on future work needed.

II. Literature Review

A. Related Works

Researchers in [5] designed, develop and deploy a system that manages the administrative, financial, and clinical aspects of a hospital. This system, developed with Borland C++ builder V6.0 and MS SQL Server 2005 helps patients to figure out what to do and where to go to, through a software interface. At the doctor's end, it can suggest what kind of laboratory test the patient should undergo which is subject to changes by a doctor who is not in agreement. The software also gives suggestions as to the diagnosis. The paper does not illustrate the methodology involved in making such a system, but from the conclusion, it can be assumed to handle a wide range of laboratory tests, help patients who do not know the hospital structure, and assist in the financial accounting of the hospital. However, it makes no mention of the use of the patient's vital signs to make an inference as to which of them is in a relatively more critical condition than the other.

In [6], a specific case of building a Hospital-Based Database system for a hospital in Bangladesh that stores almost all the possible activities that can occur in the hospital is shown. It incorporates the use of the Entity-Relationship Model (ERM), which involves cardinalities, ternary relationships, which are later transformed into a relational model. The paper gives extensive information into building robust hospital management systems that have almost all possible scenarios addressed. It does not, however, apply any vital signs monitoring, financial management, or machine learning algorithms to the data.

The use of machine learning in medical diagnostics is not a new field as it has been used in a lot of research works. The technical advances that have taken place in the healthcare sector have made available large databases of biological and chemical information [7]. With machine learning, the presence or absence of chronic heart disease can be predicted using a predictor system whose biomarker data can be adopted by healthcare administrators for better service, as seen in Nikhar and Karandikar's proposed system. The proposed system compares the accuracy of prediction of heart disease when Naive Bayes and Decision Tree classifiers were applied and concluded that, Decision Tree outperforms Naive Bayes in predicting the occurrence of the disease. Feature extraction was also proposed as a means of improving accuracy when unnecessary and irrelevant attributes are excluded from the dataset [8].

Artificial Neural Network (ANN) can also be used to classify and predict heart disease due to its ability to make meaning out of complicated or imprecise data which could be meaningful in extracting unique features or patterns and detect trends that are more complex to be noticed by humans or other computer methods as elaborated by Sonawane and Patil (2014). Here, a particular case of a competitive network, Learning Vector Quantization (LVQ), was used to create prototypes that can easily be identified and interpreted by experts in the respective application domain. LVQ, when compared to other algorithms used in this work, gave the highest accuracy of 85.55%. This accuracy though good can be increased by employing other neural network techniques since its application in the health field requires little or no room for errors [9].

An improved neural network algorithm, Multilayer Perceptron Neural Network (MLPNN) with Backpropagation algorithm (BP) used by Dangare and Apte (2012) gave a higher accuracy for the algorithms used. A data mining tool, WEKA 3.6.6, was used to experiment with the system. The system experimented with 13 and 15 heart disease attributes, which gave an accuracy of 99.25% and 100%, respectively [10], [11]. The accuracy, even though high and desired, might not be realistic as the proposed system does not check for over-fitting of data, which may result in inaccurate prediction when tested with new data.

The risk of different chronic diseases can be predicted using machine learning, as seen in Khalilia, Chakraborty, and Popescu (2011). The National Inpatient Sample (NIS) data was used to train a random forest classifier for disease prediction. The data used was imbalanced; hence, an ensemble learning algorithm based on repeated sub-sampling was employed to subdivide the data into samples while ensuring each sample is well balanced. The random forest classifier outperformed Support Vector Machine (SVM) and bagging and boosting classifiers with 88.79% accuracy when their accuracies were compared [12]. This system outlined has incorporated several other small systems hence making it useful in several applications.

In [13], the Center for Medical Services (CMS) data set is used to classify and predict 11 chronic diseases, including kidney disease, osteoporosis, heart disease, etc. This paper used the machine learning tool to reduce the number of diagnostic codes for chronic diseases, thereby reducing the health care cost for the beneficiary. Instead of studying each chronic disease separately, [13] proposed a predictive system that can be used to analyze several chronic diseases at a time. This system is similar to the one proposed by [12], but it predicts a relatively more significant number of chronic diseases. The significant difference in the accuracy of the training and testing models on each chronic disease data means there may be inconsistency in the prediction even though all the chronic diseases for training and testing had accuracy above 70%.

The paper-based approach is the traditional way of gathering and obtaining the patient's health information. Csiszar (2011), as cited by [14], submitted that medical institutions would still instead use paper to gather information from their patients and also to record surgical procedures, observations, and prescriptions. Some older medical practitioners and physicians who are not technologically-savvy find accessing digital records somewhat complicated than obtaining data on paper [12]. That situation is changing faster due to technological advancements. Manual record-keeping is very exhausting, and this may be due to the mere undeniable fact that every day, many new records are being recorded and stored in hospitals. It will be complicated to sort medical records of all patients that keep increasing every minute. The paperwork complexity often arises from errors that will considerably get new daily happenings in hospitals, clinics, and all sorts of other healthcare institutions. Apart from being time-consuming, collating and retrieving records can be difficult due to a large volume of paperwork.

Also, a large volume of space is needed to keep many paper health records. A "paperless" environment is useful because of proved medical documentation, increasing staff and instrumentation efficiency, reduction in overhead and the growth in practice, eliminating recordkeeping space, enhancing the standard of care, accumulating information for managed care contracting, standardizing an information platform for a physician group. The essential characteristics of a central database are particularly appealing to the needs of health care providers and medical practitioners. These characteristics include on-demand self-service, easy access to patient medical records everywhere, resource pooling, rapid elasticity, and measured service [14].

It is difficult finding a common platform where a patient's records will be kept for ease of accessibility in every hospital visited by the patient. That is because different hospitals and clinics use different software systems for patients' data capture and storage. These disparate heterogeneous systems do hinder the kind of seamless integration of patients' data required for proper diagnosis and healthcare service delivery. Some hospitals or clinics visited by patients do not have the appropriate electronic health records management system. This problem is further compounded by a lack of adequate training and expertise in the use of a digital record system for quick inference. Based on these issues raised, a central platform where patients' medicals records will be kept and accessed remotely by the localized servers in every hospital where our developed system will be installed is needed. The developed system is user-friendly, intuitive and easy to use with low training times.

In providing ubiquitous access to health care delivery to patients, mobile application development has taken the lead. The health sector has recently experienced a sweeping influence from the world of mobile applications. A variety of mobile applications ranging from appointment applications, medication alertness applications, medical consultation applications, and amongst others have been developed to ease the workload on health practitioners and make good use of technology. Notable amongst these new applications is the First Responder Help System. That is a system that allows mobile users to have access to health care services in case of emergencies. The cloud hosts a directory service containing a list of pre-registered professionals, such as doctors and nurses. Once a registered user sends a help message to the First Responder cloud service, a lookup operation is started to search for the most suitable medical professional, considering the location of both users. Then, the cloud service will establish/monitor a communication link between the person requesting aid and the selected professional, whereby these two users can start a chat session to exchange messages. For enhancing system availability and dealing with connection issues, SMS is provided as an alternative communication method [15]. The First Responder Help System does not populate medical diagnosis on Patients' mobile devices to assist medical practitioners in diagnosing ailments accurately, rendering some prescriptions from registered health practitioners not very accurate since modern medical diagnostics are dependent on past diagnosis. The Mobile

Medical Expert System (mMES) has a knowledge base of diagnosis, advice, and treatment of minor sicknesses such as headaches, stomach pains, and amongst others. Patients with illnesses such as minor headaches, minor stomach aches, and minor malaria need not report to a major referral Hospital such as Korle-Bu Teaching Hospital (KBTH) to release some amount of work pressure on Medical Doctors. Patients would initially register as a user through an interface and interact with the Mobile Medical Expert System and the Medical Doctor through cloud computing, mobile technology, and devices. The Medical Doctor will advise the patient through their mobile devices according to his/her interaction with the Mobile Medical Expert System[16]Komfo Anokye Teaching Hospital (KATH. The Mobile Medical Expert system for health Institutions in Ghana is indeed a change in the positive direction to aid in eradicating overcrowding in the critical health sectors in Ghana. The system stands a chance of misleading patients with the wrong diagnosis since the surety of diagnostics is never 100 percent. Patient symptoms in some cases require various laboratory tests to aid in accurate diagnostics. The Health Care application makes provisions that ensure that only registered medical practitioners execute multiple prescriptions in the medical body as against the Mobile Medical Expert System that does not make any requirement to monitor the health practitioners who perform Doctor's prescriptions.

Similar to the Mobile Expert System is the iCare android application. This application accepts the symptoms from the patients, processes the data, identifies the particular disease, and hence provides appropriate medication using a medical expert system by pattern matching techniques. The data collected from the device is evaluated using an expert system that estimates the probability of the severity of the disease. The more the patient uses the application, the better the expert system becomes in identifying various patterns of illnesses that increase the accuracy of iCare [17].

B. Proposed Work

Our system is centered on assisting doctors in medical diagnoses through the application of machine learning algorithms on stored medical data. The system also plays a role in enabling communication of medical data between various departments of a hospital. The system aims to aid in eliminating old-fashioned ways of keeping patient's data such as manual record-keeping to digitized medical health record-keeping and provide a diagnostic tool using different machine learning algorithms to help doctors in decision-making. The system accumulates patients' medical records of registered hospitals into a centralized database. The system has an android application that provides first aid measures, allows real-time consultation between patients and doctors via mobile chat platforms, and helps in improving the response time of ambulances by making service contacts of various ambulance readily available to mobile users.

III. SYSTEM DESIGN AND METHODOLOGY

Figure 1 shows a block diagram of the various components of the system and how to work with each other.

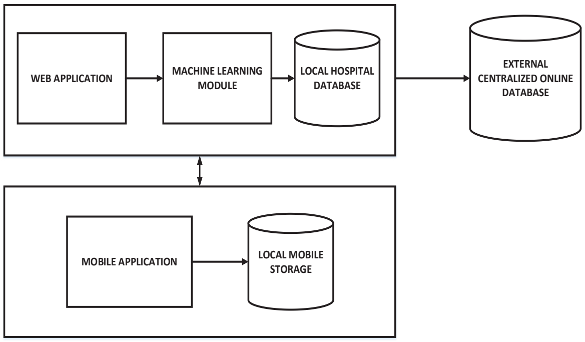


Figure 1: Block Diagram of System

A. Development of Web Application

To develop the web application, a web environment powered by Python was employed, and this came in the form of the Django framework, a Python web framework. A well-defined sub-application was built for each department of the hospital. The departments taken into consideration are, namely: Records, Reception, Laboratory, Accounts, Doctor, and Pharmacy. The principal parts of the web application are highlighted below. A flow diagram representing the sequence of activities is shown in Figure 2.

1. Department-based sub-applications - The records sub-application deals with any vital information relating to the patient who comes to the hospital due to any medical-related issue. It enables the necessary data of the patient like name, id, residence, etc. to be taken and also gives a general idea of the number of patients in line to be served for a particular day.

For the receptionist or nurse department, the vital signs of the patient are taken by a nurse and entered through the application to be stored and referenced later by the attending doctor to aid in diagnosis.

At the laboratory, a patient who has a specific lab test recommended by a doctor is tested, and a document containing the lab results is uploaded by the laboratory scientist and stored, also to be used by an attending doctor for the diagnosis.

The sub-application for the accounts department is developed to contain a form for confirming the payment of a customer for a particular service ranging from a laboratory test to acquiring a drug prescription.

For the doctor, the application first provides a list of patients whose recorded vitals and symptoms have a severity of low and medium as denoted by the nurse on the vitals form. Another list containing patients with high severity cases is provided. When a patient in the list is chosen, the full profile of the patient consisting of patient name, age, residence, current and previous vitals (if any), any readily available laboratory test results, machine learning prediction section and forms for entering the diagnosis, recommending laboratory tests and giving prescriptions.

The sub-application for the pharmacist department serves as a way for prescriptions given by the doctor to be served to the patient.

In addition to the department sub-applications, a centralized database application was created to facilitate the movement of data from the local hospital servers to a centralized database using Python Dropbox Application Programming Interface (API).

2. Internetworking of Hospital Department-based Application - For inter-hospital communication, a star topology is used where every localized database server within each hospital is connected to an online centralized database server. That provides easy movement, isolation, or interconnection with other networks. That is, the network is scalable. There are several advantages associated with the star topology, namely: a good option for new networks, low start-up costs, easy to manage, offers opportunities for expansion. An external IP address provided by the Internet Service Provider (ISP) is used to connect to the internet. Therefore each hospital connects to the centralized database server using an external IP address.

3. Storage of Data - Centralized database is a database in which data is stored and maintained in a single location. That is the traditional approach of storing data in enterprises. In a centralized database, all the data of an organization are stored in a unique place such as a mainframe computer or a server. Users in remote locations access the data through the Wide Area Network (WAN) using the application program provided to access the data. Since all the data reside in a single place, it is easier to maintain data integrity and back up data. The purpose of this centralized database is to link two or more hospitals to a central server where patient records can be obtained at different times within different hospitals. The centralized database server was implemented using the Python Dropbox Application programming interface (API).

Dropbox is a cloud storage service sometimes referred to as an online backup that is commonly used for file sharing and collaboration. That is, a user's files are available from any computer that has an internet connection connected to one's Dropbox account. Dropbox has both the capability of storing data and allowing the application to retrieve existing data from it. Files are uploaded into the Dropbox using either drag-and-drop method or writing to a file and then storing the file in the Dropbox using the file path.

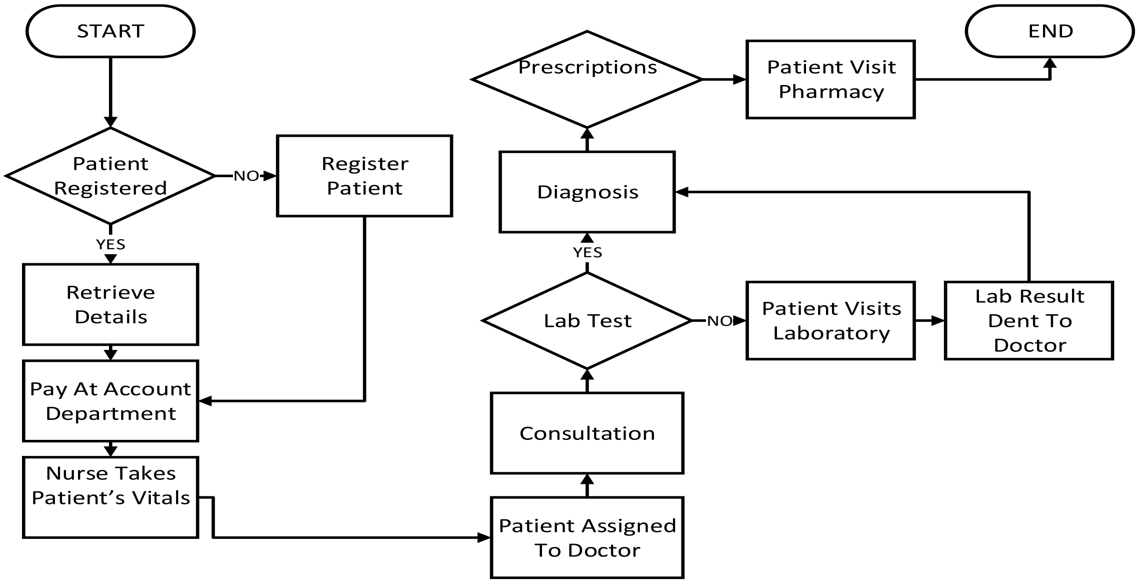


Figure 2: Flow Diagram for Web Application

B. Machine Learning Module

1. Acquiring Data - To assist health personnel with a predictive tool to aid in diagnostics, the first activity carried out was the collection and analysis of sample historical data sets. The datasets were obtained from the University of California, Irvine (UCI) containing diagnostics of patients for three different diseases, namely Breast cancer, Heart disease, and Diabetes. The demographics are white Caucasian, but it would be useful to adopt Ghanaian health data, which was not available as of writing this paper.

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms [18]. This archive was created as an FTP archive in 1987 by David Aha and fellow graduate students at UC Irvine. Since that time, it has been widely used by students, educators, and researchers all over the world as a primary source of machine learning data sets.

Currently, UCI maintains 373 data sets as a service to the machine learning community. The following are descriptions of the data sets obtained from the UCI Repository [19]:

The dataset was obtained from the Cleveland database. This data set is multivariate, with 14 distinct attributes made up of categorical, integer, and real characteristics. A total number of 303 instances are available for use, with some values missing values present. Machine learning with the Cleveland database has concentrated on merely attempting to distinguish presence (values 1, 2, 3, 4) from absence (value 0) of the heart disease condition.

The attributes used include age, sex, chest pain Type, resting blood pressure, serum cholesterol, fasting blood sugar > 120, resting electrocardiograph, maximum heart rate, exercise-induced angina, old peak slope, number of major vessels, and diagnosis of heart disease.

The breast cancer data set is obtained from the Wisconsin Breast Cancer Database [19]. That is also a multivariate dataset with 11 attributes made up of integer values. The number of instances presented here is 699, with the presence of missing values. Samples of breast cancer dataset arrive periodically, as Dr. Wolberg reports his clinical cases. The attributes used include sample code number, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, and class.

The Pima Indians Dataset is originally owned by the National Institute of Diabetes and Digestive and Kidney Diseases. In particular, all patients here are females at least 21 years old of Pima Indian heritage. This dataset is also multivariate, with eight (8) distinct attributes made up of integer and real characteristics. A total number of 768 instances are available for use, with some values missing. The attributes used include the number of times pregnant, plasma glucose concentration, diastolic blood pressure, resting blood pressure, triceps skinfold thickness, 2-hour serum insulin (mu u/ml), body mass index, diabetes pedigree function, age, and class.

2. Training Machine Learning Models with Data - The associated task ran on the dataset obtained is classification. Here, the system is trained to know which of a set of categories or sub-populations a new observation belongs, by a training set of data containing observations or instances whose category membership is known.

The data sets for the three diseases are first of all split into input and output attributes where the input attributes are made up of all the attributes except the last attribute, which is classified as the output attribute. Since the data sets contain missing values which may interfere with the machine learning algorithm, the *Imputer* library from *sci-kit-learn* was used to transform the input attributes with the missing values. Out of the instances available from the set, 67% were used for training various models, while 33% used for testing the models.

To avoid over-fitting and under-fitting of the models, feature extraction was applied where features that do not have a direct influence on the training of the system were ignored.

Seven (7) classifiers, namely; *Logistic Regression*, *Linear Discriminant Analysis*, *K-Neighbours Classifier*, *Artificial Neural networks*, *Decision Tree Classifier*, *Gaussian Naive Bayes*, and *Support Vector Machine* trained. Each classifier underwent 10-folds cross-validation and the accuracy metric recorded for each of them. The best classifier, being the one with the highest accuracy, was selected and used to fit and score the training and test sets. Summary of the classifiers used in the machine learning model for the diagnosis in HER is given below:

1. Logistic Regression Analysis

Logistic regression is a supervised machine-learning algorithm used to classify data. Simple logistic regression is less computationally intensive and relatively easy to implement. Logistic regression is a variation of ordinary regression which deals with finding a function that relates a continuous outcome variable Y to one or more independent variables $X_1, X_2, \ldots X_n$ [20]. Simple linear regression assumes a function of the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \qquad (1)$$

and finds the values of $\beta_0, \beta_1, \beta_2, \ldots \beta_n$, the regression coefficients for the corresponding independent variables $X_1, X_2, \ldots X_n$ and (β_0 is called the intercept.)

Logistic regression is useful when the observed outcome is restricted to two values (usually coded as 1 or 0, respectively). It produces a formula that predicts the probability of the occurrence as a function of the independent variables [20].

2. Linear Discriminant Analysis

Linear Discriminant Analysis is a dimensionality reduction technique used for supervised classification problems. It is used for modeling differences in groups, i.e., separating two or more classes. It is used to project the features in higher dimension space unto a lower dimension space. It is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition, and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events[21]–[23]. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before classification. LDA is closely related to the analysis of variance (ANOVA) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements [21]–[23]. Two criteria are used by LDA to create a new axis, namely: (1) maximize the distance between means of the two classes and (2) minimize the variation within each class.

3. K-Nearest Neighbours Classifier

In pattern recognition, the k-nearest neighbours' algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression: (1) with k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the objective being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is assigned to the class of that single nearest neighbor, (2) with k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors [24]–[26]. k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally, and all computation is deferred until classification. For classification and regression tasks, a useful technique can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones [24]–[26].

4. Artificial Neural networks

A biological paradigm inspires the concept of Artificial Neural Networks and hence attempts to model the information processing capabilities of the nervous system. Neurons receive signals from the environment and produce a response. The general structure of a biological neuron is shown in Figure 3. The communication among the neurons is done by electrical signals. Each neuron receives thousands of connections from other neurons and, therefore, continually receives incoming signals and responds accordingly. The neuron only generates a response if the incoming signal received by the neurons exceeds some threshold value [27], [28].

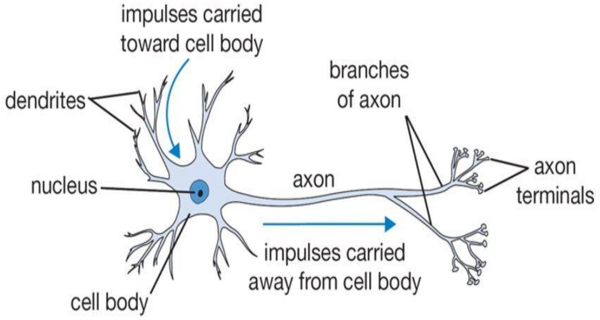

impulses carried toward cell body
branches of axon
dendrites
axon terminals
nucleus
axon
impulses carried away from cell body
cell body

Figure 3: Biological neuron

Figure 3 shows an example of a simple artificial neuron. The node is the equivalent of the biological neuron. Weights model synapses between neurons. Each input is multiplied by a weight and sent to the node. The weighted inputs are then summed to supply a node activation. The activation is compared with a threshold; if it exceeds, the node produces a high-valued output or produces low-valued output otherwise. An Artificial Neural Network can, therefore, be described as an interconnected assembly of simple processing elements, units, or nodes whose functionality is loosely based on the animal neuron [27], [28]. The processing ability of the network is stored in the inter-unit connection strength or weights obtained by the process of adaptation to or learning from a set of training patterns, as shown below in Figure 4.

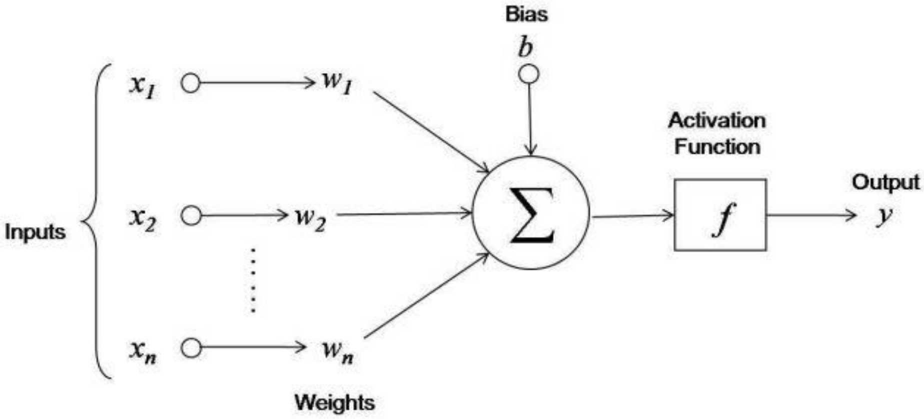


Figure 4: An artificial neuron

5. Decision Tree Classifier

In statistics, Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics, data mining, and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels, and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees [29]–[32].

With decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data (but the resulting classification tree can be an input for decision making).

6. Gaussian Naive Bayes

In machine learning, naïve Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian network models. It was introduced into the text retrieval community and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features [33], [34]. With appropriate pre-processing, it is competitive in this domain with more advanced methods, including support vector machines. It also finds application in automatic medical diagnosis. Naïve Bayes classifiers are highly scalable, requiring several parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. When the underlying data is continuous, independent, and identically distributed, the Gaussian Naïve Bayes algorithm is then used for maximum classification accuracy [33]–[36].

7. Support Vector Machines

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane, which categorizes new examples. In two-dimensional space, this hyperplane is a line dividing a plane into two parts, where each class lay on either side of the hyperplane.

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New models are then mapped into that same space and predicted to belong to a class based on the side of the gap on which they fall [37]–[40].

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

3. Testing the Model with Highest Accuracy through Web Application - The machine learning prediction section, which depends on if a laboratory test result is ready, is done through an AJAX POST call whose functionality is given by the predict function. This function uses the patient ID and test ID to query the database for the particular test that has been done, and this test result is passed to the machine learning model consisting of seven (7) different classifiers ensembled for voting thereby increasing the accuracy of the diagnosis to give a prediction of the likelihood of the specific disease being existent or not in the patient. Running the machine learning algorithm when the doctor diagnosis a patient upon the availability of the patient's laboratory results may cost some system resources. Hence the model for each of the three diseases after the training and testing were saved in a local directory. The CSV file containing the laboratory test results sent from the laboratory department is initially converted into Numpy arrays before the result is reshaped to fit a specified range between 1 and -1. Prediction is finally carried on the saved model used by the reshaped data. This prediction is returned in JSON format and visualized for the user; in this case, the doctor attending to the patient. The classes representing database tables created in the models.py were Consultation, Diagnosis, and Prescription. Consultation is designed to keep the record of the patient, the doctor taking care of the patient, any recommended laboratory tests, and any associated timestamp needed for future reference. Diagnosis is designed to keep the record of the doctor's final diagnosis of the patient and prescription, designed to keep the record of the drugs prescribed by the doctor for the patient. Figure 5 shows the flow diagram for the Machine Learning Module.

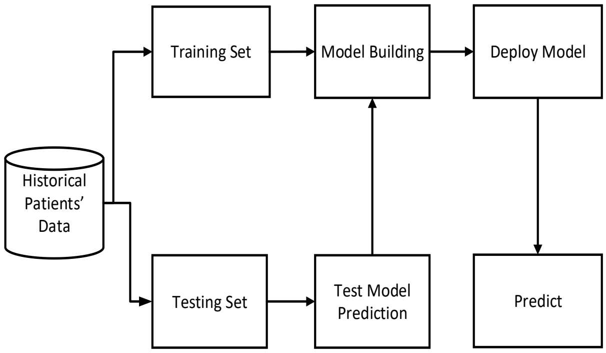


Figure 5: Block diagram for Machine Learning Module

C. Mobile Application

1. First Aid Section: The Medical Care First Aid Application is a user-friendly application that focuses on making first aid principles comprehensible through simplified grammar and videos. The first aid application provides first aid measures on a range of frequently occurring accidents. It classifies burns and scalds, sunburn, prickly heat, heat strokes, and hypothermia as heat and cold effects. It also classifies minor cuts and grazes, nose bleeding, foreign objects in the eye, severe bleeding as wounds and bleeding effects. Heart attacks, asthma, and choking are grouped under heart problems. Spinal injuries, strains and sprains and fractures are also grouped as bone problems. The Accidents covered are as follows: choking, spinal injuries, strains and sprains, fracture, head injuries, accident scene, electrocution, cardiopulmonary resuscitation, seizure, food poisoning , chemical poisoning, snake bites, insect stings, drowning, burns and scalds, foreign objects in eye, heart attack, asthma, severe bleeding, nose bleeding, minor cuts and grazes, hypothermia, heat strokes, prickly heat, sunburn.

2. Medical Aid Section: The Doctor Consultation application creates the room for a group of health practitioners to render health services via the Android platform. The Android platform is used mainly because it is the most used platform in Africa. A medical group has a doctor to render consultation services, nurses to measure patient's body vitals and dispense prescribed treatments, pharmacists to administer prescribed drugs by the doctor. The medical organization also has medical laboratory scientists who dispense laboratory prescriptions from the doctor. A medical organization is registered via registration license obtained from the authorities to enhance the security of operation. A medical organization registration is conducted via a registration page that creates fields for the name of the medical organization, the medical organization's adopted code, password of the medical organization, and the license of operation. A PHP server-side script is used to access a license from the database anytime a new medical body is created. The medical body plays the role of managing the execution of health services on the mobile platform. It also plays an essential role in either accepting or rejecting the registration of patients or other health practitioners. The steps that occur in this section are summarized in Figure 6, which shows a flow diagram of the Medical Aid section.

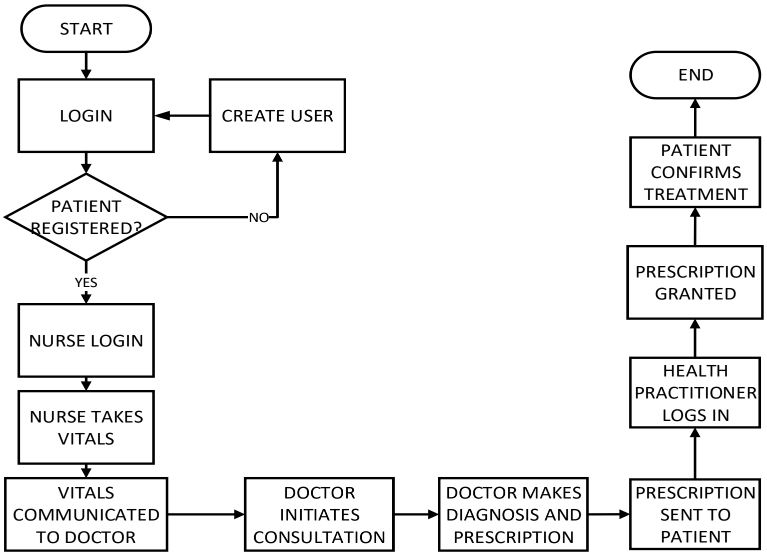

START
LOGIN
CREATE USER
PATIENT REGISTERED?
NO
YES
NURSE LOGIN
NURSE TAKES VITALS
VITALS COMMUNICATED TO DOCTOR
DOCTOR INITIATES CONSULTATION
DOCTOR MAKES DIAGNOSIS AND PRESCRIPTION
PRESCRIPTION SENT TO PATIENT
HEALTH PRACTITIONER LOGS IN
PRESCRIPTION GRANTED
PATIENT CONFIRMS TREATMENT
END

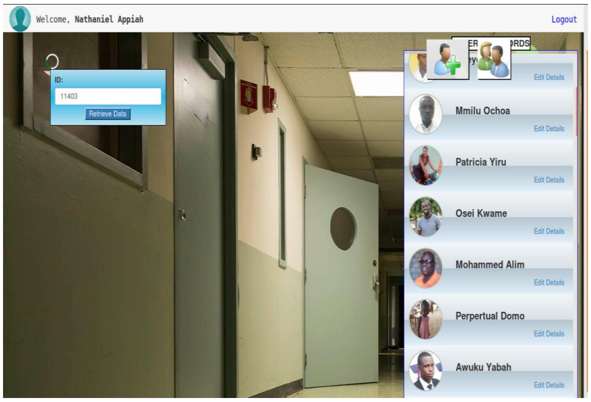
Figure 6: Flow Diagram for Medical Aid Section of Mobile Application

IV. SYSTEM IMPLEMENTATION AND TESTING

This section presents the system implementation and testing of the various submodules developed . It includes the visualization of the results obtained from the research undertaken and discusses them. Tests run to attain the results are also visualized and explained.

A. Web Application

Figure 7 shows a sample result where a patient who may have visited various hospitals can have all his/her records centralized and accessed by their consent in another hospital to let doctors not make uninformed decisions and to aid the machine learning module have more data from the patient's previous records for accurate predictions. Figure 8 shows the sorting of the patient queue according to the level of severity attached by the nurses who take their vital signs.


Welcome, Nathaniel Appiah

ID:

11403

Mmilu Ochoa

Patricia Yiru

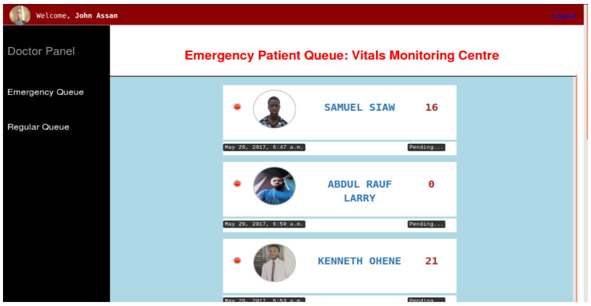
Osei Kwame

Mohammed Alim

Perpertual Domo

Awuku Yabah

Figure 7: The User interface for acquiring patient records if any


Welcome, John Assan

Emergency Patient Queue: Vitals Monitoring Centre

SAMUEL SIAW 16

May 29, 2017, 6:47 a.m. Pending...

ABDUL RAUF LARRY 0

May 29, 2017, 6:50 a.m. Pending...

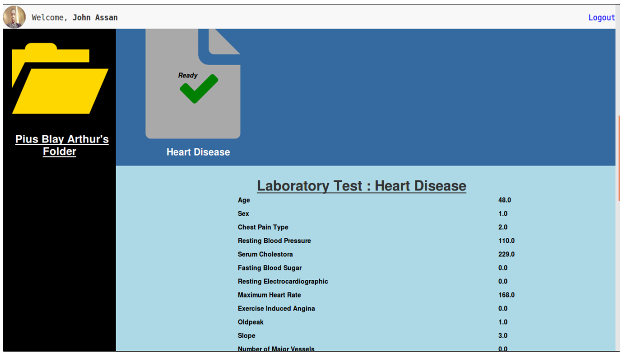
KENNETH OHENE 21

May 29, 2017, 6:53 a.m. Pending...

Figure 8: Patient Queue Sorted According Severity

B. Diagnosis with Machine Learning

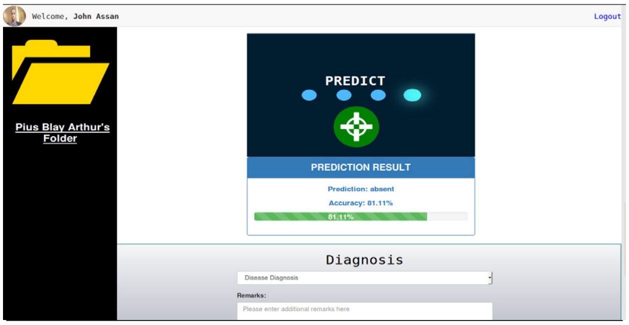
Sample lab tests sent from the laboratory department application were tested against the trained machine learning module, and the classifier with the highest accuracy was chosen as the possible diagnosis. In figures 9 and 10, a resulting laboratory test of heart disease is tested against the trained model, and from the results, the model gives a diagnosis of heart disease being absent with an accuracy of 81.11%. Accuracies going into the ranges of 90% and above were also realized, depending on the lab test result supplied as a test.


Welcome, John Assan

Laboratory Test : Heart Disease

Age	48.0
Sex	1.0
Chest Pain Type	2.0
Resting Blood Pressure	110.0
Serum Cholestora	229.0
Fasting Blood Sugar	0.0
Resting Electrocardiographic	0.0
Maximum Heart Rate	168.0
Exercise Induced Angina	0.0
Oldpeak	1.0
Slope	3.0
Number of Major Vessels	0.0

Figure 9: Sample Lab results for Heart Disease


Welcome, John Assan

PREDICTION RESULT

Prediction: absent

Accuracy: 81.11%

81.11%

Diagnosis

Disease Diagnosis

Remarks:

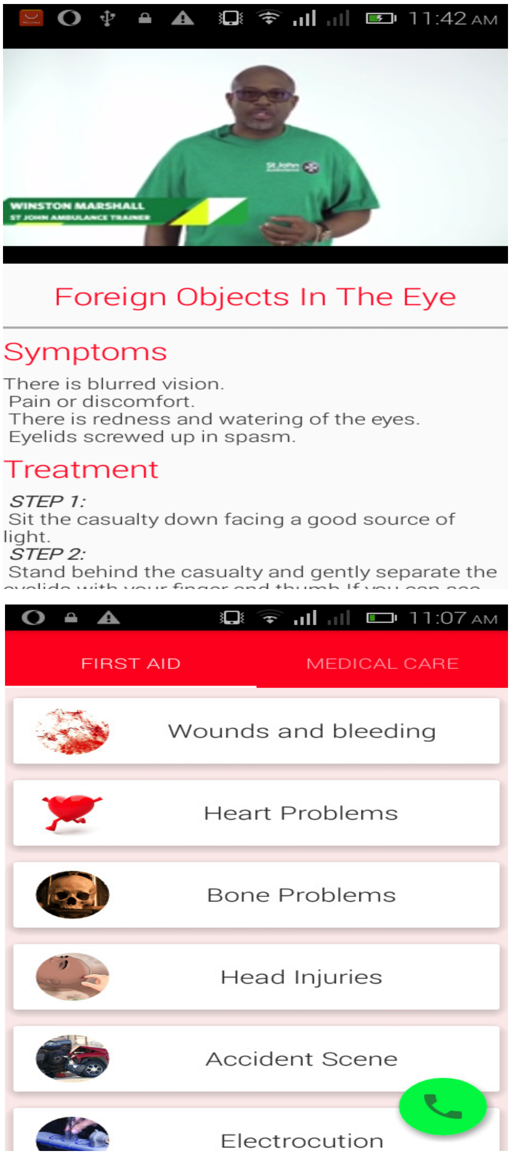
Please enter additional remarks here

Figure 10: Prediction for sample lab test shown in Fig. 9

C. Mobile Application

With the help of the first aid section, various common accidents were able to be handled through the comprehensive step-by-step guide of first aid measures to take. Online videos bolstered textual guides, if not illustrative enough. The calling option provided by a floating action bar also came in handy as ambulances could be contacted. That is shown in Figure 11.

About the medical aid section, shown in figures 12 and 13, patients would be able to have their vitals taken by certified professionals hence reducing the time spent in queueing at hospitals and increasing the rate of quality treatment.



Foreign Objects In The Eye

Symptoms

There is blurred vision.
Pain or discomfort.
There is redness and watering of the eyes.
Eyelids screwed up in spasm.

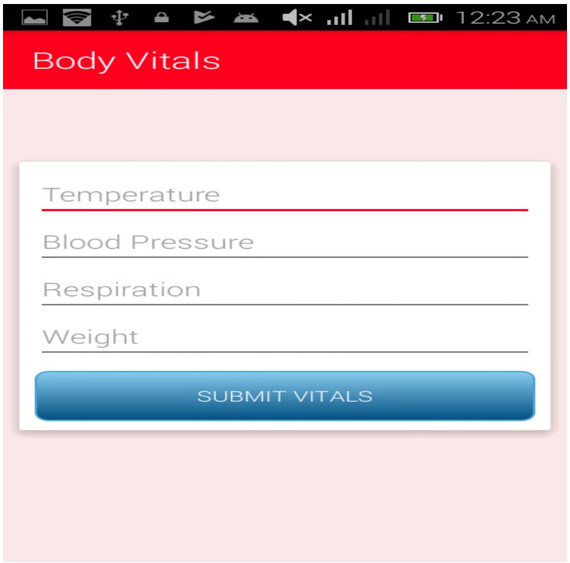
Treatment

STEP 1:
Sit the casualty down facing a good source of light.
STEP 2:
Stand behind the casualty and gently separate the

FIRST AID | MEDICAL CARE

- Wounds and bleeding
- Heart Problems
- Bone Problems
- Head Injuries
- Accident Scene
- Electrocution

Figure 11: List of Possible Accidents with Ambulance call option (left) and Use of YouTube API for Getting Visual First Aid Measures(right)



Body Vitals

Temperature

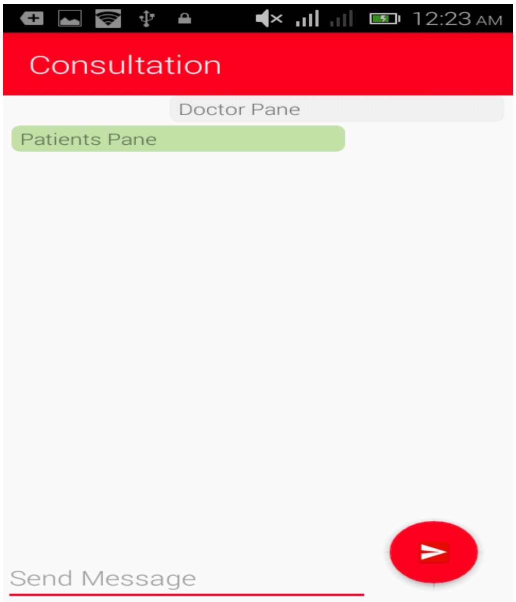
Blood Pressure

Respiration

Weight

SUBMIT VITALS

Figure 12: Form for Nurses to Input Patient Vitals to be sent to Available Consulting Doctors

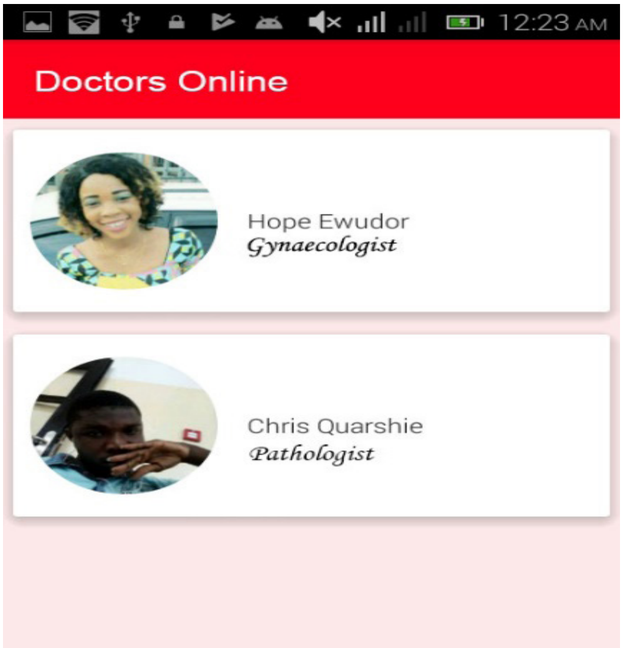


Consultation

Doctor Pane

Patients Pane

Send Message


12:23 AM

Doctors Online

Hope Ewudor
Gynaecologist

Chris Quarshie
Pathologist

Figure 13: Doctors Available for Remote Consultation with Patients (left). Consultation in session (right).

V. CONCLUSION

The design and development of the predictive diagnostic tool with machine learning have sub modules integrated. The seven (7) classifiers developed were implemented via web and mobile applications for intended users to enhance interactivity. The highly accurate machine learning algorithms implemented were based on sufficient training data. Each algorithm worked on the data samples independently; the algorithm that had the highest recall and precision was ultimately chosen for the specific task. This approach led to the accurate determination of a diagnosis of certain kinds of ailments, which was very promising for medical doctors who rely on Electronic Health Records.

The web application developed in this paper enables real-time communication and fast information access between patients and doctors. It also contributed meaningfully to the usefulness of machine learning in medical diagnosis while computing numerical summaries such as statistics in medical diagnostics across networked hospitals. It has also facilitated patient medical record retrieval across the networked hospitals that enable mobility of patients across the healthcare facilities.

It allows knowledge management since all patients' medical records are put on one web application for analytics purposes. The mobile app has a user-friendly interface for interaction between patients and doctors on an Android platform. It also enables the dispensing of medical prescriptions via authorized logins by functional actors in the proposed system.

Even though the system has been well-tested and validated to be efficient and effective in managing health records as well as helpful in diagnostics and making informed decisions based on the statistics generated. For future enhancements, we recommend the following, namely: (1) making the mobile application cross-platform so that mobile users on other platforms can have access to the app, (2) implementing machine learning module directly on mobile application to assist doctors in diagnoses during consultation, (3) scaling the machine learning algorithm for it not to only predict the presence or absence of a disease but also suggest possible drugs to be administered, and (4) using Global Positioning System (GPS) to locate the closest ambulance service for the ambulance speed dial service.

References

S. M. Saif, S. A. Wani, M. Maheswran, and S. A. Khan, "A Network engineering Solution for Data sharing across healthcare providers and protects patients health data privacy using EHR System," *J. Glob. Res. Comput. Sci.*, vol. 2, no. 8, pp. 67–72, 2011.

"What is an electronic health record (EHR)? | HealthIT.gov." [Online]. Available: https://www.healthit.gov/faq/what-electronic-health-record-ehr. [Accessed: 30-Nov-2019].

M. A. Makary and M. Daniel, "Medical error—the third leading cause of death in the {US}," *BMJ,* vol. 353, p. i2139, 2016.

D. S. Jones, S. H. Podolsky, and J. A. Greene, "The Burden of Disease and the Changing Task of Medicine," *N. Engl. J. Med.*, vol. 366, no. 25, pp. 2333–2338, 2012.

B. Koyuncu and H. Koyuncu, "Intelligent Hospital Management System (IHMS)," *Proc. - 2015 Int. Conf. Comput. Intell. Commun. Networks, CICN 2015*, pp. 1602–1604, 2016.

R. S. Khan and M. Saber, "Design of a Hospital-Based Database System (A Case Study of BIRDEM)," *Int. J. Comput. Sci. Eng.*, vol. 02, no. 08, pp. 2616–2621, 2010.

J. F. McCarthy *et al*., "Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management," *Ann. N. Y. Acad. Sci.*, vol. 1020, pp. 239–262, 2004.

H. D. Masethe and M. A. Masethe, "Prediction of Heart Disease using Classification Algorithms," *Int. J. Adv. Eng. Manag. Sci.*, vol. 2, no. 6, pp. 617–621, 2016.

J. S. Sonawane and D. . Patil, "Prediction of Heart Disease Using Learning Vector Quantization Algorithm," *Conf. IT Business, Ind. Gov.*, pp. 0–4, 2014.

C. S. Dangare and S. S. Apte, "a Data Mining Approach for Prediction of Heart Disease Using Neural Networks," *Int. J. Comput. Eng. Technol.*, vol. 3, no. 3, pp. 30–40, 2012.

I. H. Witten, E. Frank, Mark A. Hall, and C. J. Pal, *Data mining : practical machine learning tools and techniques*, 4th ed. Cambridge, MA: Morgan Kaufmann, 2017.

M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest.," *BMC Med. Inform. Decis. Mak.*, vol. 11, no. 1, p. 51, 2011.

D. Gupta, S. Khare, and A. Aggarwal, "A method to predict diagnostic codes for chronic diseases using machine learning techniques," *2016 Int. Conf. Comput. Commun. Autom.*, pp. 281–287, 2016.

A. A. Abayomi-Alli, A. J. Ikuomola, I. S. Robert, and O. O. Abayomi-Alli, "An Enterprise Cloud-Based Electronic Health Records System," *J. Comput. Sci. Inf. Technol. J. Comput. Sci. Inf. Technol.*, vol. 2, no. 22, pp. 21–36, 2014.

K. P. Rao, M. A. Hanash, and G. A. Al-aidaros, "Development of Mobile Phone Medical Application Software for Clinical Diagnosis," *Int. J. Innov. Sci. Mod. Eng.*, vol. 2, no. 10, pp. 5–8, 2014.

N. Asabere, "mMES: A Mobile Medical Expert System for Health Institutions in Ghana," *Int. J. Sci. Technol.*, vol. 2, no. 6, pp. 333–344, 2012.

S. Singh, P. Khadamkar, M. Kumar, and V. Maramwar, "Healthcare Services Using Android Devices," pp. 41–45, 2014.

K. Bache and M. Lichman, "UCI Machine Learning Repository," *University of California Irvine School of Information*, 2013. .

A. A. A.Frank, "UCI Machine Learning Repository: Data Sets," *Univ. Calif. Irvine Sch. Inf.*, 2007.

A. Y. Ng and M. I. Jordan, "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes," *in Advances in Neural Information Processing Systems 14 (NIPS 2001)*, 2001, pp. 841–848.

T. Hothorn and B. Lausen, "Double-bagging: Combining classifiers by bootstrap aggregation," *Pattern Recognit., vol. 36*, no. 6, pp. 1303–1309, 2003.

J. A. *Sanz et al.*, "Lamb Muscle Discrimination Using Hyperspectral Imaging: Comparison of Various Machine Learning Algorithms," *J. Food Eng.*, vol. 174, pp. 92–100, Nov. 2015.

Y.-H. Liu and Y.-T. Chen, "Face recognition using total margin-based adaptive fuzzy support vector machines.," *IEEE Trans. Neural Netw.*, 2007.

David Barber and D. Barber, "Bayesian Reasoning and Machine Learning," *Mach. Learn.*, p. 646, 2011.

S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica, vol. 31*, pp. 249–268, 2007.

E. Alpaydın, "Introduction to machine learning," *Methods Mol. Biol.*, vol. 1107, pp. 105–128, 2014.

S. Samarasinghe, "Neural Networks for Applied Sciences and Engineering," p. 15, 2006.

J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks, vol. 61*, pp. 85–117, 2015.

P. A. Chou, "Optimal Partitioning for Classification and Regression Trees," *IEEE Trans. Pattern Anal. Mach. Intell.*, 1991.

L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers - A survey," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, 2005.

B. Hssina, A. Merbouha, H. Ezzikouri, Erritali, and Mohammed, "A comparative study of decision tree ID3 and C4.5," *Int. J. Adv. Comput. Sci. Appl.*, pp. 13–19, 2014.

P. Rai, "Supervised Learning : K -Nearest Neighbors and Decision Trees," *Mach. Learn.*, vol. 2011, pp. 1–20, 2011.

V. Metsis, I. Androutsopoulos, and G. Paliouras, *Spam filtering with Naive Bayes—which Naive Bayes*? 2006.

A. Y. Ng and M. I. Jordan, *On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes.* 2002.

S. J. Russell, P. Norvig, and J. Canny, *Artificial intelligence : a modern approach, 2nd ed. Prentice Hall*, 2003.

R. Caruana and A. Niculescu-Mizil, *An empirical comparison of supervised learning algorithms*. 2006.

C. Soguero-Ruiz et al., "Support Vector Feature Selection for Early Detection of Anastomosis Leakage From Bag-of-Words in Electronic Health Records," *IEEE J. Biomed. Heal. Informatics*, vol. 20, no. 5, pp. 1404–1415, Sep. 2016.

E. E. Bron, M. Smits, W. J. Niessen, and S. Klein, "Feature Selection Based on the SVM Weight Vector for Classification of Dementia," I*EEE J. Biomed. Heal. Informatics*, vol. 19, no. 5, pp. 1617–1626, Sep. 2015.

L. Han, S. Luo, J. Yu, L. Pan, and S. Chen, "Rule Extraction From Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes," *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 2, pp. 728–734, Mar. 2015.

J. Zhang, Z. Yin, and R. Wang, "Recognition of Mental Workload Levels Under Complex Human–Machine Collaboration by Using Physiological Features and Adaptive Support Vector Machines," *IEEE Trans. Human-Machine Syst.*, vol. 45, no. 2, pp. 200–214, Apr. 2015.