ELECTROCARDIOGRAM(ECG) BIOMETRIC AUTHENTICATION USING MACHINE LEARNING FRAMEWORK

Dr.Kanaka Durga Returi[1]., Devireddy Sai Veenarani[2]., M.Prathyusha[3]., P.Sai Shivani Yashaswini[4] ., V.Sai Pooja[5]

1 Professor, Department of CSE., Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India (✉ durga1210@gmail.com)

2, 3, 4, 5 B.Tech CSE, (19RG1A0514, 19RG1A0532, 19RG1A0546, 19RG1A0559),

Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India

ABSTRACT; To create reliable biometric authentication systems based on Electrocardiograms (ECGs), this research presents a framework for adopting and tuning Machine Learning (ML) methods. Using the suggested framework, researchers and developers working on biometric authentication methods based on electrocardiogram (ECG) data may more precisely specify the scope of necessary datasets and get high-quality training data. Use case analysis is used to establish the scope of datasets. Three unique authentication categories are defined based on three use cases derived from diverse application scenarios using ECG-based authentication. The accuracy of ML-based ECG biometric authentication methods may be improved by providing more quality training data to the appropriate machine learning systems. In this setup, high-quality ML training data is gathered using the ECG time slicing approach with the R- peak anchoring. The quality of ML training and testing data is assessed using four novel metrics presented in the proposed methodology. In addition, a Matlab toolbox is created and made accessible for future research, which includes all suggested mechanisms, measurements, and example data along with demos utilizing different ML approaches. The suggested framework may help researchers organize the necessary ML settings and the ML training datasets, as well as three specified user case scenarios, for developing ML-based ECG biometric authentication. To generate high-quality ML-based training and testing datasets and make use of novel measure metrics, the suggested framework remains relevant for researchers who are using ML approaches to create new schemes in various study fields.

INDEX TERMS; Biomedical signal processing, electrocardiogram (ECG), identification, MATLAB, statistical learning, neural network, regression, and other related terms

INTRODUCTION

Due to the widespread availability of Internet connectivity among modern application systems, biometric authentication has quickly become the norm for user access. As a result, biometric authentication has exploded in popularity as a field of study. Electrocardiogram authentication provides the benefit of incorporating live user body signals during identification, unlike other biometric methods like fingerprint scanning and face recognition. By collecting the user's real-time ECG data, a verification model for person identity may be built using machine learning methods. There have been many recent state-of-the-art literatures [1-4] on biometrics based on the electrocardiogram. A number of issues with ECG biometrics, including authentication classification, pre-processing for data quality improvement, data collections, and selection on Deep Learning, still need further research.

(DL) and other Machine Learning classification approaches[5].

This study introduces a machine learning (ML) framework for ECG-based biometric authentication in an attempt to solve the existing issues with ECG authentication. In order to better understand potential application settings, it is necessary to describe core use cases for ECG authentication. According to the proposed architecture, there are three major use cases for ECG authentication: hospital, security check, and wearable devices. There are also some innovative data pre-processing techniques provided, such as the baseline correction of frequency abnorm=alities in the ECG, a method for reducing noise from ECG data in the case of Power Line Interference (PLI), and a system for flipping the ECG signal if the electrodes are misplaced. In addition, the system uses temporal slicing methods to get ML-based predictions available.

training datasets and fresh metrics for gauging authenticity accuracy. The proposed approach introduces four novel measures to gauge data quality. Accuracy Percentage within Ranges (APR), Accuracy per UCL (APU), Upper/Lower Range Control Limits (UCL/LCL), and Mean Absolute Error Rate (MAER).

You can see how the new framework model for ML-based ECG biometric authentication is laid out in Figure 1. A number of different ML methods are used inside the central part of the process, including Decision Tree (DT) and Support Vector Machine (SVM) for regression, and Artificial Neural Network (ANN) and Convolutional Neural Network (CNN) for classification. A time slicing approach is also created for ECG data and linked to the primary procedure.

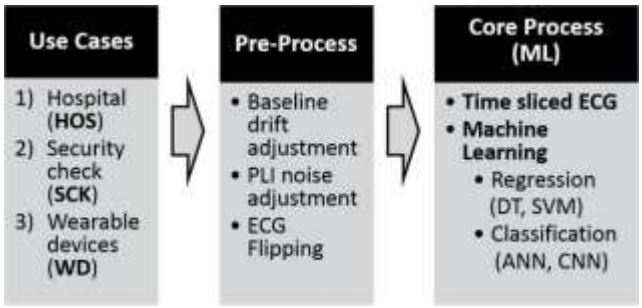
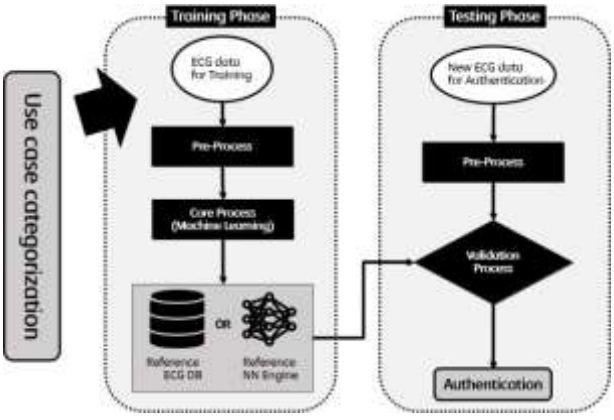


FIGURE 1. Overview of the New Framework Model for ECG based Biometric Authentication

The data flow diagram for the ECG biometric authentication framework utilizing Machine Learning methods is shown in Figure 2. The training phase of the proposed framework begins with the selection of an appropriate user category for the target system environment, followed by the acquisition of matching ECG data from target users. After receiving the training data, the pre-process procedures are used to them to provide filtered, higher-quality (less noisy) data. In order to produce an authentication assessment model, the filtered data is fed into one of the specified core process mechanisms shown in Figure 1. As an assessment model for any ML-based ECG authentication system, the proposed core process presently supports both an ECG reference database and a trained Neural Network (NN) reference engine. After the training process is done, a reference database (or NN Engine) is produced. A user authentication request based on ECG data is created during testing, and the ECG data must first use data pre-processing procedures in order to provide filtered data with better quality (less noise). To make a final determination about this authentication request, the filtered data are sent into a validation procedure, which subsequently consults a reference database (or the NN Engine).



The study article comes with a brand-new Matlab toolbox that contains all the recommended procedures, measurements, and example data used to demonstrate the various ML algorithms. The developed and publicly available toolbox is now open to further research.**FIGURE 2. The Framework Processing Flow for EGC Biometric Authentication using Machine Learning Technologies**

The article is divided into seven parts. In Section II, we will talk about three different types of authentication that are grounded in specific use cases. Three innovative pre-processing approaches to improve the quality of obtained ECG data are shown in Section III. In Section IV, we provide two methods for training ML data and the time slicing methodology for ECG data to enhance the performance of the applicable ML schemes during training. In Section V, we provide four fresh measures of data quality. Our Matlab toolbox, whose features are outlined in Section VI, is used to test how well our suggested framework works in practice. Finally, Section VII provides a brief overview of the proposed framework and our contributions to it.

I. AUTHENTICATION CATEGORIZATION BASED ONUSE CASES

There has been a lot of research on utilizing ECG data for biometric authentication [6-11]. Despite the plethora of works on ECG-based authentication, all of them rely on a variety of user contexts and ECG detecting equipment. Data from electrocardiograms (ECGs) are often collected [6-8] by researchers with competence in medical engineering. In contrast, researchers with backgrounds in electrical engineering often deploy wearable devices equipped with basic ECG sensors [9-11] to collect electrocardiogram data. Because of this, it is crucial to take into account the user's surroundings and the potential ECG detection equipment while designing an ECG based biometric authentication strategy.

A use case is a documented scenario that illustrates how the intended system will be used. System needs may be determined early on in the design phase with the help of use case analysis, which also helps to explain crucial information for system processes [12].

Use case analysis might be used to classify the many use cases that could arise. Patients in a hospital, identity checks at building entrances, and continuous authentication for personal use (also known as HOS, SCK, and WD; see Figure 1) are the three authentication categories identified using use case analysis on potential application scenarios for ECG based user authentication. The next sections discuss the related user environments and assumptions for each class. Please take note that the authentication speed and accuracy rate requirements for each category are uniquely based on the systems that will be using the authentication.

A. Each category's description and associated requirements are provided below. While there may be similarities across authentication scenarios, distinct performance metrics will be required in each situation.

B.

HOSPITAL (C) PATIENT IDENTIFICATION AND AUTHORIZATION Typically, an electrocardiogram (ECG) is used to determine whether a patient has cardiac disease or heart stress. Obtaining high-quality ECG data from a patient requires extensive and expensive equipment used in medical diagnostics. Therefore, numerous leads are needed during an ECG test, and the sampling period for acquiring ECG data is rather lengthy (from a few of minutes to hours, depending on the kind of ECG test). Category 1 (Hospital Operating System) use cases now include patient identification through ECG testing. Patients are assumed to pre-register both their identities (i.e., names or legal identification numbers) and their prior ECG records. For an ECG based authentication strategy to function properly, it is also expected that the measured ECG signals from the same patient are consistent enough (i.e., the recorded ECG signal values within a normal range) to be used for both the registration (training) and verification (testing) phases. The hospital may then use an ECG-based biometric authentication technique to positively identify such individuals during their subsequent hospital visits. Contrast the time it takes a nurse to ask for a patient's name and legal identification number with the time it takes a well-trained ECG user authentication model (or scheme) to identify a patient by evaluating live ECG signals (less than a couple of seconds) [13]. Patients who have lost consciousness in the emergency department may be readily identified using ECG-based user (or patient) authentication. One of the most common settings where ECG-based authentication techniques might be used is in patient authentication in hospitals. This HOS application sees the highest usage in the healthcare and medical research sectors [14]. Many databases, such the PhysioBank database [15], are open to the public and include archived ECG data.

The obtained electrocardiogram (ECG) data may be inverted or include sounds as a result of the intricacy of the data collection methods.

resulting from faulty electrode placement or PLI. Because of this, the quality of the ECG data must be improved by data pre-processing procedures before it can be used to train a user assessment model. patient) Authorization Function.

PERSONNEL AUTHENTIFICATION AT THE LOBBY OF A BUILDING (SCKSecurity check (Category 2; SCK use case) for building access and room entrance is enhanced by the second user authentication use case based on user ECG data. Most businesses nowadays have security checkpoints where staff and visitors must provide proper identification. Now that portable ECG detection devices or ECG detection sensors are readily available, ECG based biometric authentication systems can join fingerprint scanning, facial recognition, voice identification, iris scanning, and retina scanning as options for user authentication at security checkpoints. An ECG-based authentication system may be utilized in this SCK scenario to distinguish between known and unknown individuals (such as guests). The ECG authentication system works on the premise that all legitimate workers have already registered their identities (i.e., names or legal identification numbers) and associated ECG data. Furthermore, it is expected that the same employee's recorded ECG signals are stable enough to be used in both the registration and verification processes.

C. CONTINUOUS AUTHENTICATION FOR PERSONALUSAGE (WD)

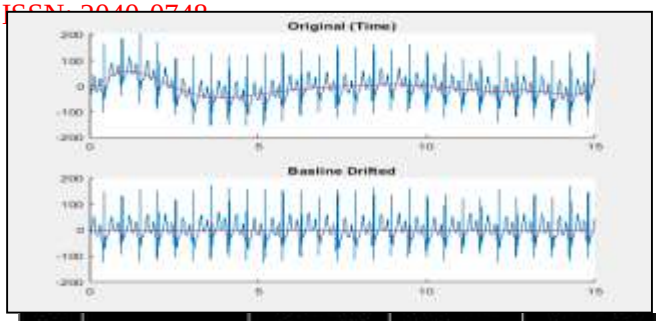
A third category of WD use cases for ECG-based user authentication is personal wearable gadgets (like a smart watch) equipped with ECG sensors to continually check if the person using the item is its legitimate owner. Generally speaking, wearable tech merely requires constant proof of ownership authentication. A new framing technique may be needed to normalize the received ECG signals by filtering out potential signal noise caused by user body status, as the heart beat period (R-R peak period) and amplitude of ECG signals of a person may change dramatically when the person is under different body statuses like walking, running, and sleeping. A wearable device that includes an ECG sensor and an ECG authentication module may provide the user with a means of two-factor authentication. A WD user may have greater command over the safety of their wearable device by using a two-factor authentication mechanism (for example, authentication with both user password and user ECG data).

Table I displays an overview of the many types of use cases. Measure metrics, person identification limits, and referenced ECG data properties all need to be tailored to each use case type. Table I shows that although the HOS example doesn't take the possibility of patients having uncertain identifications into account, the SCK case's authentication method does.

There were several misidentifications. Different bodily states even from the same individual need thinking about how to cancel out the noise in the ECG data in the WD instance.

TABLE I
DATASET BOUNDARIES CATEGORIZED BY AUTHENTICATION USE CASES

ISSN: 2049-0748



Cat No.	Cat. Name	Known ID classification	Unknown ID	Personal Status
1	Hospital (HOS)	O	X	X
2	Security Check (SCK)	O	O	X
3	Wearable Devices (WD)	X*	O	O

** There is only one known ID in the WD case*

Based on the use case types, researchers may take into account further criteria. In contrast to hospital patients, security checkpoint personnel should have a shorter ECG sampling time. Because the goals for using the obtained ECG data are different, the sampling frequency for ECG signals in the WD situation should be substantially lower than the ECG sampling frequency provided by typical medical measuring equipment on ECG signals. Furthermore, as different kinds of ECG equipment are used between the HOS instance and WD case, human operation problems such misplacing leads will not be included in WD category.

PREPROCESSING ECG DATA TO IMPROVE QUALITY

III. Prior to beginning the main process (that is, the machine learning process), it is necessary to make certain adjustments to the data. Since ECG data might be seen as signals, the signal processing methods have been extensively utilized to altering ECG data. Signal processing techniques, such as filter designs [16–18] and Fourier Transforms [19, 20], are used to improve ECG detection. While various pre-processes are used to improve signals before machine learning is used, only three are approved for improving ECG data.

A. ADJUSTMENT OF THE STANDARDS Breathing while wearing electrically charged electrodes causes a low-frequency distortion in the ECG known as the baseline drift (or baseline wander) [21]. The baseline drift is being corrected for via the baseline adjustment procedure. In most cases, raising the high-pass filter's cut-off frequency is necessary for full baseline wander reduction. in the signal, higher than the lowest frequency. Most methods for eliminating baseline wander have the property of canceling out the signal's low-frequency components. In terms of how often

A high-pass filter for preventing baseline wander is often tuned to a frequency just about 0.5 Hz [22]. Although these methods have been explored extensively and are commonly used [23], it is still necessary to identify in advance the frequency that would effectively eliminate baseline drift. Baseline drift during ECG data collection may be caused by factors other than low-frequency noise, such as the applicant's own movement. If we do not have a good indication of the threshold frequency or the expected motion of an application (HOS example), then these filtering strategies may be useless.

These problems might be avoided if a baseline correction was performed using a curve fitting technique. Curve fitting refers to the process of creating a curve that best fits a collection of data points [24], and polynomial curve fitting is a technique that provides a more accurate and smooth fit to the data [25]. Figure 3 demonstrates how the fitted curve data may be subtracted from the original ECG data to baseline ECG should be adjusted.

FIGURE 3. Baseline adjustment by using the polynomial curve fitting [15]

This method of adjusting the baseline by fitting a polynomial curve is helpful if the low-frequency noise is a source of concern beyond only the baseline drift. ECG amplitude changes are also prevented by the automated baseline adjustment to zero. Movements of the applicant during the collection of ECG data are an unusual source of the baseline wander beyond a low frequency noise.

A. A. CUTTING DOWN ON POWER LINE DISTURBANCE AND NOISE Multiple sources of noise signals, such as baseline wander and PLI noise that is connected to signal cables, might be problematic. Cables used to transmit data between the exam room and the monitoring equipment are vulnerable to electromagnetic interference (EMI) of frequency due to uninterrupted power lines [26]. This is a regular occurrence in healthcare facilities. Power line electromagnetic fields are a frequent source of noise in the electrocardiogram (ECG), often manifesting as sinusoidal interference at 50 or 60 Hz and potentially accompanied by a number of harmonics. This kind of narrow-band noise is problematic.

D. relying on waveforms with low amplitudes, since this might lead to misleading and incorrect results [27]. To get rid of PLI noise, the Infinite Impulse Response (IIR) notch filter is often used [28]. Signals in a certain frequency range, known as the stop band frequency range, are rejected or attenuated by a notch filter, while signals above and below the stop band frequency range are passed unaltered [29]. These filters might be used to eliminate the baseline drift, but only after a suitable target frequency has been established. However, a notch filter achieves comparable results without requiring knowledge of the target frequency in advance by eliminating aberrant peak points in the frequency domain. This method might be used to get rid of the PLI noise

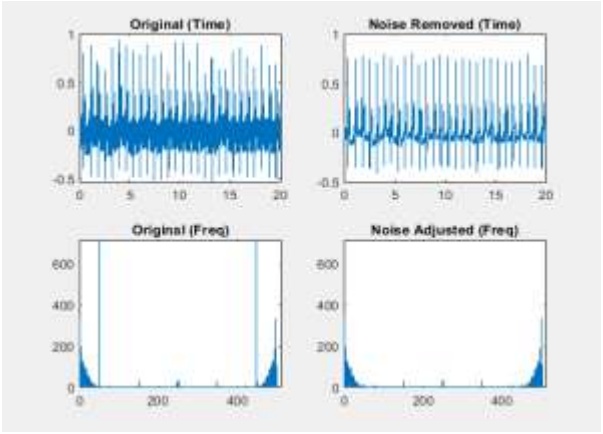
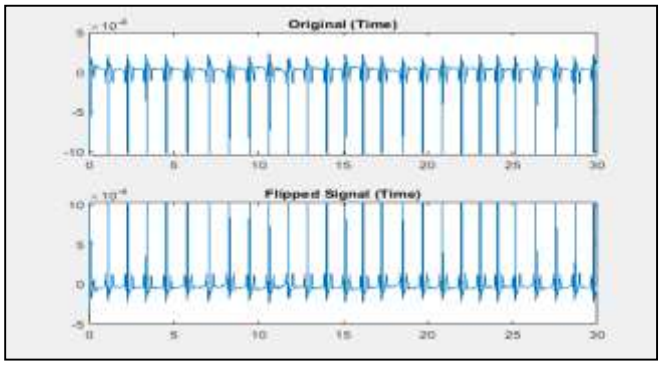


Figure 4 since PLI has the highest peaks in the frequency domain.

FIGURE 4. **Noise adjustment for removing PLI frequency [30]**

The Fourier Transform has been used to locate out-of-the-ordinary peaks in the frequency spectrum. It is possible that the PLI frequencies and other out-of-the-ordinary peaks in the frequency domain constitute noise, and that their removal might improve the original ECG data. Noises might be eliminated altogether, or only the frequency at which they occur. In addition to the PLI at 50 Hz, baseline wander at 1 Hz was also noted above. Even without filters, the Fourier Transform may be used to enhance the ECG data by eliminating certain frequencies. It should be observed that the PLI occurs in the HOS scenario but never occurs in the WD situation. Wearable device authentication may not benefit from the PLI noise elimination method since PLI occurs only when medical ECG equipment is used, even in the SCK instance. An ECG receiver using a straightforward sensor type device should not worry about the pre-process



for PLI noise elimination.

Professionals are often needed to set up and measure ECG signals on medical devices like an electrocardiograph. However, even experts may make blunders, such as incorrectly positioning electrodes. Due to human error in hospitals, ECG signals can invert. Thus, it is preferable to verify whether or not a target ECG data (either training or testing phases) is flipped, as shown in Fig. 5. If the original ECG data is inverted, this pre-process will correct it so that it reads correctly. The flipping state of ECG data may be quickly determined by using a simple mean check rather than a thorough analysis of the whole data set.

FIGURE 5. **Flipping ECG data example [31]**

It depends on the authentication scenarios whether or not all three pre-processes need to be applied simultaneously. For instance, the WD example specifies that the "flipping signal" cannot be used on wearable devices. Other operations, such as upgrading the company's entry system and adjusting the sample frequencies for recorded ECG data (a reference database), are also regarded as preliminary steps. After all the preliminary steps have been taken, the primary procedure, consisting mostly of ECG time slicing and machine learning training, may begin. following section provides the details.

CUTTING TIME IN HALF AND MACHINE LEARNING

III. For the purpose of creating the machine learning training dataset, the time slicing method is being examined. This method works particularly well when amassing data for use in machine learning. A slice (window) time (more often referred to as a sliding window) is used to anchor the ECG data to the R-peak. Each piece of data generated in this way may be used as a training input for a machine learning algorithm. In addition to being easily combined with other training inputs, the time-sliced ECG dataset may be used for a wide variety of purposes.

Use the Machine Learning (ML) Training Techniques outlined in Section B.

A. A. RECORDINGS OF ECG CHANGES OVER TIME FOR MACHINE LEARNING

The QRS complex is composed of the three most prominent graphical deflections on an electrocardiogram and hence serves as its focal point and primary visual indicator [32]. The highest point of a QRS complex, denoted by an R-peak. Indicative of a single cardiac cycle, the R-peak is often used to anchor the QRS complex, using related techniques such as R-peak identification and sliding window time optimization [33]. In order to layer the ECG signal segments based on the R-peak moment (i.e., R-peak anchoring), we may use time slicing, which is essentially slicing ECG data in the time domain, to cut the ECG signal from an R-peak moment to the sliding window period. Each R-peak anchored slice from Figure 6 is used as a training sample for the ML algorithm. Here, we use a slicing time (i.e., sliding

window) of 0.6 seconds, which is similar to a heart rate of 100 beats per second (bps), based on the average minimum of a heartbeat interval from atypical heart rate [34]. Some metrics of machine learning performance are sensitive to the amount of time spent slicing an electrocardiogram and may be adjusted in different ways depending on the project's goals. Although improving window duration for biometric purposes is not currently under consideration, it is one of the many exciting research opportunities in the field of ECG-based security.

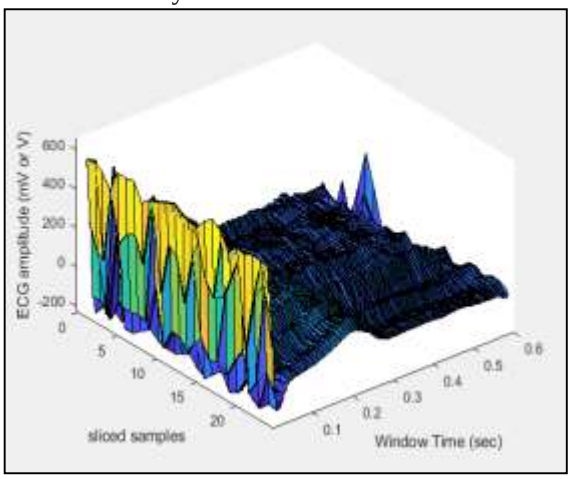


FIGURE 6. ECG Time Slicing with the R-peak anchoring [15]

Data-driven machine learning for training purposes

Machine learning is a branch of AI that allows computers to learn how to solve problems on their own, without being given any precise guidelines. Predictions (regressions) or choices (classification/pattern recognition) may be made by machine learning algorithms [35] without the need for human intervention in the form of explicit programming. Using AI methods to examine electrocardiogram data has been the subject of much research [5, 6, 8, 36–39]. The goal of this study is to apply machine learning methods to train time-sliced ECG data. Several methods are shown to illustrate the integration of a machine learning model into the ECG-based biometric authentication process.

REGRESSION-FOCUSED MACHINE LEARNING APPROACH

Predicting a target value using a model constructed from many variables is the goal of the data mining work known as multi-variable regression. SVM stands for "support vector machine" and is a "discriminative classifier" [40]. Since the 1990s [41], this method has been used for the regression. Regression models in the shape of trees may also be constructed using the Decision Tree (DT) algorithm. It divides a dataset into subgroups while simultaneously growing a decision tree [42, 43]. Table II shows the results of using the DT method vs the SVM method to construct a regression model using data from the PhysioBank database [15].TABLE II

COMPARISON TABLE OF DT AND SVM

DT		SVM	
Results		Results	
RMSE	33.761	RMSE	36.168
R-Squared	0.59	R-Squared	0.54
MSE	1139.8	MSE	1308.1
MAE	19.164	MAE	19.01
Prediction speed	~740000 obs/sec	Prediction speed	~600 obs/sec
Training time	5.2194 sec	Training time	2008.2 sec

There are many additional machine learning models that might be used to create a regression model in addition to these two. Additionally, a DT model may be the optimum solution just for a certain computer system configuration and specific (time sliced) ECG data sets. A variable dataset or computer architecture may affect which machine learning model is best suited for a certain regression technique.

1) 1) A PROCESS BASED ON MACHINE LEARNING FOR CLASSIFICATION
2) The goal of classification is to determine which set of categories applies to a given collection of observations [44]. Each identification corresponds to a distinct category set, and the whole collection of identifications may be thought of as a single category. Using the time-slices of the ECG dataset as inputs, neural network (NN) models might be developed (either Artificial Neural Networks [45] or Convolutional Neural Networks [46]; see Figure 7). Network model performance should vary, and building an effective NN model is yet another area of study [47–50]. Both neural network models may utilize the same input nodes and the same output nodes regardless of the NN model(s) being used. Both models may use input samples from the time-sliced ECG dataset created using the R-peak anchoring.
3)
4)

The purpose of categorization is to assign a group of observations to a predetermined set of categories [44]. It is possible to see the whole set of identifications as a single category, with each individual id mapping to its own unique category set. Neural network (NN) models (Artificial Neural Networks [45] or Convolutional Neural Networks [46]; see Figure 7) might be created using the time-slices of the ECG dataset as inputs. There should be variation in network model performance, and there is more research into developing efficient NN models [47–50]. The input and output nodes of one neural network model may be shared with another NN model. The time-sliced ECG

dataset generated with the R-peak anchoring may be used as input for both models

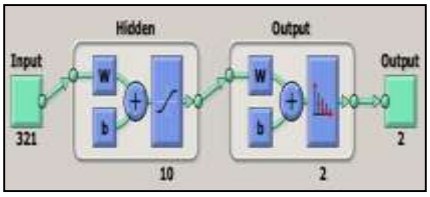


FIGURE 7. Design of 10 hidden layer Convolution Neural Network (WD)

It should be emphasized that this study does not discuss any specific models; rather, it aims to provide suggestions for creating more effective neural network (NN) authentication systems.

In the next part, we will discuss the metrics used to evaluate the soundness of a machine learning model and its accompanying training dataset.

III. III. MEASUREMENTS OF DATA QUALITY

- Delivering high-quality samples is crucial for evaluating machine learning algorithms, which is an integral aspect of every ML project [51]. Values from evaluating the quality of sample data and regression algorithms used in based authentication systems are used to create certain performance measurements (evaluation metrics) [52]. Common indicators of data quality and performance for a regression strategy include:
-
- • Error in Squaring the Sum
- • Total Sum of Squares
-
- • Error, Mean Squared
- Consider the M.A.E.
-
- The study presents a number of novel measuring criteria in addition to the usual quality indicators. These metrics of quality are used not just as criteria for validating samples, but also in the evaluation of machine learning systems that use a

$$UCL = \bar{R}_Y + \sigma(b)\bar{R}_Y,$$

ed metrics are as

$$LCL = Max\left(0, \bar{R}_Y - \sigma(b)\bar{R}_Y\right),$$

Accuracy Percentage within Ranges; Upper and Lower Range Control Limits; Mean Absolute Error RateLimit of Control Accurac New measurements of data quality are described in further depth below.

AVERAGE PERCENTAGE OF ERRORS (MAER)

Measures of error that may be used to the assessment of a machine learning model include the Mean Square Error (MSE) and its root, the root mean squared error (RMSE).

The MSE measures how different the two estimated datasets are on average [53]. Both the MSE and its square root, the RMSE, are common metrics for evaluating machine learning models. Despite the fact that both MSE and RMSE are often used for most regression

where n is the standard (usually the mean) of the data set Yn. It should be noted that division by zero is something the formula (1) has been constructed to prevent. The MAER uses the mean values of the prediction after machine learning training as the reference values (n, n = 1,..., N). The MAER is computed not using a fixed mean or arbitrary differences, but rather on the basis of the moving mean (that is, the reference values). It's a good indicator of how well the regression is doing since it gives you the range's reference value component. The MAER has a fixed range from 0 (zero) to, with a lower number indicating superior performance.

Not only may a regression-based machine learning model be evaluated using the MAER, but the quality of the ML training samples themselves. Outliers are values in the data set that fall beyond the MAER tolerance band, and it is possible to improve the quality of the data set by removing them. Subsection B) discusses how a quality engineering approach (namely, Statistical Process Control) may be used to determine the appropriate MAER value to use as a data quality threshold.

CONTROL RANGE BOUNDARIES (UCL, LCL) In quality engineering, statistical process control (SPC) is often used as a synonym for statistical quality control (SQC) due to its same definition [54]. Important uses of SPC in quality control include control charts and acceptance sampling [55]. There are two distinct kinds of control charts: X-charts, which focus on central tendencies, and R-charts, which focus on dispersion. A control chart's control limits (Upper Control Limit and Lower Control Limit) are two horizontal lines, often drawn three or six standard deviations (SD) from the chart's mean. Both the reference values and no reference values might be used in the construction of the R-charts. Outliers may be filtered out using the control values shown in both charts before a dataset is utilized for training. You may determine the upper and lower bounds (UCL and LCL) of the values in range R in the following ways:

(2)

where (b) is the fraction of the normal distribution's sigma level at which range-based acceptance of data is appropriate (usually 3 sigma; b = 3; (b) = (b) (0); (z) is the normal distribution function). The R-bar indicates an average over a specified interval. The control limits might be established using either the mean squared error (MSE) or the mean absolute error (MAER, MAE). The UCL becomes the threshold range value for judging the datasets, and a lower

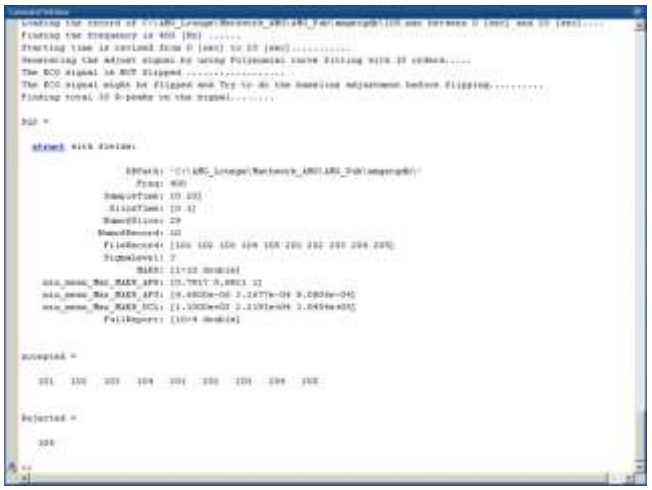
UCL is indicative of greater performance. Training and testing datasets may be properly organized from raw data using the SPC's R-chart as a cutoff.

A. A. RELIABILITY (IN/OUT OF RANGES)) data samples.

B. How much $APR = \frac{n(\{R \leq UCL\})}{n(\Omega_R)}$ all within the thresholds c onfidence limit (UCL) is the Accuracy Percentage within Ranges (APR). Numbers inside ranges are counted and compared to the total number of ECG segments.(4)

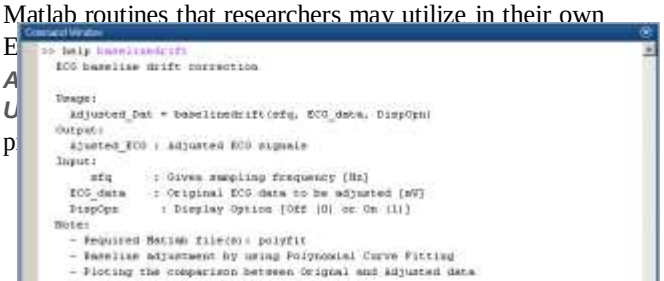
Even before data is validated, the APR shows the quality of the time sliced ECG data, with a bigger APR indicating better performance. When comparing the testing ECG data to the reference data, this number will also serve as the cutoff for rejecting the testing data. The higher the APR, the better the datasets and ML systems are.

ACCURACY AS MEASURED BY THE UNIFORM CONTROL LOCALITY (APU) The ratio of the APR to the UCL reveals how much progress has been made in training; this is shown by the Accuracy percentage within ranges per the upper control limit (APU). though the UCL is also decreasing, then the APR may be low even though the accuracy is good. The APU is determined by the steps below: These metrics are effective for evaluating the quality of both training and testing datasets across a variety of engineering disciplines.



TOOLBOX FOR MATLAB (IV) Incorporating ML approaches into ECG-based biometric identification systems requires the core process and essential components described in Sections I–IV. Each section's recommended Toolbox for illuminating those processes and approaches is, in practice, Matlab functions. This section presents some of the tools available in the amgecg Toolbox (Amang ECG Toolbox), a collection of Matlab routines that researchers may utilize in their own E

A. A
U
p



following Matlab functions: baseline drift correction (Figure 3), noise adjustment (Figure 4), and ECG data flipping (Figure 5).

By using Matlab's "help" feature, one may learn how to make use of the program's many tools. Figure 8 displays the Matlab command window, where users may learn specifics about how to utilize each function.The APU is an additional metric for assessing the efficacy of your data and ML setup.Quality metrics already developed for ML datasets may still be utilized carefully. For regression-based Machine Learning systems used in biometric authentication, the novel data quality measurements proposed in the research may provide an alternate assessment metric. Depending on the kind of use case and the goals of the project, the right set of quality metrics to use will be a question of personal preference. New measures are being recommended, and they're based on methods used in

FIGURE 8. Using 'help' function in the Matlab command window

The function of the time slicing of the ECG data also provided in the *amgecg Toolbox* as shown in Figure 6 and this*timeslicedecg* function is one of core for adopting ML techniques into ECG authentication projects and it gives flexibilities to adopt various ML techniques for ECG data. The ML system could be designed by using time sliced ECG data and each sliced data is considered as the input of any ML systems. It is noted that actual Machine Learning implementations such as the DT regression and CNN classification (Figure 7) are not included in the *amgecg*

Toolbox but users can implement ML systems by using this Toolbox.

Data Quality Tools B.

The utilities for doing such an analysis are also included in the amgecg Toolbox. The toolbox includes the data quality indicators discussed in Section V, as well as the following associated operations:

• rangecontrol • maer

• mseamg • maerdataqualityengine • msedataqualityengine

The data quality functions are a hybrid of more fundamental and unified operations. The "help" function in Matlab is where users may learn more about the program's features. Figure 9 further shows that the demo file may be found inside the Toolbox.

FIGURE 9. Demonstration of MAER based Data Quality

Readers are reminded that the Matlab source codes (i.e., amgecg Toolbox) are publicly accessible on GitHub1. In addition, you may see examples of how to use the amgecg Toolbox's features on YouTube2.

VI. CONCLUSION;

In the near future, ECG based biometric authentication will be used on enormous application systems all over the globe as new ECG detection devices become portable, lightweight, embeddable with smartphones and wearable devices, and connectable with distant servers using wireless technologies. In order to construct a more reliable assessment model for ECG-based biometric authentication, ML approaches are often utilized. Here we provide a generic machine learning framework.

introduction of a biometric identification system based on the electrocardiogram. Researchers will find the suggested framework useful for quickly designing and evaluating an ML-based ECG user authentication strategy since it details the overall data processing flow of such a mechanism. These features consist of a publicly accessible Matlab Toolbox (i.e., amgecg Toolbox), as well as four new data quality measures, a temporal slicing approach to create high-quality ECG datasets, and three new data pre-processing techniques. The proposed framework offers several data pre-processing techniques and newly defined measure metrics that can aid researchers in speeding up the development of ML-based schemes, even if they aren't specifically interested in ECG based biometric authentication.

ACKNOWLEDGMENT;

Some of the ECG datasets used in this work were generously provided by A. Khandoker and H. Alsafar, and the authors would like to express their gratitude. For his insightful feedback on our novel ML adaption technique for ECG biometrics, Jiankun Hu deserves special recognition. Grant Number 8474000137-RC1-C2PS-T3 from the Center for Cyber-Physical Systems at Khalifa University is greatly appreciated, as is Grant Number MOST 107-2218-E-011-002 from the Taiwan Information Security Center (TWISC) and the Ministry of Science and Technology, Taiwan.

REFERENCES

[1] The article "HeartID: A Multiresolution Convolution Neural Network for ECG-Based Biometrics Human Identification in Smart Health Applications" was published in IEEE Access, volume 5, pages 11805-11816, 2017.

[2] In 2018, IEEE Access published "Evolution, Current Challenges, and Future Possibilities in ECG Biometrics" by J.R. Pinto, J.S. Cardoso, and A. Lourenco.

[3] In 2018, the IEEE Transaction on Information Forensics and Security published "Learning Deep Off-the-Person Heart Biometrics Representations" by E.J.S. Luz, G.J.P. Moreira, L.S. Oliveira, W.R. Schwartz, and D. Menotti.

[4]

[5] S. Y. Chun, J.-H. Kang and et al., "ECG Based User Authentication For Wearable Devices Using Short Time Fourier Transform", in Proc. 39th IC-TSP, Vienna, Austria, 2016, pp. 656-659.

[6] A. F. Hussein, A. K. AlZubaidi and et al., "An IoT Real-Time Biometric Authentication System Based on ECG Fiducial Extracted Features Using Discrete Cosine Transform", [Online]. Available: https://arxiv.org/abs/1708.08189

[7] usability.gov, Use Cases, [Online]. Available: https://www.usability.gov/how-to-and-tools/methods/use- cases.html, Accessed on Feb. 1, 2019.

[8] E. Zaghouani and et al., "ECG based authentication for e-healthcare systems: Towards a secured ECG features transmission", in Proc. 13th IWCMC, Valencia, Spain, 2017, pp. 1777-1783.

[9] F. SufiIbrahim, K. Hu, "ECG-Based Authentication", Handbook of Information and Communication Security, pp. 309-331, 2010.

[10] A. L. Goldberger, L. A. N. Amaral and et al., PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals, Circulation vol. 101, no. 23, pp. e215-e220, June, 2000. [Online]. Available: http://circ.ahajournals.org/content/101/23/e215.full.

[11] Y. -W Bai, W. -Y. Chu and et al., "Adjustable 60Hz noise reduction by a notch filter for ECG signals", in Proc. IEEE Instrumentation and Measurement Technology Conference, Como, Italy, 2004, pp. 1706-1711.

[12] A.R. Verma, Y. Singh, "Adaptive Tunable Notch Filter for ECG Signal Enhancement", Procedia Computer Science 57, pp. 332-337, 2015.

[13] E. Ebrahimzadeh, M. Pooyan and et al., "ECG signals noise removal:Selection and optimization of the best adaptive filtering algorithm based on various algorithms comparison", Biomedical Engineering Applications Basis and Communications vol. 27 no 4, pp. 1-13, July, 2015.

[14] P. Chen, M. Chang and et al., "Study of Using Fourier Transform to Capture the ECG Signals between Awakeness and Dozing", in Proc.IS3C, Xian, China, 2016, pp. 1055-1058.

[15] H. Gothwal1 and et al., "Cardiac arrhythmias detection in an ECG beat signal using fast fourier transform and artificial neural network",
J. Biomed. Sci. and Eng. vol. 4, pp. 289-296, 2011.

[16] G. Lenis, N. Pilia and et al., "Comparison of Baseline Wander Removal Techniques considering the Preservation of ST Changes in the Ischemic ECG: A Simulation Study", Computational and Mathematical Methods in Medicine Vol. 2017, 13 pages, 2017.

[17] Y. Luo, R. H. Hargraves and et al., "A Hierarchical Method for Removal of Baseline Drift from Biomedical Signals: Application in ECG Analysis", The Scientific World Journal, Vol. 2013, 10 pages, 2013.

[18] F. P. Romero, L. V. Romaguera and et al., "Baseline wander removalmethods for ECG signals: A comparative study", [Online]. Available:https://arxiv.org/abs/1807.11359

[19] Arlinghaus, Sandra L. (ed.) Practical Handbook of Curve Fitting. CRC Press, [Online] Available: http://hdl.handle.net/2027.42/58759,
Accessed on Feb. 1, 2019

[20] Mathworks, polyfit, [Online] Available:
https://www.mathworks.com/help/matlab/ref/polyfit.html,
Accessedon Feb. 1, 2019

[21] H. Limaye and V.V. Deshmukh, "ECG Noise Sources and Various Noise Removal Techniques: A Survey", Int. J. of App. or Inn. in Eng.and Mgt, vol. 5, no. 2, pp. 86-92, 2016.

[22] vHeart, "How to deal with ECG noise", [Online] Available: https://vheart.io/blog/2017/how-to-deal-with-ecg-noise, Accessed

on Feb. 1, 2019.
[23] H. K. Jayant, K. P. S. Rana and et al., "Efficient IIR notch filter designusing Minimax optimization for 50Hz noise suppression in ECG", inProc. ISPCC, Waknaghat, India, 2015, pp. 290-295.
[24] everythingRF, "What is a Notch Filter?", [Online] Available: https://www.everythingrf.com/community/what-is-a-notch-filter, Accessed on Feb. 1, 2019
[25] S. D. Greenwald, "Improved detection and classification of arrhythmias in noise-corrupted electrocardiograms using contextual information", Ph.D. dissertation, Harvard-MIT Division of Health Sciences and Technology, 1990.M. H. Imam, C. K. Karmakar and et al, "Detecting Subclinical Diabetic Cardiac Autonomic Neuropathy by Analyzing Ventricular Repolarization Dynamics", IEEE J. Biomedical and Health Informatics, vol. 20 no. 1, pp. 64-72, 2016.
[26] A. Gacek and W. Pedrycz, "ECG Signal Processing, Classification and Interpretation", Springer, New York, NY, 2012.
[27] A. T. Bhatti and J. H. Kim, "R-Peak detection in ECG signal compression for Heartbeat rate patients at 1KHz using High Order Statistic Algorithm", J. of Multidisciplinary Eng. Sci. and Tech., vol. 2, no. 9, pp. 2509-2515, 2015.
[28] American Heart Association, "All About Heart Rate (Pulse)", [Online], Available: https://www.heart.org/en/health-topics/high- blood-pressure/the-facts-about-high-blood-pressure/all-about-heart- rate-pulse, Accessed on Feb. 1, 2019.
[29] F. H. Bennett and et al., "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming" Artificial Intelligence in Design 96 Springer, Dordrecht. pp. 151-170, 1996.
[30] A. Minchole and B. Rodriguez, "Artificial intelligence for the electrocardiogram", Nature Medicinevolume, vol. 25, pp. 22-23,2019.
[31] E. Tatara and A. Cinar, "Interpreting ECG data by integrating statistical and artificial intelligence tools", IEEE Engineering in Medicine and Biology Magazine, 21:1, pp. 36-41, Jan.-Feb. 2002.
[32] L. Wieclaw, Y. Khoma and et al., "Biometric Identification from Raw ECG Signal Using Deep Learning Techniques", IEEE Proceedings ofIDAACS, pp. 129-133, 2017.
[33] I. Chamatidis, A. Katsika and G. Spathoulas, "Using deep learning neural networks for ECG based authentication,", IEEE Proceedings of ICCST, pp. 1-6, 2017.
[34] S. Patel, "Chapter 2: SVM (Support Vector Machine) -- Theory",Machine Learning 101, [Online] Available: https://medium.com/machine-learning-101/chapter-2-svm-support- vector-machine-theory-f0812effc72, Accessed in Feb. 1, 2019.
[35] H. Drucker, C. J. C. Burges and et al. (1996, December), Support Vector Regression Machines, Advances in Neural Information Processing Systems 9 (NIPS) [Online] Available: https://papers.nips.cc/
[36] P. Gupta, "Decision Trees in Machine Learning", Towards DataScience, [Online] Available: https://towardsdatascience.com/decision-trees-in-machine-learning- 641b9c4e8052, Accessed in Feb. 1, 2019.
[37] J. R. Quinlan, Induction of Decision Trees. Mach. Learn. 1, 1, pp. 81-106, 1986.
[38] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, NY, 2006
[39] M. Gerven1, S. Bohte, "Artificial Neural Networks as Models of Neural Information Processing", Front. Comput. Neurosci., 19 December 2017 [Online] Available: https://www.frontiersin.org/articles/10.3389/fncom.2017.00114/full
[40] L. Yann, "LeNet-5, convolutional neural networks", [Online] Available: http://yann.lecun.com/exdb/lenet/, Accessed on Feb. 1, 2019
[41] W. Yu, K. Yang and et al., "Visualizing and Comparing Convolutional Neural Networks", [Online]. Available: https://arxiv.org/abs/1412.6631
[42] H. Yu, T. Xie and et al., "Comparison of different neural network architectures for digit image recognition", in Proc. IC-HSI, Yokohama, Japan, 2011, pp. 98-103.
[43] S. M. Weiss, I. Kapouleas, "An Empirical Comparison of Pattern Recognition, Neural Nets, and Machine Learning Classification Methods", in Proc. IJCAI, [Online] Available: https://www.ijcai.org/Proceedings/89-1/Papers/125.pdf, Accessed inFeb. 1, 2019.
[44] T. Szandal, "Comparison of Different Learning Algorithms for Pattern Recognition with Hopfield's Neural Network", Procedia Computer Science, vol. 71, pp. 68-75, 2015.
[45] A. Mishra, "Metrics to Evaluate your Machine Learning Algorithm", [Online] Available: https://towardsdatascience.com/metrics-to- evaluate-your-machine-learning-algorithm-f10ba6e38234, Accessed in Feb. 1, 2019.
[46] A. Avati, "Evaluation Metrics", [Online] Available: http://cs229.stanford.edu/section/evaluation_metrics.pdf, Accessed in Feb. 1, 2019
[47] E. L., Lehmann and G. Casella, Theory of Point Estimation 2nd Ed.,Springer, New York, NY, 1998.
[48] ASQ, "What is Statistical Control Process (SPC)?", [Online] Available: https://asq.org/quality-resources/statistical-process- control, Accessed in Feb. 1, 2019.
[49] J. M. Juran, A History of Managing for Quality: The Evolution, Trends, and Future Directions of Managing for Quality, The American Society for Quality Control Press, Milwaukee, WI, 1995.