Demographic Group Prediction Based on Smart Device User Recognition Gestures

Adel R. Alharbi and Mitchell A. Thornton
Department of Computer Science and Engineering
Southern Methodist University, Dallas, Texas, USA
Email: {aalharbi, mitch}@lyle.smu.edu

***Abstract*—We propose a novel demographic group prediction mechanism for smart device users based upon the recognition of user gestures. The core idea of our proposed approach is to utilize data from a variety of the internal environmental sensors in the device to predict useful demographics information. In order to achieve this objective, an application with several intuitive user interfaces was implemented and used to capture user data. The results presented here are based upon the data from fifty users. These captured data are integrated or fused, pre-processed, analyzed, and used as training data for a supervised machine learning predictive approach. The data reduction methods are based upon principal component analysis (PCA) and linear discriminant analysis (LDA). PCA/LDA were implemented to reduce the data feature dimensions and to improve the k-nearest neighbors (KNN) supervised classification predictions. The results of our experiment indicate that high accuracy is achieved from this method. To the best of our knowledge, this is the first research that uses user recognition gestures to predict multiple demographic groups.**

***Index Terms*—Demographic group prediction, machine learning, PCA, LDA, and KNN classification model.**

I. Introduction

MANY smart device applications contain capabilities and features that require the use of private and sensitive user demographic information that is subsequently stored directly on the smart device or within the cloud. The presence of this sensitive data represents a target for malicious exploitation and causes increased security concerns among the user base. As an example, numerous smart device applications rely upon location-based services. Location-based application services allow users to explore, search, and share geographic information with stakeholders. Furthermore, industries or other third-parties can also create and serve customized advertisement services based upon this and other demographic data. However, these services typically require smart device users to register accounts that involve entering personal demographics information to enable sharing their experiences with other new users and for developing friend recommendation services [1]. Due to increasing privacy concerns and generally, the increased awareness among the user base of issues regarding computer security and privacy, applications that require such personal data to be disclosed are often provided with intentional inaccuracies, in a limited fashion (i.e., not allowing the location service to be enabled), or not used at all. The research presented here investigates alternative means for an application to determine user demographic data. Specifically, the extraction of user demographic information from data inherently associated with user gestures of smart devices rather than requiring explicit disclosure is considered. The rich and diverse set of environmental sensors present in modern smart devices serve as generators of data that can be fused into a feature vector and subsequently used in a supervised learning approach to inherently predict the demographic classification of a particular user when explicit demographic information is not provided, or to detect the presence of inaccurate explicitly provided data. This inherently determined user demographic information can also be used to validate or authenticate explicitly disclosed demographic data when present, or to replace such data when it is not present.

Based upon our results, demographic user classes exhibit unique gestures when they use smart device applications. Each gesture carries comprises individual behavior recognition patterns that can be distinguishable from those of a different class. By implementing machine learning (ML) techniques based on the behavior recognition patterns, we created an automatic, intuitive, and effective demographic group predication mechanism. Our mechanism can predict user demographic information such as gender, age, native language, nationality, etc. that can be used to serve several demographic services and goals without breaking requiring a user to enter explicit demographic information into their device.

A. Summary of Previous Work

Our previous work hypothesized that smart device sensors could provide biometric data that could further be used to predict a user's demographic class [2]. The work utilized common sensors available in most devices including the touchscreen and accelerometer. The data collection portion of the study involved the use of an Android-based biometric data collection application that included seven different user exercises. The application requested the human subject to engage in tasks such as entering routine information, re-typing sentences, re-typing random character patterns, reading an article and answering questions about it, drawing on the touchscreen using the user's finger, zooming in/out on a picture, and playing a game that involved spatially reorienting the device. This application allowed us to collect user gestures and store the data on the device's SD card. The data were collected from a single smart device, the AUSU MEMO Pad 7 tablet, to avoid biases among multiple devices. The collected data

were captured in a constrained environment where participants held the device in a fixed position, seated in a chair. Other data collection constraints that were fixed included the screen brightness and keyboard type. The period of time for the data collection exercise was approximately 25 to 30 minutes. During the data collection exercise, the users were under the impression that they were simply evaluating the useability of the device and were informed of the true nature of the study after the data was collected. Demographic truth data were obtained from a consent form that each subject completed prior to using the data collection application. The data collection phase was approved by an Institutional Review Board (IRB) for human subject experimentation at our university. Hence, the subjects had the option to withdraw from the study after they were informed of its true purpose.

This previous work utilized a special biometric approach, a behavioral-based approach because it was hypothesized to be more effective, feasible, and did not require excessive calculations. The idea of the behavioral-based approach was to implement machine learning (ML) methods on the collected data as it was defined in the application without extracting any new features. More explanation of the behavioral-based approach is discussed in section II of this paper. The previous study extracted thirteen features from the data collection application. Table I contains summary descriptions of the features of the keystrokes (F1:F3), the touch-screen features (F4:F9), and accelerometer features (F10:F13). In the Table I, all of the features were continuous variable types except F1 and F4 which were categorical. Furthermore, the previous work used PCA and decision tree (C4.5) classifiers as the ML approach for twenty-two users and predicted five different user demographics yielding promising results and motivating further study as reported here. More detail about the previous study can be found in [2].

TABLE I: Summary of the Extracted Features

F#	Feature	Description
F1	KS_Acts	Indicates pressed key actions.
F2	KS_Code	Indicates pressed key ASCII code.
F3	KS_T	Indicates pressed key time-stamp.
F4	$Touch\_Acts$	Indicates touched finger actions.
F5	$Coor\_X$	Indicates position of x coordination.
F6	$Coor\_Y$	Indicates position of y coordination.
F7	$Touch\_Pr$	Indicates touched finger pressure.
F8	$Touch\_S$	Indicates finger area covered size.
F9	$Touch\_T$	Indicates touched finger time-stamp.
F10	$Assel\_X$	Indicates x axis horizontal movement.
F11	$Assel\_Y$	Indicates y axis horizontal movement.
F12	$Assel\_Z$	Indicates z axis outside movement.
F13	$Assel\_T$	Indicates arm movement time-stamp.

B. Contributions of Our Current Work

The work described here is a continuation of the previous study. This work increased the number of participants up to fifty users thus increasing the number of data samples. It also increased the number of the represented demographic groups up to eight groups. Figure 1 presents the eight demographic groups as bar charts. Each chart shows grouped data with bars that have lengths proportional to the value that they represent. The bar lengths represent a percentage of membership within a particular demographic group. The percentages are calculated based on the total number of users. For example, the gender bar chart has two bars which are the male and female users, and their percentages are 74% and 26%, respectively. Moreover, the names of the eight demographic groups are given in the figure as titles for each chart.

For conciseness, we encoded every user demographic group as D1 through D8 since we have a multiple number of groups as shown in Table II. This Table also provides the cardinality of each of the demographic group, which is equivalent to the number of bars that are presented in Figure 1.

TABLE II: Number/Labels of the Classification Problem

D#	Demographic	Number of the classification problem
D1	Genders	2
D2	Languages	2
D3	Operating System	2
D4	Nationalities	15
D5	Ages	6
D6	Social status	2
D7	Education levels	3
D8	Handedness	3

Finally, the essential contribution of this work is to devise an ML methodology that can predict user demographic information by applying the PCA or LDA data reduction algorithms before implementing the KNN classification model in order to improve the demographic group predictions.

II. Other Related Work

Other work uses data similar to that we use in our work, but for the purpose of human authentication as opposed to demographic classification [1], [2]. Authentication is related to the problem of demographic classification but with several important differences. The relationship between demographic classification and authentication is that in the latter case, the goal is to identify a user as being a member of a set with unity cardinality. That is, the authentication problem attempts to classify a user as being a member of a set containing exactly one element and whereby the collection of all possible sets is disjoint. In contrast, the demographic classification problem classifies users as belonging to a set generally composed of many members and whereby the collection of all possible sets may be non-disjoint. For example, both males and females may also be members in the set of native English speakers. From this point of view, the universe of authentication sets is one where each individual is represented by a set containing one element that uniquely identifies that individual whereas the universe of demographic classification sets is comprised of sets that have multiple elements and whereby an individual may have membership in more than one set.

The demographic group prediction concept can use authentication biometric approaches because it depends on the recognition of smart device user gestures. We can divide biometric approaches into two categorizes: dynamic-based and

behavioral-based approaches. Most of the studies in [1], [3]–[5] concentrate on using the dynamic-based approach. The dynamic approach calculates extracted features and introduces a new set of features based on statistical measures such as mean, median, etc.. However, this approach requires more calculation time for the ML methods. Alternatively, some studies in [6]–[8] use the behavioral-based approach because it is a faster and simpler approach. This approach acquires extracted features and immediately applies them to ML methods. In fact, the study by Lin et al. [6] could be an example for the behavioral-based approach where they extracted data from the touch-screen features and applied the KNN classifier. They achieved accuracy rates of 90.7% with only 30 user touches.

III. General Overview and Goals

In this section, we provide a high-level overview of our general idea and goals of the demographic group prediction mechanism as shown in Figure 2. Users interact with the built-in embedded sensors in the smart device while they accomplish various activities. The sensors capture data during the interaction, which is monitored and collected in the smart device's local database storage. When the data is collected in the smart device database, it can be directly forwarded to the pattern recognition algorithms in order to detect and confirm whether the operation is executed by the appropriate user or not. If the pattern is confirmed, the resident behavioral pattern recognition component computes the difference between the currently collected data and the stored patterns for future analysis and computational purposes. If a pattern is not confirmed, it could mean that a new user pattern has been discovered. The system will update the new pattern and categorize it as belonging to different demographic group or groups. All the stored patterns resulting from the training phase that are present in the smart device database can be retrained to gain the faster access to the applications. However, if there are a large number of patterns, only the most frequently used patterns can be stored in the smart device storage system, while other patterns can be kept in the operators' cloud storage. When the application is enabled to have a real-time internet connection, it can update the patterns automatically by using application background hiding techniques without interrupting the user's usage of a particular application [9].

Using the demographic group prediction mechanism opens a pathway to several goals. First, mobile device or service operators can send suitable categorized advertisements and services corresponding to the appropriate demographic group. Second, some operators or authoritative entities can monitor users in order to identify, classify, or understand other characteristics of a user without prior training data obtained specifically from that user. Thirdly, when a user is entering explicit demographic information, it can be validated by comparing it with the inherently computed demographic information provided by the predictive classifier. Another goal is security. If the operators or some other supervising authority notice any behavioral changes indicating that an unauthorized user may be masquerading as an authorized user, they can take appropriate action. Operators or device owners can also use the proposed mechanism as an authentication technique in a dynamic and continuous manner rather than the more common static technique provided by log-in procedures. Finally, many operators exert effort and spend time collecting customer and employee demographic information in order to make their business investments more suitable for certain demographic groups. This method can be used instead of or to augment the effort and time of this effort by the operators to predict demographic information from the use of customer and employee smart device applications.

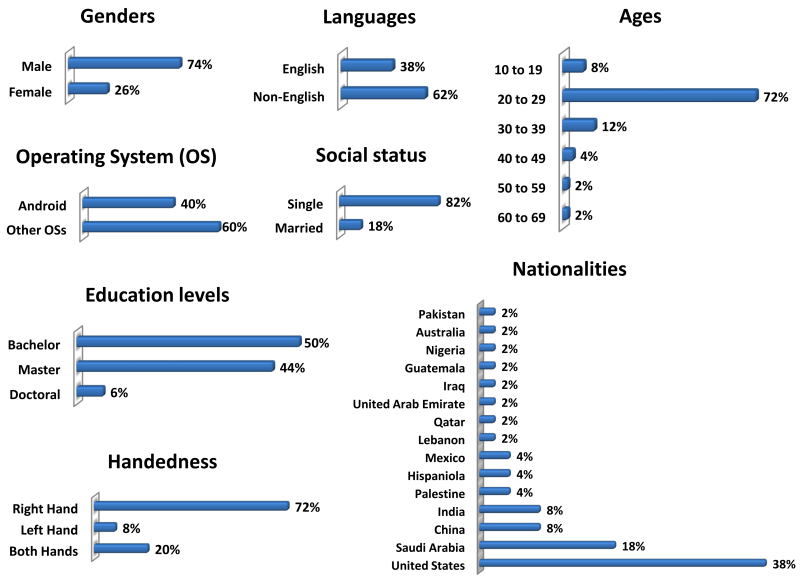


Fig. 1: Subject Demographic Group Histograms

IV. Data Acquisition

The data acquisition process for our study is concentrated on how the participants' application data and demographic group information are gathered. The application is used to collect unsupervised data from our subjects comprised of university students and faculty members. The data collector assigned an individual user identification number (User-ID) to each participant. The User-ID is used to correlate the data collected from the participant's use of the application and their answers to the consent form that included explicit demographic data considered as the "truth" data. In order to enable the supervised learning approach, every User-ID label is included in the application data and the truth data. The consent form included questions, which were as follow: (D1) What is your gender? , (D2) What is your native language? , (D3) What type of operating systems do you use? , (D4) What is your nationality? , (D5) What is your age? , (D6) What is your social status? , (D7) What is the highest education level your have completed? , and (D8) Are you left-, right- ,or both-handed?

V. Methodology

The study's ML methodology was implemented by a statistical and machine learning software tool in the R programming language [10]. This section provides and explains the sequence of the methodology operations in detail.

A. Data Integration

Our study took advantage of using R language to connect, extract, and modify the structured query language (SQL) data tables from the SD card of the device into the R space. There were a total of nine data tables generated from the keystroke, screen-touch, and accelerometer of the application user interfaces where each of those interfaces hold three tables. Our primary objective was to integrate all of the nine tables into one large data table. We called this large data table the multi-feature (MF) table because the gesture data originated from various user interfaces in the application. The integration process was performed in three simple steps. The first step created three biometric tables based on the keystroke, screen-touch, and accelerometer data. After creating these biometric tables, we inserted data rows from each of the nine generated tables into the appropriate biometric tables. The second step was performed to select each user based on the signed User-ID as explained in section IV from the biometric tables. After selecting the user from each of the biometric tables, we could combine the columns of each features and then save the produced user table separately. This step was repeated until all the users were covered. The final step was performed to merge all the user separated files into one large data table. Table III presents an example of the MF data frame with all thirteen features with the correspond User-ID. The Table has some missing values, which are represented with a value of "not available" (NA) due to the lengths of the features being different.

B. Data Pre-processing

This phase involves several operational steps that ensure the clearness and uniqueness of the data features, which are:

1) Numeric Transform: We converted the categorized values to numerical values for the F1 and F4 features to simplify and equalize them with other sets of numerical features [2].
2) Remove Missing Values: Some features contain NA values. If the work converted the NA values to zero or replaced them to their mean values, this would introduce new values especially when we apply the normalization and scaling steps. We chose to remove the rows that

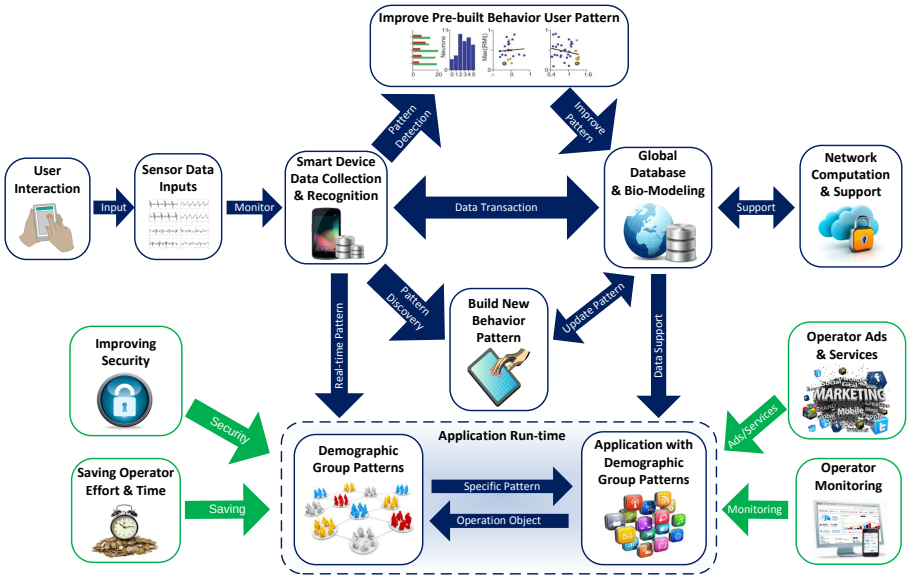


Fig. 2: Demographic Group Prediction Mechanism General Idea and Goals

TABLE III: MF Data Table Example after Implementing the Data Integration Process

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	User-ID
...	...	...	...	...	...	...	...	...	...	...	...	...	...
On	49	34288405	Down	157.0	284.0	0.1425	0.1311	34450086	0.17842	0.35681	9.02457	34450300	2
After	49	34288503	Move	318.0	246.0	0.6151	0.1231	34450162	-0.12418	0.36789	9.08632	34450302	2
Before	49	34288611	Move	337.2	290.0	0.6632	0.1101	34450208	-0.36742	0.46678	9.08974	34450304	2
NA	NA	NA	Cancel	349.0	310.1	0.6781	0.0636	34450316	1.00846	6.65234	0.18907	34450308	2
NA	NA	NA	NA	NA	NA	NA	NA	NA	1.16543	7.79073	0.24657	34450311	2
On	101	34288660	Down	357.0	315.2	0.7511	0.1311	34450354	1.29860	8.80920	0.35650	34450318	3
After	101	34288707	Move	366.0	325.0	0.7536	0.6131	34450377	1.29790	8.80930	0.35650	34450322	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...

contain the NA values. Overall, we omitted less than 3.1% of the data per user based upon this choice [6].

3) Movement Smoothing: This step is an optional step. Accelerometer features (F10, F11, and F12) may have some extreme noise values (i.e., spikes) because users may move the device unintentionally. The solution was to apply a weighted moving average (WMA) filter with a small sliding window of size five to overcome this noise problem [8].
4) Normalization: This step was used to normalize all the features so that they conform to a same scale of distribution. This step used the standard z-score transformation, which subtracts the feature means and then divides them by the feature's standard deviations to have an average of zero and a standard deviation of one [2]. This step is an important step before applying any data reduction methods because it will normalize data variances.
5) Scaling: All features are initially in different scales. They have to be in the same range of zero to one in order to avoid bias and to avoid exceeding unacceptably large values. Hence, the KNN classifier requires the data to be scaled because it depends on distance functions [6].

C. Data Reduction

Smart devices have data storage limitations. The research target is to design a feasible solution with limited data dimensions. The collected data must be as small as possible in order to be stored in the device's memory or the cloud. We chose to use two data reduction techniques, which are:

1) PCA: PCA is a technique that has been used to reduce the complexity of MF data and represent the data within a smaller number of dimensions [11]. It is also a technique that removes redundancy and data that is not significant by retaining only the most significant principal components, which include the maximum variance of the MF data. By implementing the PCA technique in our study, we had the ability to rank the new components and choose only the first ten components and omitted the last three components that have low variance values.

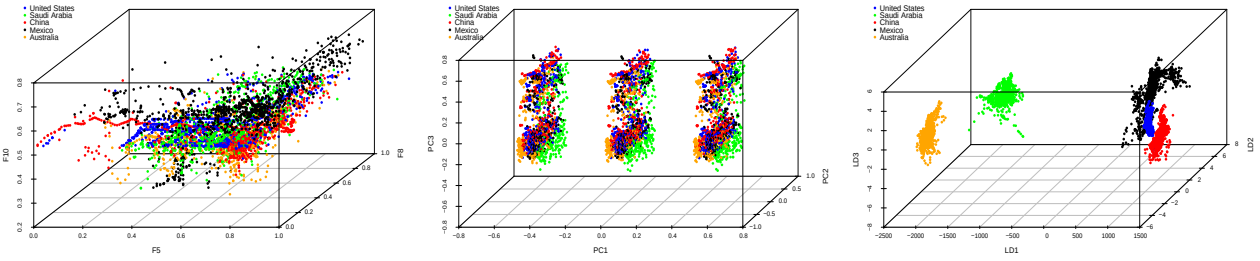
2) LDA: We selected the LDA technique because it reduces feature dimensions based on the provided demographic class labels. LDA separates the MF gaussian distributions linearly. It also discovers a linear combination of data features that best separate two or more classes of targets [5]. Therefore, the LDA technique produces a new vector space whose dimension is, at most, equivalent to the same number of dimensions of the original MF. The reason for using the new vector space is that instances corresponding to same demographics are grouped as close as possible and projected as far as possible from instances of other user demographics [12].

The primary reason for choosing both the PCA and LDA techniques was due to similar studies [3], [5] that indicated that PCA/LDA improve the classification method performance without losing useful information. Many past classification algorithms also have utilized these techniques to design specific classification systems that are used in the biometric applications. In our case, PCA and LDA techniques have been used to discover the hidden patterns and extract new dimensions that are smaller than the original data. The PCA technique performs as an unsupervised transformation (ignores class labels), while LDA technique is a supervised transformation. In order to illustrate the differences between the PCA and LDA techniques in our MF space, we visualize them using a three-dimensional (3-D) scatter plot. We picked five user nationalities: Untied States, Saudi Arabia, China, Mexico, and Australia. Each nationality is represented with different colors: blue, green, red, black, and orange, respectively. Figure 3 shows three 3-D scatter plots before implementing the PCA and LDA techniques and after. Plot (a) represents three selected features (F5, F8, and F10) where the points are distributed randomly in the space. Plot (b) presents the three highest principle component dimensions, where the points are distributed without considering the nationality classes. Plot (c) presents the three linear discriminant dimensions, where the points are distributed with consideration of the nationality classes.

D. Supervised Classification

We chose the KNN supervised classification model for several reasons. One reason is that it is a robust classifier with respect to the noise that might exist in the feature data since it is based on the distance measurements. Second, it provides fast classification and demographic predictions. Finally, it was chosen regrading to the explained related study by Lin et al. [6] in section II, where our approach is similar to their.

The KNN classifier uses the local neighborhood between each new observation (here, the demographic observation) and projects the observation into the feature space with respect to the training observations. The local neighborhood is based on distance measurements to compute the similarity of the



(a) Original Features (b) PCA (c) LDA

Fig. 3: Three 3-D Scatter Plots for 5 User Nationalities Before and After Applying PCA and LDA

nearest observation neighbors. We used the L_2 Euclidean distance measurement because it is fast, simple, and yields good performance in many smart device biometric projects [5]. The classifier has the ability to memorize the K training observations that are more similar to the new observations. Hence, it chooses the label that has the majority of the K closest training observations. This classifier, in essence, stores all training observations and labels, which may be a limitation where large data sets are used. In our case, this limitation is not a problem because our approach initially uses either PCA or LDA data dimensionally reduction techniques to store fewer dimensions than those present in the initial demographic observations. In our study, we use a value of K equal to the number of the classification problems as discussed in Table II.

VI. Multi-Feature Data Analysis

The total number of observations were 45,163, where each participant who used our application generated more than 868 observations. In this section, we focus on providing some basic analysis on the MF data.

The MF Normality: The study applied the shapiro-wilk and anderson-darling normality tests [13]. The tests resulted with p-values of less than $2.2e^{-16}$ for all the thirteen features. These very small p-values tell us that the data is not normality distributed. We reject the null hypothesis of normality.

The MF Ranking: This study uses the f-selector function that chooses the best combination of features based on information gain theory of the feature's entropy [14]. The reason for choosing is because the weights assigned by other classification methods are different, which might effect the ranking process. The study's primary purpose for using this technique was to sort the features in descending order and discover the best feature combination. Table IV shows the results of using the method for the first five demographics and provides the best MF combinations, where the best feature is on the right and the least contributing feature is on the left. For example, in D1 the best feature is F3 and the least contributing feature is F4. By considering this set of demographics, the timing features (F3, F9, and F13) are ranked as the best. It is also the case that each demographic has its own combinations.

The MF Relationships: In order to understand the MF relationships and which features provide hidden information,

TABLE IV: Ranking of Features for the First 5 Demographics

D#	The best feature ranked combination
D1	F3+F9+F13+F12+F11+F10+F6+F7+F8+F5+F1+F2+F4
D2	F3+F9+F13+F11+F12+F10+F6+F5+F7+F8+F2+F4+F1
D3	F3+F9+F13+F11+F12+F10+F6+F5+F7+F8+F4+F1+F2
D4	F9+F3+F13+F11+F12+F6+F10+F5+F7+F8+F2+F4+F1
D5	F3+F9+F13+F11+F12+F6+F10+F5+F7+F8+F4+F2+F1

we depict the correlation coefficients [15]. Figure 4 shows the correlation coefficients of all pairs of the features in a color-coded plot. White squares indicate that the feature pair is not correlated, blue indicates a positive correlation, and red indicates a negative correlation. On the right side of the correlation matrix plot, the legend color scale ranged from -1 to 1 which indicates the correlation coefficient ranges with the corresponding color codes. The darker a color is, the larger the correlation coefficient between the feature pairs is. With this plot, one feature can be highly highly correlated with other features. For example, feature F3 is highly correlated with features F9 and F13 that because they are timing based features.

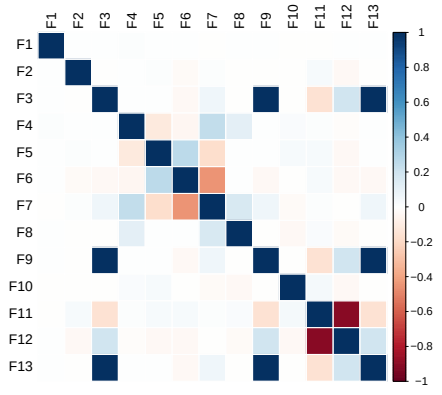


Fig. 4: Correlation Matrix for all Features

VII. Experimental Results

We carried out three different experiments for the proposed approach. We used the $cross-validation$ strategy to validate all of our experimental results. This strategy was based on ten-fold cross validation repeated ten times.

A. Model Validation and Comparison Results

In this experiment, we validated the KNN classifier and achieved the accuracy rates in situations where the classifier was implemented without PCA/LDA data reduction and also when PCA and LDA techniques were implemented before classification, as shown in Table V. This Table indicates that the PCA technique achieves roughly the same accuracy rates of the classifier from D1 through D6, while it improves the classifier accuracy rates in D7 and D8. The LDA technique decreases classifier accuracy rates for all the demographics except in D4 and D5.

TABLE V: Comparison of Accuracy Rates

D#	KNN	PCA+KNN	LDA+KNN
D1	86.65%	86.57%	65.45%
D2	85.71%	85.62%	63.02%
D3	83.00%	82.78%	57.53%
D4	72.40%	72.29%	82.18%
D5	84.89%	84.81%	92.00%
D6	91.88%	91.81%	78.45%
D7	81.53%	82.43%	62.65%
D8	83.69%	85.13%	70.70%

Based on the comparison of the accuracy rates of Table V, PCA and LDA techniques improve the KNN classifier to the extent that they have significant accuracy rates for all of demographic groups with limited data dimensions. These results guided our decision to implement either the PCA or LDA technique on the demographic groups where the accuracy rates were improved. The PCA technique was implemented for D1, D2, D3, D6, D7, and D8. The LDA technique was implemented for D4 and D5.

B. Performance Results for Different Number of Users

In order to gain more insight into the impact of the number of users on our approach performance, we observed how the accuracy of demographic predictions are influenced by increasing the number of subjects [4]. This experiment selected the first three demographic groups (D1, D2, and D3) to record the accuracy rates for each different number of user data sets. We began the experiment with five subjects in order to have enough data samples to validate. Then, we increased the number of subjects by five until we reached the total number of available subject data samples. For each increasing number of subjects, we performed cross-validation, as explained previously, to produce the accuracy rates as it shown in Table VI. This Table shows that the accuracy rates decrease gradually when we add more users.

C. Validation Results

We obtained validatation information by using the biometric performance measurements in [16]. The measurements were based on the receiver operatic characteristic (ROC) curve, also known as the area under the ROC curve (AUC). The AUC demonstrates the classification model's ability to differentiate between positive and negative classes. The AUC relies on two performance measurements: sensitivity and specificity. Sensitivity measures the number of samples from the positive (first) class that were truly predicted correctly. Specificity measures the number of samples from the negative (second) class that were truly predicted correctly. A perfect classification model that has a high sensitivity, specificity, and AUC measurement rates is 100% if the model has made all predictions perfectly. If the model results in rates closer to 50%, then it is no better than a random guess.

TABLE VI: Accuracy Rates Versus Number of Subjects

Number of subjects	D1	D2	D3
5	92.07%	93.77%	90.28%
10	94.57%	97.48%	91.42%
15	90.64%	91.02%	87.86%
20	88.98%	89.39%	87.36%
25	86.12%	87.31%	87.01%
30	87.38%	87.92%	87.46%
35	87.00%	88.55%	87.10%
40	86.25%	86.54%	84.16%
45	86.58%	87.12%	83.68%
50	86.57%	85.62%	82.78%

The ROC or AUC measures are only fit binary classification problems. However, our study has multi-classification problems for some demographics (D4, D5, D7, and D8) as in Table II. Thus, we used only the two highest group percentage classes from those demographics. For example, in D4, the study considered only the United States and Saudi Arabia user classes because they comprised the highest number of subject data samples. It is noted that we changed the value of K in the KNN classifier to make it binary (two). Table VII shows the high performance measurement rates of the AUC, sensitivity, and specificity for the proposed approach.

TABLE VII: The AUC, Sensitivity, and Specificity Rates

D#	AUC	Sensitivity	Specificity
D1	89.37%	75.58%	90.73%
D2	90.99%	81.34%	88.29%
D3	88.44%	76.96%	86.51%
D4	96.51%	90.93%	96.56%
D5	94.22%	97.62%	83.42%
D6	90.51%	74.55%	95.38%
D7	89.43%	84.18%	81.86%
D8	87.60%	70.39%	91.82%

VIII. Limitations and Potential Extensions

A. Limitations

This study was based on collecting data from a single tablet device. Due to the differences that exist from one device to another, the study should consider collecting data samples from multiple smart devices in order to capture variety of data. This includes both different devices of the same model where variation among the internal sensors may be present as well as the use of different tablet models. In terms of using multiple versions of the same device, variation of accuracy rates with respect to different tolerances present in the internal sensors could be established. In terms of using different models and manufacturers of smart devices, an opportunity to compare and identify the type of device that the user is using could be studied. Is it a phone or tablet? Which version of device is it?

Another limitation is that the study only considered collecting the user gestures when the subject utilized the data collection application for the first time. As a subject gains familiarity with a particular application, their corresponding gesture behavior will likely change. A study where data is captured from subjects that gain increasing familiarity with a given application would allow for characterization of how captured gesture data evolves with respect to application familiarity [3].

B. Potential Study Extensions

Our approach is based upon integrating or fusing three different commonly used device sensors. One potential extension is to utilize additional sensors such as the gyroscope, magnetometer, heart rate, and other sensors that are becoming common in modern devices. By adding more sensor data to our initial collection phase, we could extract more features into the MF data space which may increase the predication accuracy results. For example, the heart rate sensor may be useful for age demographic because younger users usually have a higher heart rate than older users.

Our demographic group prediction method depends upon categorizing users into groups. In our study, we categorized subject demographics based on samples from a university population and we used a predetermined set of demographic group cardinalities. Hence, the categorizing mechanism may be different in other domains where other subject populations are included. Some populations may include more subjects with lower educational levels or different percentages of native languages. In terms of set cardinality, it may be desirable to use different cardinalities associated with each demographic group. For example, it may be a desire to categorize subject gender with respect to a set of cardinality three such as male, female, and others. Categorizing additional user groups necessarily results in adding additional classification problems. Our approach yielded results indicating that having a large number of the classification problems may not be a drastically limiting constraint. This is because our approach applies the aforementioned data reduction mechanisms (i.e., PCA and LDA) to reduce the subject group data dimensionality to support the classifier performance. For example, D4 resulted in fifteen classification problems, and the LDA technique improved the classifier accuracy rates from 72.40% to 82.18% as shown in Table V.

IX. Conclusion and Future Work

In this investigation, we describe an effective mechanism for demographic group predictions for smart device users based on the use of user gestures. An advantage of this approach is that it does not require any additional specialized hardware since it depends only on the sensors already present in smart devices. Our approach has many potential applications ranging from serving customized advertisements or other data, continuous and non-obtrusive authentication, validation of explicitly entered demographic data, and others. Our ML approach integrates the data from multiple on-board sensors, pre-processes and reduces the data into a smaller-dimensioned space using PCA or LDA, and then performs supervised demographic classification. Experimental accuracy rates indicate that the best results are achieved by implementing either the PCA or LDA technique for different demographic groups prior to invoking the KNN classification method. The achieved accuracy rates were all greater than 80% for the demographics.

In future work, we plan to continue capturing more subject data samples to refine our results and observe the accuracy rate versus the number of users. We will also continue to discover more demographic groups that may be discernable using our approach. Our future work will also consider using more and different statistical hypothesis tests [15]. We also plan to investigate the use of alternative classifiers such as use of a decision tree or other classifiers to determine performance versus accuracy trade-offs regarding ML approaches for demographic group predictions.

References

[1] J. J.-C. Ying, Y.-J. Chang, C.-M. Huang, and V. S. Tseng, "Demographic prediction based on users mobile behaviors," *Mobile Data Challenge*, 2012.

[2] A. R. Alharbi and M. A. Thornton, *Demographic Group Classification of Smart Device Users*. IEEE International Conference on Machine Learning and Applications (ICMLA), 2015.

[3] Z. Sitová, J. Šeděnka, Q. Yang, G. Peng, G. Zhou, P. Gasti, and K. S. Balagani, "Hmog: New behavioral biometric features for continuous authentication of smartphone users," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 5, pp. 877–892, 2016.

[4] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Cell phone-based biometric identification," in *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*. IEEE, 2010, pp. 1–7.

[5] I. de Mendizabal-Vzquez, D. de Santos-Sierra, J. Guerra-Casanova, and C. Snchez-vila, "Supervised classification methods applied to keystroke dynamics through mobile devices," in *Security Technology (ICCST), 2014 International Carnahan Conference on*, Oct 2014, pp. 1–6.

[6] C.-C. Lin, C.-C. Chang, and D. Liang, "A novel non-intrusive user authentication method based on touchscreen of smartphones," in *Biometrics and Security Technologies (ISBAST), 2013 International Symposium on*. IEEE, 2013, pp. 212–216.

[7] A. A. E. Ahmed and I. Traore, "A new biometric technology based on mouse dynamics," *IEEE Transactions on dependable and secure computing*, vol. 4, no. 3, p. 165, 2007.

[8] M. Muaaz and C. Nickel, "Influence of different walking speeds and surfaces on accelerometer-based biometric gait recognition," in *Telecommunications and Signal Processing (TSP), 2012 35th International Conference on*. IEEE, 2012, pp. 508–512.

[9] K. W. Nixon, X. Chen, Z.-H. Mao, Y. Chen, and K. Li, "Mobile user classification and authorization based on gesture usage recognition," in *Design Automation Conference (ASP-DAC), 2013 18th Asia and South Pacific*, Jan 2013, pp. 384–389.

[10] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[11] I. K. Fodor, "A survey of dimension reduction techniques," 2002.

[12] M. Welling, "Fisher linear discriminant analysis," *Department of Computer Science, University of Toronto*, vol. 3, 2005.

[13] N. M. Razali, Y. B. Wah *et al.*, "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests," *Journal of statistical modeling and analytics*, vol. 2, no. 1, pp. 21–33, 2011.

[14] T. Cheng, Y. Wang, and S. H. Bryant, "Fselector: a ruby gem for feature selection," *Bioinformatics*, vol. 28, no. 21, pp. 2851–2852, 2012.

[15] T. Neideen and K. Brasel, "Understanding statistical tests," *Journal of surgical education*, vol. 64, no. 2, pp. 93–96, 2007.

[16] M. Kuhn, "Caret package," *Journal of Statistical Software*, vol. 28, no. 5, 2008.