**Alcatraz, an end-to-end SaaS product for complete computer and network security**

**Dr. Annapurna P Patil J Geetha Reddy**

Professor, Assistant Professor,

Department of Computer Science and Engineering, Department of Computer Science and Engineering,

Ramaiah Institute of Technology Ramaiah Institute of Technology

Bangalore - 560054 Bangalore - 560054

**Ayush Kumar Bineet Kumar**

Student, Department of CSE, RIT Student, Department of CSE, RIT

**Aman Mehta Deepak Jayaprakash**

Student, Department of CSE, RIT Student, Department of CSE, RIT

***Abstract--***The demand for modern tools and techniques to restrict access to applications and services which contain delicate data is increasing exponentially each year. People can no longer rely on older methods of computer and network security as they fail in dealing with the challenges posed.

The product is aimed at providing an architecture and open source code to developers so that they can embed these features into their applications to enhance the security. The services provided are top notch and cover the broad spectrum of computer and network security. All the features of the product involve the application of **Data Mining** and **Machine Learning** techniques onto the domain of **Computer Security**.

The most prominent features of the product are listed below:

- **Keystroke Biometric Authentication**: Ensembles result of 3 statistical models and classifies user based on Dynamic Thresholding

- **Network Intrusion Detection:** Data mining on NSL-KDD data set, build 3 models, accuracy comparison and data visualization.

- **Anti-key logging:** Uses authentication based on GET request exchange and adds a double layer of AES + Hashing to guarantee both Data Integrity and Security.

- **O-Auth with Hashing:** Traditional O-Auth modified with additional layer of Hash to impart Digital Signature.

The complete product with the features is based on dynamic-dispatch, a kind of “on-demand software” inspired ideology, hence the term SaaS. The entire product is released as open source with detailed documentation on GitHub. It is released as a stable live version with a guide for developers on how to contribute.

*Keywords: Keystroke dynamics, zero variance removal, O-Auth, Network Intrusion Detection, Anti-key logging, Statistical modeling*

1. INTRODUCTION

The biggest necessity of a product of this nature arises from the fact that developers prioritize attributes such as functionality, UI, memory and efficiency while designing and creating applications of both small and large scale. But they comparatively turn a blind eye or assign a low level priority to security. This is understandable because making an application/ service secure requires a lot of programming effort and expertise. The objective of the product is hence to provide an overall architecture, design and code to application developers so that they can embed these onto their modules.

The features are also on a demand and get basis so only those services which are required can be taken. Each feature is an application of statistical modeling and ML algorithms on layers of computer security. The services use highly efficient mechanisms ranging from Dynamic Thresholding, Zero-variance removal, AES, Hashing, Data Mining to Ensemble learning, asynchronous communication and accuracy analysis.

The entire product is versioned and released on GitHub. Each feature is a novel implementation because all of them are not a single model-classifier service. They are an ensemble learnt multi classifier accumulation, wherein each classifier/ statistical model is designed to outperform one another so that the drawbacks of one are overcome in another thereby leading to a more accurate result. Enough mathematical support is given to all the algorithms used. Some of the concepts used are linear programming, statistical mean, median, variance, k-cross validation, iterative averaging, etc.

1. LITERATURE SURVEY

The features offered in the product belong to well researched and hot topics lying in the domains of Biometrics, Data Mining, Machine Learning, Statistical modelling and Computer/ Network Security. Prominent work has been done in the field of keystroke biometrics. This is because of the rising significance in the area of digital signatures, data integrity and security. Many existing approaches implement the static text method of keystroke dynamics [5]. The existing live implementations for “free text” are very hard to find, which can be attributed to the low accuracy rate and other problems like bias, over-fitting, false recognition rates and non-significance

First true work on keystroke dynamics originated from Monrose and Rubin [15]. Their learning models were implemented by using statistics like digraph/ trigraph mean and variance

Researchers have presented their work on choosing different distance metrics ranging from Manhattan to Euclidian and combined them with mean, median and tested their suitability on the biometric authentication. For the implementation both traditional and modern classifiers have been used, including knn[5], Bayes Point Machines, K-mediods/means methods [11], genetic algorithms, neural networks , and SVMs if the attributes are non linearly dependant.

Network Intrusion Detection happens to be one of hot topics in the domain of Network (both wired and wireless) security. Prominent data mining algorithms like Naïve Bayes, Decision Tree, K-nearest neighbor, K-Means and Fuzzy C-Mean clustering algorithms and machine learning strategies like Support Vector Machine (SVM), neural nets have been applied to study the classification. The existing work can be broadly classified into:

- Machine Learning Based IDS

- The Multiple Classifiers (MCS)

- Boosting based IDS

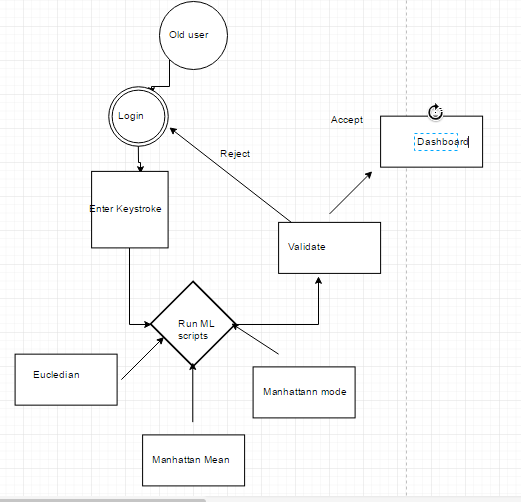
- Data Mining and Rule based IDS

There has been quite few implementations of Anti-Key logging systems and have been offered as service or an application supporting some Operating Systems. The research done on this include, Behavior based detection technique using KLIMAX (Kernel- Level Infrastructure for Memory and execution profiling), Anomaly Based Detectors, Signature based detectors, Subnet Mask Scanners and Host Analyzers.

III. PROPOSED SYSTEM

The problem with the existing software and services available is that they are not bundled properly. The right combination of features and a documented source of reference are nowhere presented. Therefore we propose a complete encapsulated bundle of features and open source code to make integration of security modules much easier.

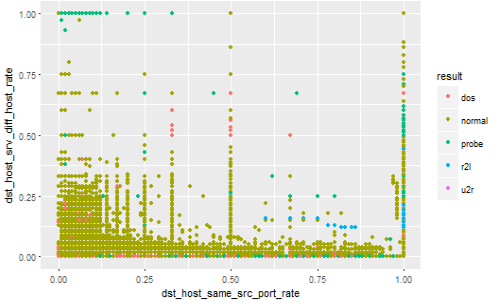
The design of one of the modules is shown in figures **Fig 3.1.**



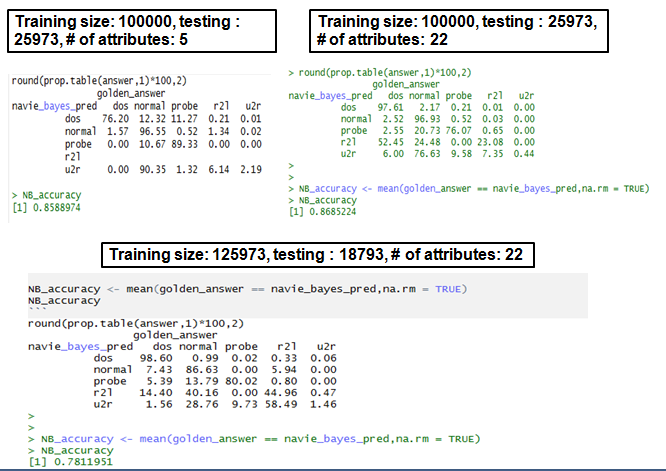
**Fid 3.1: Design of Keystroke Authentication**

Some key milestones in the architectural design:

* **Data and feature extraction:** In keystroke authentication it corresponds to the derivation of tuples taken from the user to train the ML models. In NID it maps to the labeling of data, attribute selection, zero-variance removal, dimensionality reduction, etc.
* **Pre-processing and metric selection:** This includes selection of statistical parameters like Manhattan, Mahalanobis, Euclidean, etc. The plotting of attributes for k-cross comparison to analyze the interdependencies between them is also done in this phase. Fig 3.2 shows one of the plots.
* **Building Classifiers (statistical and DM/ML models):** The heart of the product lies in this stage. This involves using multiple algorithms like Naïve Bayes, Decision Trees and Random Forests for NID and Manhattan mean, Euclidian mean and Manhattan median for Keystroke Authentication.
* **Accuracy Comparison and Analysis**: The idea of building multiple models is that one can overcome the drawbacks of the other, then ensembling it into one cumulative result. A set of accuracies arise due to the combination of number attributes, data size length, etc. One set of accuracies are as shown in Fig 3.3
* **Visualization and version control:** The last phase involves data visualization and handling multiple versions of the code and documentation.



**Fig 3.2: Plot to analyze interdependency between attributes**



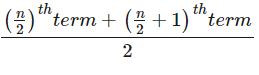
**Fig 3.3: Accuracy comparison for one of the models**

***3.1 Metrics for statistical models***

Distance measurements are in Manhattan and Euclidian statistic. Manhattan can be expressed as the modulus of the absolute difference between two points in space. Euclidian is the RMS distance of two points.

The arithmetic metrics used to train models is on mean and median. Equations (1) and (2) represent the formulae to calculate the mean and median of n points. The points that are being referred to are the data points for the training set. Assume that the space is made up of 6 data points each of them which would stand for a set of attribute array in our database. Now pick a data point a random and calculate the distance of this particular point from each of the rest of data points.

That would give the threshold that must exist for this data point. Now choose another data point and repeat this calculation. At the end we would end up with six different difference vectors. Now take the average of the vectors and conclude to a single point in space. This point is cumulative distance equivalent of all the vectors combined.

….(1) ….(2)

***3.2 Algorithmic flow for modified O-Auth and Anti-key logging***

The architecture is divided into client and server side protocols. The host will a set S of forms in N number of web pages. Only s subset s of S will need anti key logging. This information will be stored in a JSON document maintained at the server of the form

{ form\_name: [Name],

isAuthenticationRequirred:[True/False]

}

The steps flow as shown below:

i. For the set {S-s} proceed with the default mechanism.

ii. For the set s which requires Anti-Key logging, capture all the form elements and encrypt them using the standard AES encryption algorithm.

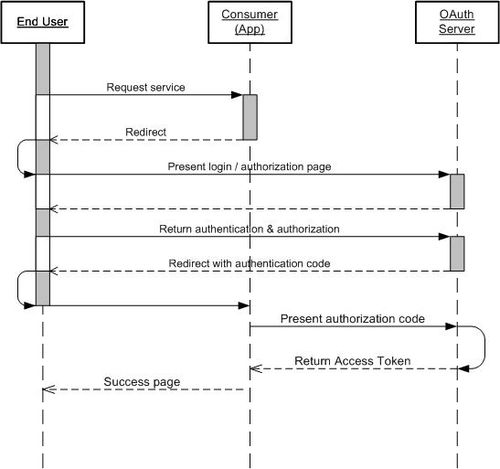
iii. The trick to the anti key logging is the key used in AES. Here, the key is dynamically generated by hashing a random r out of the R form elements.

iv. In OAuth terms a GET request is sent to the server, with a [client\_id, {subset of form elements}]. The server has a has function which replies with a hashCode.

v. This value is passed as the key to AES. This guarantees data integrity.

|  |
| --- |
|  |
|  |  |

Fig 3.4 shows the sequence diagram of the flow of communication.



**Fig 3.4: Sequence diagram for Anti-Keylogging and modified OAuth**

1. IMPLEMENTATION

The product follows a factory design pattern with very systematic modular oriented highly coherent features. It consists of 5 major modules namely:

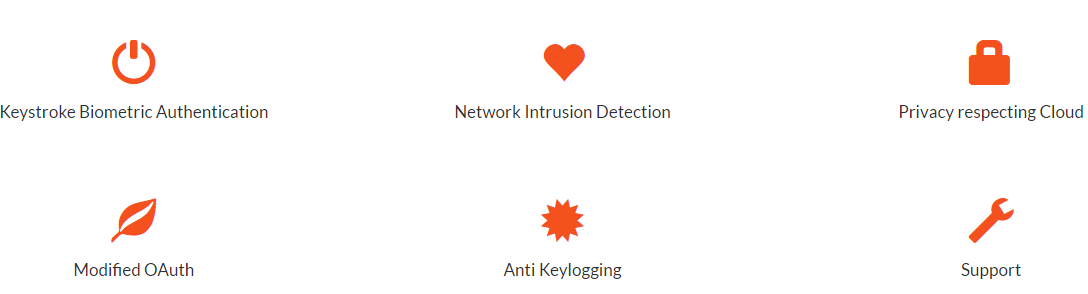
i. **DB module**: MySql, JSON in conjugation with PHP. Responsible for maintaining user profiles, keystroke timing features, .csv to data frame transformation, etc

ii. **UI module:** Involves data visualization for decision trees and plotting attributes, along with Dashboard for each feature, front-end JS validations, etc

iii. **Back-end scripts:** Profiling user information, connecting DB with front-end, asynchronous communication for GET, POST requests, etc

iv. **Machine Learning module:** The three classifiers for keystroke authentication belong in here. It also involves the Ensemble algorithm.

v. **Data Mining module:** The classifiers used in Network Intrusion Detection along with dimensionality reduction, pre-processing and building models.



**Fig 4.1: Services offered in the product**

Fig 4.1 shows the features offered which is listed on the landing page of the product. The implementation for some of the features is explained in detail.

**4.1. Keystroke Biometric Authentication:** The three statistical models implemented here are

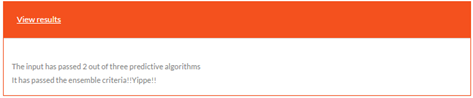
-***Manhattan Mean***: Simple, not robust, not adaptive

-***Euclidian Mean***: Adaptive, dynamic Thresholding, not robust

- ***Manhattan Median***: static threshold, not adaptive, robust

Fig 4.2 shows the Dashboard with information about the algorithms and test case result.





**Fig 4.2: Dashboard and test case results**

**4.2** **Network Intrusion Detection**

This feature has the following steps explained briefly

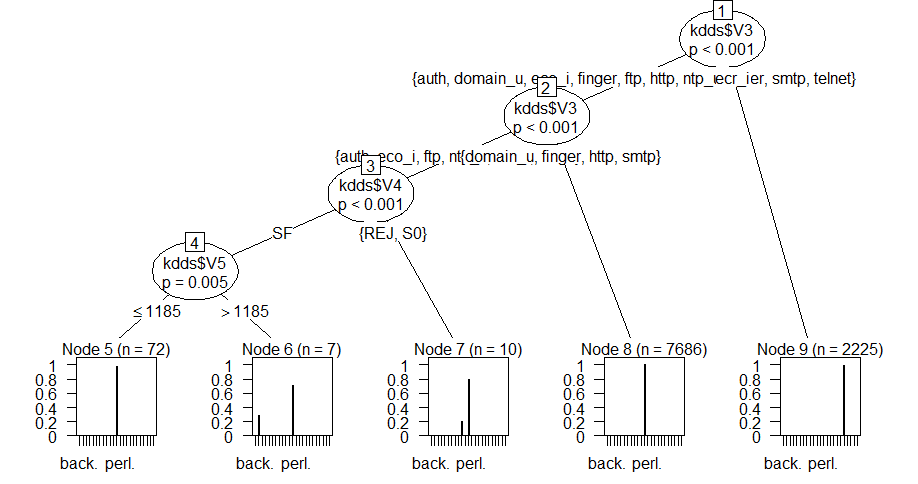
i. ***Setting up of workspace and data:*** The NSL-KDD has a total size of 125000 tuples when compared to the original KDD set which had around 250000 tuples in the 20%subset

ii. ***Pre-processing and attribute selection*:** This involves label the attacks into 4 main classes and also to normalize some of the attributes to the range of 0-1. It also has zero-variance removal which reduces attribute size almost by 70%.

iii. ***Multi-classifier building:*** 3 models Naïve Bayes, Decision trees and Random Forests are built with bagging.

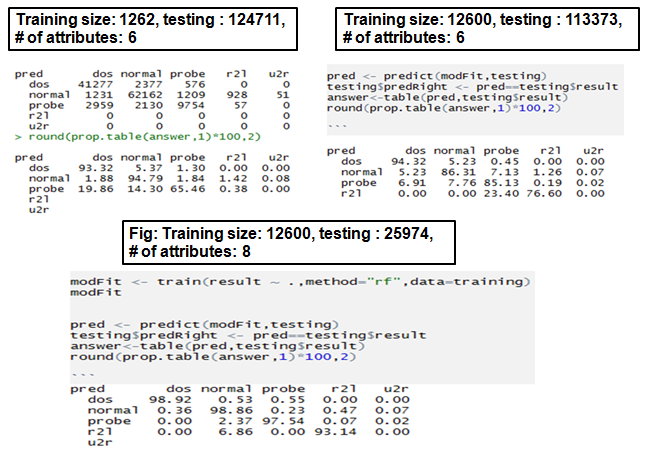
iv. ***Accuracy Analysis:*** The test result are compared with labeled set and correct combination of number of attributes and data set size is derived.

Fig 4.3 shows one of the decision trees plotted.



**Fig 4.3: Decision Tree with 8 attributes**

The challenge lies in finding out the correct attribute size and data size because there exists a trade-off between time taken to build the models and accuracy. Fig 4.4 shows the accuracies for different combination for Random Forests, as the #attributes and data set increases, the time taken will grow exponentially. The average accuracy of all the models included came out to be around ~94.5%.



**Fig 4.4: Accuracies for Random Forests model**

The algorithmic flow for modified O-Auth and Anti-Key logging is explained in section 3.2 itself.

1. COMPARISON WITH EXISTING WORK

The product when compared to existing software applications like Snort, SpyShelter and other biometrics have advantages like

- No external hardware

- One complete encapsulated bundle

- Open source code

- Cloud Integration

- Detailed documentation and support

Significant progress and research has happened in the field of Keystroke Biometrics. Killourhy and Maxion used 14 keystroke dynamics anomaly detector to authenticate users, 11 of which were previously proposed, and 3 where classic recognition patterns using various distance statistic models. (Euclidean, Manhattan, and Mahalanobis distance measures). The features from each sample included delays and latencies mentioned above and achieved and EER of 9.6%.

Our model used two distance statistics Manhattan distance and Euclidean distance. Our dataset composed of 40, who typed the password for 6 times and authenticated using the model for 21 times during 3 sessions.

Prominent data mining algorithms like Naïve Bayes, Decision Tree, K-nearest neighbor, K-Means and Fuzzy C-Mean clustering algorithms and machine learning strategies like Support Vector Machine (SVM), neural nets have been applied to study the Network Intrusion Detectors.

But there has been no detailed comparison of accuracies and multi classifier approach in Network Intrusion Detection, so our work would be the first to analyze 3 different models and derive qualitative conclusions.

1. CONCLUSION AND FUTURE WORK

The product provides you with different features which essentially provide end to end computer, network and web application security. We applied new combinations of distance metrics and learning models and there is existed a few anomalies and false predications. This can be attributed to the problem of over fitting the data onto the model. But some of the models out-performed the existing keystroke dynamic classifiers which use traditional algorithms and distance metrics.

The pre-processing, construction of classifiers comparison, analysis gave us some insights. The redundancy can be eliminated by finding out the zero-variance features and not is including them in the training set as it leads to biased results.

Once threshold efficiency reached, any increase in the training set would only over-fit the data. Therefore the results would tend to deteriorate if training set is kept increasing.

With respect to Keystroke Dynamics, we analyzed the existing attributes and algorithms and came up with a new adaptive model and statistic which would change the threshold according to the user’s dissimilarity in his/her typing patterns. Since we have eliminated the outliers, made the models adaptive, and shown different algorithms implementations, the models proved to be more accurate than existing models.

The problems and short comes in the existing features can be attributed to the variations in the data and existence of outliers. Another possibility is that the degree of ensembling can be analyzed and fixed to a value that avoids over-fitting of data.

The O-Auth can be improved and the GET request exchanges can be made asynchronous. The communication is usually supported across cross domains, which means developers face a difficulty called CORS(Cross Origin Resource Sharing). This can be fixed via a third party gateway.

***References***

[1] Araújo, Lívia CF, Luiz HR Sucupira, Miguel Gustavo Lizarraga, Lee Luan Ling, and João Baptista T. Yabu-Uti. "**User authentication through typing biometrics features**." *IEEE transactions on signal processing* 53, no. 2 (2005): 851-855.

[2] Bergadano, Francesco, Daniele Gunetti, and Claudia Picardi. "**User authentication through keystroke dynamics**." *ACM Transactions on Information and System Security (TISSEC)* 5, no. 4 (2002): 367-397.

[3] Cho, Sungzoon, Chigeun Han, Dae Hee Han, and Hyung-Il Kim. "**Web-based keystroke dynamics identity verification using neural network**" *Journal of organizational computing and electronic commerce* 10, no. 4 (2000): 295-307.

[4] Zhong, Yu, Yunbin Deng, and Anil K. Jain. "**Keystroke dynamics for user authenticatio**." In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 117-123. IEEE, 2012.

[5] Leggett, John, and Glen Williams. "Verifying identity via keystroke characterstics." *International Journal of Man-Machine Studies* 28, no. 1 (1988): 67-76.

[6]Kevin S. Killourhy and Roy A Maxion, **“Comparing Anomaly-Detection Algorithms for Keystroke Dynamics** Proc. of the 3rd Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing, pp. 61-64, 2007.

**[7]** Xu, Xin. "Adaptive intrusion detection based on machine learning: feature extraction, classifier construction and sequential pattern prediction." International Journal of Web Services Practices 2, no. 1-2 (2006): 49-58.

[8] Obaidat, Mohammad S., and Balqies Sadoun. "**Verification of computer users using keystroke dynamics."***IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 27, no. 2 (1997): 261-269.

[9] Ajith Abraham, Ravi Jain, Johnson Thomas and Sang Yong Han, “D-SCIDS: Distributed soft computing intrusion detection system”, Elsevier, Journal of Network and Computer Applications, Vol.30, No.1, pp. 81–98, 2007..

[10] Monrose, Fabian, Michael K. Reiter, and Susanne Wetzel. "**Password hardening based on keystroke dynamics.**" In *Proceedings of the 6th ACM Conference on Computer and Communications Security*, pp. 73-82. ACM, 1999.

[11] Li, Yilin, Baochang Zhang, Yao Cao, Sanqiang Zhao, Yongsheng Gao, and Jianzhuang Liu. "**Study on the BeiHang keystroke dynamics database.**" In *Biometrics (IJCB), 2011 International Joint Conference on*, pp. 1-5. IEEE, 2011.

[12] S. Prabhakar, S. Pankanti, and A. K. Jain, “**Biometric Recognition: Security and Privacy Concerns**”, IEEE Security and Privacy Magazine, Vol. 1, No. 2, pp. 33-42, 2003.

[13] T. Sim and R. Janakiraman, “**Are digraphs good for freetext keystroke dynamics?** ”, IEEE CVPR, pp. 17-22, 2007.

[14] Davoudi, Homa, and Ehsanollah Kabir. "**A new distance measure for free text keystroke authentication."** In *Computer Conference, 2009. CSICC 2009. 14th International CSI*, pp. 570-575. IEEE, 2009.

[15] Monrose, Fabian, and Aviel D. Rubin. "**Keystroke dynamics as a biometric for authentication**." *Future Generation computer systems* 16, no. 4 (2000): 351-359.

[16] KDD Cup 1999 Data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[17] Boukerche, Azzedine, Kathia Regina Lemos Jucá, João Bosco Sobral, and Mirela Sechi Moretti Annoni Notare. "An artificial immune based intrusion detection model for computer and telecommunication systems." Parallel Computing 30, no. 5 (2004): 629-646.

[18] Montalvao, Jugurta, Carlos Augusto S. Almeida, and Eduardo O. Freire. "**Equalization of keystroke timing histograms for improved identification performance**." In *Telecommunications Symposium, 2006 International*, pp. 560-565. IEEE, 2006.

[19] Kumar, Vipin, Himadri Chauhan, and Dheeraj Panwar. "K-means clustering approach to analyze NSL-KDD intrusion detection dataset." International Journal of Soft Computing and Engineering 3 (2013).

[20] Sahu, Santosh Kumar, Sauravranjan Sarangi, and Sanjaya Kumar Jena. "A detail analysis on intrusion detection datasets." In Advance Computing Conference (IACC), 2014 IEEE International, pp. 1348-1353. IEEE, 2014.

[21] Mahar, Doug, Renee Napier, Michael Wagner, William Laverty, R. D. Henderson, and Michael Hiron. "**Optimizing digraph-latency based biometric typist verification systems: inter and intra typist differences in digraph latency distributions**." *International journal of human-computer studies* 43, no. 4 (1995): 579-592.

[22] Woodward, John D., Nicholas M. Orlans, and Peter T. Higgins. ***Biometrics:[identity assurance in the information age****]*. New York: McGraw-Hill/Osborne, 2003.