

KIET JOURNAL OF COMPUTING AND INFORMATION SCIENCES



ISSN (P): 2616-9592

ISSN (E): 2710-5075



Volume: 4

Issue: 1

Jan - Jun

2021



KIET JOURNAL OF COMPUTING AND INFORMATION SCIENCES

Volume 3, Issue 2, 2020

ISSN (P): 2616-9592

ISSN (E): 2710-5075

Frequency Bi-Annual

Editorial Board

Patron

Air Vice Marshal (Retd) Tubrez Asif, HI(M) - President, KIET

Editor-in-Chief

Prof. Dr. Muzaffar Mahmood

Associate Editor

Dr. Muhammad Affan Alim

Managing Editor

Prof. Dr. Muhammad Khalid Khan

Manager Production & Circulation

Mr. Muhammad Furqan Abbasi



College of Computing & Information Sciences
Karachi Institute of Economics & Technology

College of Computing & Information Sciences

Vision

To develop technology entrepreneurs & leaders for national & international market

Mission

To produce quality professionals by using diverse learning methodologies, aspiring faculty, innovative curriculum and cutting edge research, in the field of computing & information sciences.



AIMS AND SCOPE

KIET Journal of Computing and Information Sciences (KJCIS) is the bi-annual, multi-disciplinary research journal published by **College of Computing & Information Sciences (CoCIS)** at **Karachi Institute of Economics and Technology (KIET)**, Karachi, Pakistan. **KJCIS** aims to provide a panoramic view of the state of the art development in the field of computing and information sciences at global level.

It provides a premier interdisciplinary platform to researchers, scientists and practitioners from the field of computing and information sciences to share their findings and contribute to the knowledge domain at global level. The journal also fills the gap between academician and industrial research community.

KJCIS focused areas for publication includes; but not limited to:

- Data mining
- Big data
- Machine learning
- Artificial intelligence
- Mobile applications
- Computer networks
- Cryptography and information security
- Mobile and wireless communication
- Adhoc and body area networks
- Software engineering
- Speech and pattern recognition
- Evolutionary computation
- Semantic web and its application
- Data base technologies and its applications
- Internet of things (IoT)
- Computer vision
- Distributed computing
- Grid and cloud computing

OPEN ACCESS POLICY

For the benefit of authors and research community, this journal adopts open access policy, which means that the authors can self-archive their published articles on their own website or their institutional repositories. The readers can download or reuse any article free of charge for research, further study or any other non profitable academic activity.

PEER REVIEW POLICY

Peer review is the process to uphold the quality and validity of the published articles. KJCIS uses double-blind peer review policy to ensure only high-quality publications are selected for the journal. Papers are referred to at least two experts as suggested by the editorial board. All publication decisions are made by the journal's Editors-in-Chief on the basis of the referees' reports. We expect our Board of Reviewing Editors and reviewers to treat manuscripts as confidential material. The identities of authors and reviewers remain confidential throughout the process.

COPYRIGHT

All rights reserved. No part of this publication may be produced, translated or stored in a retrieval system or transmitted in any form or by any means; electronic, mechanical, photocopying and/ or otherwise the prior permission of publication authorities.

DISCLAIMER

The opinions expressed in **KIET Journal of Computing and Information Sciences (KJCIS)** are those of the authors and contributors, and do not necessarily reflect those of the journal management, advisory board and the editorial board. Papers published in KJCIS are processed through double blind peer-review by subject specialists and language experts. Neither the **CoCIS** nor the editors of **KJCIS** can be held responsible for errors or any consequences arising from the use of information contained in this journal, instead; errors should be reported directly to the corresponding authors of the articles.

Academic Editorial Board

Dr. Ronald Jabangwe <i>University of Southern Denmark, Denmark</i>	Dr. Sardar Anisul Haque <i>Alcorn State University, USA</i>
Dr. M. Ajmal Khan <i>Ohio Northern University, USA</i>	Dr. Yasser Ismail <i>Southern University Louisiana, USA</i>
Dr. Suliman A. Alsuhibany <i>Qassim University, Saudi Arabia</i>	Dr. Manzoor Ahmed Hashmani <i>University of Technology Petronas, Malaysia</i>
Dr. Wael M El-Medany <i>University of Bahrain, Bahrain</i>	Dr. Atif Tahir <i>FAST NUCES, Pakistan</i>
Dr. Asim Imdad Wagan <i>Mohammad Ali Jinnah University, Pakistan</i>	Dr. Maaz Bin Ahmed <i>Karachi Institute of Economics & Tech, Pakistan</i>
Dr. Salman A. Khan <i>Karachi Institute of Economics & Tech, Pakistan</i>	Dr. Taha Jilani <i>Karachi Institute of Economics & Tech, Pakistan</i>

Advisory Board

Dr. Andries Engel brecht <i>University of Pretoria, South Africa</i>	Dr. Mohamed Amin Embi <i>University Kebangsaan, Malaysia</i>
Dr. Rashid Mehmood <i>King Abdul Aziz University, Saudi Arabia</i>	Dr. Anh Nguyen-Duc <i>Norwegian University of Technology, Norway</i>
Dr. Ibrahima Faye <i>University of Technology Petronas, Malaysia</i>	Dr. Tahir Riaz <i>Data Architect, SleeknoteApS, Denmark</i>
Dr. Faraz Rasheed <i>Microsoft, USA</i>	Dr. Mostafa Abd-El-Barr <i>Kuwait University, Kuwait</i>
Dr. Abdul Naser Mohamed Rashid <i>Qassim University, Saudi Arabia</i>	Dr. Mohd Fadzil Bin Hassan <i>University of Technology Petronas, Malaysia</i>
Dr. Syed Irfan Hyder <i>Institute of Business Management, Pakistan</i>	Dr. Bawani S. Chowdry <i>Mehran University, Jamshoro, Pakistan</i>
Dr. Jawad Shami <i>FAST - NUCES, Pakistan</i>	Dr. Nasir Tauheed <i>Institute of Business Administration, Pakistan</i>

Table of Content

<b style="font-size: 2em;">1 1- 14	Influence of Online Banking on consumer and its perception and acceptance during Covid-19 <i>Jahangir Hameed</i>
<b style="font-size: 2em;">2 15-26	A Comparative Study of Curvature-Based and Differential Versions of Dotter Raster-stereography Techniques <i>Muhammad Wasim, S. Talha Ahsan, Lubaid Ahmed</i>
<b style="font-size: 2em;">3 27-43	Concept Drift in Streaming Data: A Systematic Literature Review <i>Tatheer Fatima, Dr. Tariq Mahmood</i>
<b style="font-size: 2em;">4 44-56	Two-Dimensional Wavelet based Medical Videos using Hidden Markov Tree Model <i>Rubab Fatima Bangash, Imran Tauqir, Azka Maqsood</i>
<b style="font-size: 2em;">5 57-70	Keystroke dynamics Based Technique to Enhance the Security in Smart Devices <i>Farman Pirzado, Shahzad Memon, Lachman Das Dhomeja, Awais Ahmed</i>

Influence of Online Banking on consumer and its perception and acceptance during Covid-19

Jahangir Hameed¹

Abstract

Online Banking Growth. If service quality/ Technological factors are increased this will result in online banking growth in pandemic situation. Lastly, Social, Cultural and demographic Factors has beta value -0.138 , t-value = -1.601 and sig value 0.111 . These values tell that Social, Cultural and Demographic Factors have negative insignificant relationship with Covid19 and Online Banking Growth. Which means if Social, Cultural and Demographic Factors are increased this is result in decreased online banking growth during COVID-19 situation.

Technology has overlapped the globe within the past two decades, from ordering the food through mobile to making sophisticated and high level transactions through e-commerce, everything have been on the fingertips of human being. This enable banking sector to enhance their services and upgrade their practices to online banking system. This study is about Influence of Online Banking on consumer and its perception and acceptance during Covid-19. Research hypothesis and research questioner had been developed to undertake the study for the research hypothesis, primary data has been used through the sample size and Convenience sampling was carried out where the quota was determined according to the share of the total number. To answer the questions statistically five hypotheses were created and tested through the questionnaire. Results shows that majority of the respondents have agreed and established trust on online banking transactions, their ease with respect to traditional banking, time saving and convenience, while showing a neutral behavior about fraudulent activities.

Keywords: Online Banking, Customer Satisfaction, Technological Factor, Demographic, Covid-19.

1 Introduction

Online banking has been the top most recent advancement which enables customers with shopping, fund transfers and many more, all through online platform. Online banking has made it easy for people as they donot need to stand in ques and wait for hours. One can have access of all banking records and information 24/7 within reach through online banking system. People can carry out online transactions; sending and receiving amount with surity of clearance mainly at their homes. If one has personal computer or mobile devices, it becomes easy for him to engulf in the handy process of online banking system. Today in the modern world, people has given different attributes to online banking such as electronic, web, internet or mobile banking. Online banking is a connectivity channel between customers and benefactors.

For the purpose of online banking different banks across the globe have established applications and websites through which online banking could be taken place. These online and web based

¹ Karachi University Business school | jhangirhameed@gmail.com

forms on banking has facilitated the customers with the best possible ways. Through online banking cost of human resources has reduced. We can say that online banking system has helped banks to gain favour from their customers with improved services and revise traditional methods of banking. Globally, there is a surge in investment as a result of online banking system and different banks have facilitated their customers with huge profits. This modern way of banking is possible when customers comply with the technological advancement in field of banking.

As far as online banking is concerned in Pakistani context, its scope seems not so high as people are still not familiar of modern ways of banking. The aim of this research is to investigate acceptance of Online banking in Pakistan on consumer perceptions and its acceptance. The study was limited in its scope because of Covid-19. The major objective of the study was to investigate acceptance of online banking in Pakistan during covid-19. The major influences on this relationship were security and privacy issues, people's lack of knowledge, problem in accessing internet, satisfaction of customers, adaptation of modern technology, socio-cultural influences including Islamic and traditional values along with demographic representations.

To validate the objective of research, we have used Research hypothesis and research questioner. For the research hypothesis, we used primary data from data collection. Primary Data is also collected through the sample size of and was chosen based on sources. Convenience sampling was carried out where the quota was determined according to the share of the total number. To answer the questions statistically five hypotheses were created and tested through the questionnaire. The major research instrument from which we decide whether our business research hypothesis is accepted or not for that we collect data through questioner.

2 Problem Statement

In traditional banking people sometimes pay more amount on transactions than the actual amount being deposited or withdrawn. They have to stand for hours in banks to get their transactions done. Online or web/app-based banking system has facilitated customers to carry out economic and secure banking. In modern developing world various banks has shifted from traditional to online systems which turned out be beneficial for customers saving time reducing the cost. Also, technology is user friendly which has enabled users to use it in abundance. According to the user and customers perspective online banking is customer oriented which is accessible 24/7 to the customers.

With the rampant use of online banking there are some hurdles as well. People are observed with lack of awareness about technology, they have difficulty in operating online system, they have legal issues, fragile relation between banker and customers. According to some people online banking is customer oriented and safe but for other people in Pakistan who comes from rural areas it is very challenging. They have lack of awareness, security threats and difficulty in access. Therefore, it becomes significant to investigate perception of Pakistani people towards acceptance of online banking.

3 Study Objectives

The Primary Objective of this study is to find out the level of impact of online banking on consumer behavior and its perception during covid-19 through customer satisfaction, technological factors, social, cultural, economic and demographic factors.

A *Secondary Objective*

The objective is to identify the usage of online during covid-19.

The objective is to identify what kind of barriers and resistance occurs on consumer mind for online banking

The objective is to identify how we can improve online growth in Pakistan.

B *Scope of Study:*

The scope of the study is to discover the acceptance of Online banking in Pakistan on consumer perceptions. The scope of the research was limited Due to the Pandemic situation. The objective of the research was to find acceptance of online banking in Pakistan during covid-19. The major factors affecting this relationship were congruency to the awareness and information marketed by banks, security issues, privacy issues, demographic factors etc.

C *Hypothesis*

H1: Trust, Security, and privacy issue among the acceptance of Online banking customers in Pakistan.

H2: Customer Satisfaction influence the acceptance of Online banking among banking customers in Pakistan.

H3: Economic factors and demographic factors are one of the challenges for implementation and development of Online banking in Pakistan.

H4: Management and banking issues are one of the challenges for implementation and development of Online banking in Pakistan.

H5: Technological, Social, and cultural barriers are one of the challenges for implementation and development of Online banking in Pakistan.

4 Literature Review

The literature suggests that many factors influence the acceptance and consumer perception of online banking during pandemic. Different authors suggest different opinions about adoption of online banking system in a country.

In internet banking privacy and security plays a vital role. If these both factors are provided to customers than it increases comfort of customers towards the bank. Privacy and protection of

online banking transactions and confidentiality of personal information are critical interest to both internet banking customers and the banking industry. Aladwani (2001) Online banking analysis, prospective clients have classified internet security and consumer privacy as the most imperative potential problems faced by banks. In studies of banking, the importance of security and confidentiality to the acceptance of online banking has been specified. (Roboff and Charles, 1998; Sathye, 1999; Hamlet and Strube, 2000; Tan and Teo, 2000; Polatoglu and Ekin, 2001; Black et al., 2002; Giglio, 2002; Howcroft et al., 2002). Roboff and Charles (1998) discovered that, despite their knowledge of the risks and their dependence on their bank on privacy issues, individuals have a fragile understanding of online banking security risks. Similarly (Fatima Mazhar, 2014) studied the impact of various factors on the adaptation of online banking through TAM (technological acceptance model) in Pakistan, specifically in non-urban regions

More precisely, Privacy and protection were found to be crucial barriers to the adoption of online banking, several studies have shown that protection and privacy are important indicators of the desire towards online banking (Yousafzai et al., 2009; Abu-Shanab & Abu-Baker, 2011). This mindset indicates that Internet banking networks can have protection measures and reduce the possibility of leakage of user-related information leading to scam (Ameme, 2015).

(Mohammad Alafeef, 2012) claims that, most talked external factor that is considered to be an obstacle for adaptation of technology is trust, and believes that people will live upto his expectations and will not take advantage of it, McGoldrick and Laforet explains that possible reason for people to not go for adaptation of technology or online banking are privacy and security concerns. Peralta and Novak (1999) says issues such as security fear, fear of invading private information, less profitable business, less security for payment are all obstacles in way of online banking's adaptation. In lessening these concerns, plays a crucial role. It helps to reduce financial transaction-related complexities, ambiguities, threats and fraud, thus increasing the probability that customers will go for online banking.

According to (Syed Sheheryar Ali Kazmi, March 2015) has defined issue and challenges that Pakistan facing on online banking sector. First current issue is lack of infrastructure investment in Information technology , mostly banks in Pakistan still behind from other countries banks they have not adopt data base system or integration of enterprise software like SAP and oracle which is also a big cause towards digitization of e-banking in Pakistan , similarly issue like trust issue of customer towards e-banking, security and privacy issue which is causing cyber-attack on online banking also a big obstacle in adoption of online banking in Pakistan.

Other major issue like money laundering which is also facing by banks because it's easy to hack the account and transfer the payment to other accounts of person it occurs due to the lack of basic infrastructure of Information technology systems and untrained staff of banking industry which is also a big barrier facing by banks in Pakistan.

In marketing major concern is satisfaction of the customers (Safeena R, 2015) . it increases the purchase and attitudes of the customers such as repetition and loyalty to the brand. According to Oliver (1980), when clients equate their understanding of real brand results with expectation, the sense of satisfaction emerges. A variety of concepts have been reported on consumer

satisfaction. . Oliver (1980) describes satisfaction as an emotional evaluative decision about a good or service after consumption. Similarly, Tse and Wilton (1988) identified consumer satisfaction as a "customer reaction to the evaluation of the potential difference between expectations and end result after usage." Gratification can also be defined as the evaluation of a post-purchase evaluation of the quality of certain service/product and compared to the expectation of previous purchase Kotler & Keller (2011)

According to the studies so far consumers sense of satisfaction is displayed in customers attitude to adopt the platform of online banking be it internet banking, or mobile banking.

(Jamil Hammoud, 2018) reveals in his research that demographics influences are also associated with adoption of online banking. According to one study conducted in US, which suggests that online banking is affected by level of education and income positively and being affected negatively due to age. Rogers (1995) says that young generation with high skills and knowledge are more inclined to adopt technological advancements and new methods. Also genders seems to affect online banking directly. Income level also determine the adoption of online banking people with more stable income have more use as compare to low income level customers. Then the lack of investment in banking sector specifically in the area of online banking facilities and ongoing economic crisis also the barriers in acceptance of online banking in Pakistan.

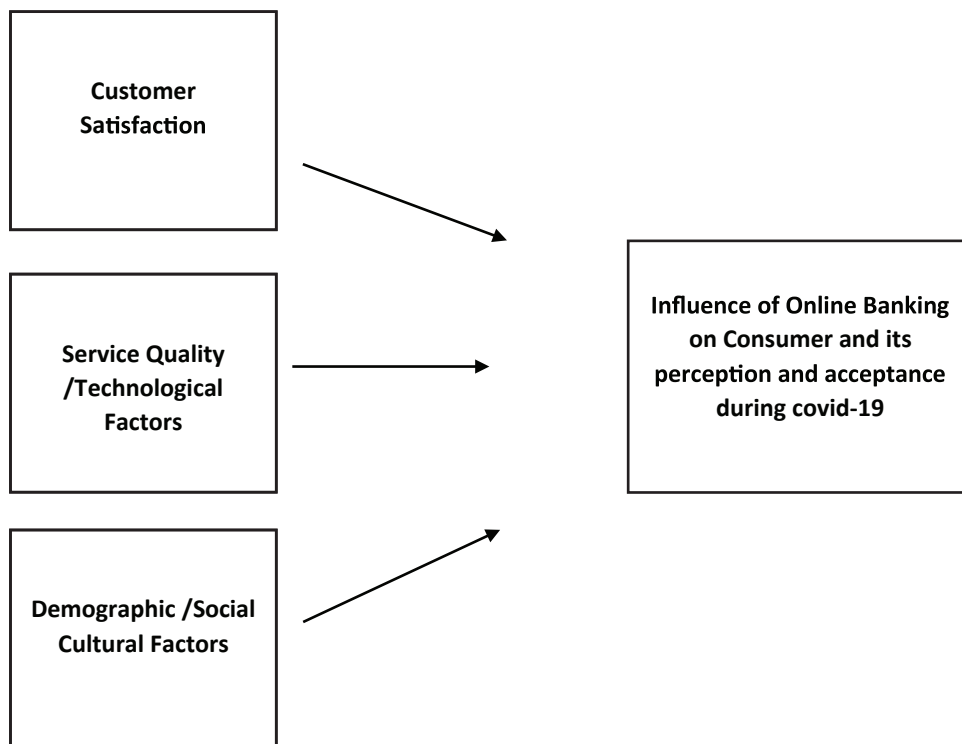


Figure 1: Theoretical Framework:

The Covid-19 is an independent variable in the study and its impacting the influence of online banking. Customer Satisfaction, Service quality, technological factors, social cultural factors and demographic factors are dependent variables in the present study. However as per the literature

of the study, it has been observed that covid-19 its affected consumer business and each sector adversely because of the pandemic and there is no prediction related to it. The pandemic majorly impacting countries' economies and as well customer income is affected. On the other hand, covid-19 has also emerged an opportunity for online industries like e-commerce and online banking has growth enormously in this pandemic around the world.

5 Banking Channels Review

As per the SBP Apr-June quarterly data on alternative delivery channels of banks (ADC) for banking payment sector payments and transactions.

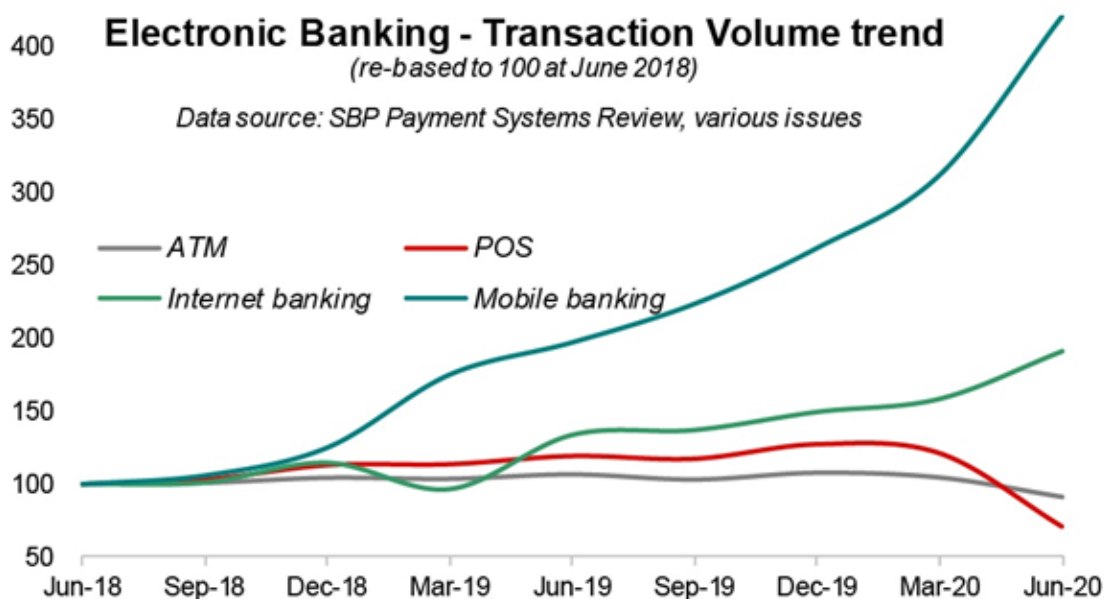


Figure 2: Electronic Banking - Transaction Volume Trend

As per figures it is observed ATM cash withdrawal during pandemic is going down due to the lockdown and pandemic situation but on the other hand mobile banking channels has been growing well in recent year and the register mobile banking transactions had reached 8.5 million as of June 2020 with 50 percent growth in just a year of 2020, some 27 banks has offering this ADC. The mobile banking momentum continued amidst the lockdowns. For instance, transaction volume of 29 million in Apr-Jun was higher by a third on quarterly basis and it more than doubled on a yearly basis. The transaction value of Rs622 billion in the same quarter followed similar levels of quarterly and yearly spikes. There is continued growth across the board, but mostly in intra-bank and inter-bank fund transfers

The third pillar of e-banking is internet banking, which is another fast-growing ADC, offered by 28 banks to nearly 4 million users as of June 2020. The lockdowns had accelerated the growth in this channel relative to recent quarters. At 17 million transactions in Apr-Jun, the volume was higher 21 percent over previous quarter and up 43 percent year-on-year. At Rs894 billion, the

transaction value in the corona quarter was 20 percent and 64 percent higher on quarterly and yearly bases, respectively. Just as with mobile banking, fund transfers take the lead in expanding transaction pie for internet banking.

And finally, the ever-shrinking POS machines, which had reduced by 14 percent year-on-year to 49,000 as of June 2020 and embraced by mere 9 banks. The pandemic quarter's troubles have set back quarterly POS transactions even more than two years. The lockdown did not help and customer are afraid to use POS machine Due to pandemic mutation.

This analysis show that some ADC channel growth has been decreases and some of ADC channels like mobile and internet banking has tremendously growth during pandemic which is good signed for banking sector growth.

6 Methodology

A Research Philosophy

In view of Silverman (2016), research philosophy is the first step of methodology as it creates the base of the data collection, extraction, and process. Through this, researchers explore the possibilities of systematically gathering information from different sources to systematically complete study objectives and questions. Some of the research philosophies are positivism, realism, and interpretivism as it provides structure to study. Interpretivism philosophy supports the qualitative study, and it allows researchers to extract detailed information. However, realism philosophy is critical because these objectives are related to the real event, and researchers need to capture data from authentic sources.

For the present study, information has been derived from positivism philosophy as it supports quantitative research, and in this, the researcher capture data from numerical sources. The researcher derives information from a quantitative source as it allows the researcher to measure the relationship among variables with the help of software to derive results accurately.

B Approach

In view of Gray (2016), deductive and inductive are approaches of study, and they are connected with the philosophy of research and further allow analyst to derive data from accurate sources. According Taylor ml (20 IS), the inductive approach supports interpretivism philosophy, and in this, detailed information is extracted. This approach is considered because it is one of the flexible sources and allows researchers to derive information from multiple ways in detail. In an inductive approach, an analyst emphasizes specific observation, and as per that, they derive information in detail to complete the study effectively.

However, in this study, the researcher focuses on the deductive approach because it supports positivism philosophy y and it is captured through the quantitative source. The deductive method has a complete structure that is being followed by researchers globally; it starts with a hypothesis of the study, and further information is collected to accept or reject it as per sources. However, quantitative data provide numerical data, which helps in concluding results and allow

the researcher to gather knowledge about the relationship among variables and complete findings effectively.

C Data Collection Method

In view of Gray & Alins (2016), primary and secondary are data collection methods that researchers consider completing the study systematically and fulfill the requirements of objectives effectively. The primary techniques provide more specific information because, in this, first-hand knowledge is gathered, which is related to a study. However, secondary sources are considered to get detailed information, and it helps to complete the literature of research.

In this study, the researcher adopted both methods to complete the study because the analyst mainly needs to capture information related to re-education of online banking influence during Covid-19 and for that data is required, and secondary data further complete the requirement of literature. The secondary information has been derived from books, literature, researchers, journal, articles, and other online sources to get the authentic information. For the primary source, the questionnaire has been considered by the researcher to capture information from customers as they are people who use the Online banking from different parts backgrounds.

D Sample Size and Technique

According to Martens (2014), the sample size of a study is one of the critical things as it needs to be selected with complete concentration because it impacts the research findings. The study sample size of the study is 250 as it presents the opinion of all customers includes students, businessman, job-oriented peoples etc. However, as per opinion of Creswell (2014), two sampling techniques are commonly considered by researchers, which are probability and non-probability. The probability sampling techniques are simply random, random cluster, and systematic sampling.

In the current study, the researcher has been adopted non-probability sampling because the analyst does not have information about the total population. When the researcher does not exact information about the population non-probability sampling technique is considered, and from further techniques, the Convenience sampling method is adopted. Through Convenience sampling, the researcher can judge people and gather information from people who use online banking during covid-19 or before pandemic as well.

E Ethical consideration

In view of Worthington and Bodie (2017), ethics of the study are considered by researchers because it has an impact on the findings and credibility of the study. In this research, numerous ethics are considered because they can affect conclusions and allow the researcher to systematically complete the study. The researcher provides credit to the study authors and derives information from authentic sources to enhance credibility. Further, the researcher does not manipulate any data from primary and secondary sources to maintain the findings' integrity. However, respondents are part of the study and the researcher respected each individual, and information has not been derived forcefully. The researcher derived information from customers

and treated them anonymously as they were not willing to share personal information, which is part of the study's ethics.

7 Result

This chapter summaries the analysis or the result generated. SPSS was used to assess the data. Regression analysis, correlation analysis, reliability and descriptive analysis was used to analyze data.

A *Reliability*

Table 1: Reliability Statistics

Cronbach's Alpha	N of Items
.903	17

Above table tells about the reliability and internal consistency of the instrument. Table tells that the value of Cronbach's Alpha is .903. Acceptable range of Cronbach's Alpha is 0.7 to 0.9. As the value of instrument is equal to 0.9 this mean

That instrument used is acceptable and reliable and can be used to research. 0.9 value of instrument reliability tells that instrument used is highly reliable.

Descriptive Analysis:

Table 2: Descriptive Statistics

	N	Minimum	Maximum	Mean	Std Deviation
Customer Satisfaction	260	1	5	3.82	.675
Service Quality/Technological Factors	260	1	5	3.73	.682
Social, Cultural and Demographic Factors	260	1	5	3.80	.646
Covid19 and Online Banking Growth	261	1	5	3.62	.828
Valid N (listwise)	260				

Descriptive analysis is used to find out the distribution of the data. It also tells about the correlation of variables. Its summaries the data, however it cannot be used to draw any conclusion. The above table tells that responses collected or used in this research are 260; N=260. Similarly, it tells the mean of variables and from the above it can be seen that all variables have mean values close to each other. Standard deviation tells about the skewness of the data which is how much your data is spread. The greater value of standard deviation shows that data is greatly spread. However, from the above table it can be seen that descriptive analysis doesn't contain higher values of SD which means our data is not very much spread.

B Correlation Analysis

Table 3: Correlations

		Customer Satisfaction	Service Quality/ Technological Factors	Social, Cultural & Demographic Factors	Covid19 and Online Banking Growth
Customer Satisfaction	Pearson Correlation	1	.650**	.707**	.549**
	Sig. (2-tailed)		.000	.000	.000
	N	260	260	260	260
Service Quality/ Technological Factors	Pearson Correlation	.650**	1	.880**	.879**
	Sig. (2-tailed)	.000		.000	.000
	N	260	260	260	260
Social, Cultural and Demographic Factors	Pearson Correlation	.707**	.880**	1	.747**
	Sig. (2-tailed)	.000	.000		.000
	N	260	260	260	260
Covid19 and Online Banking Growth	Pearson Correlation	.549**	.879**	.747**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	260	260	260	261

** . Correlation is significant at the 0.01 level (2-tailed).

Correlation matrix is used to evaluate the relationship between variables. It not only tells about the relationship between independent variable and dependent variable but it also tells about the relationship among independent and dependent variables. It helps to conclude if variables have significant relationship or not. Above table depicts that customer satisfaction has significant relationship with all variables used in the research. Since the sig value of all variables used to determine their relationship with customer satisfaction is 0.000. Moving on if we analyze the relationship of Service Quality/ Technological Factors with rest of the variables, this can be concluded that Service Quality/ Technological Factors have significant relationship with other factors as the sig value of all variables is 0.000. Similarly, table tells that Social, Cultural and Demographic Factors have significant relationship with other variables; sig value of these variables is 0.000. Finally, if we evaluate the relationship of dependent variable which is COVID 19 and online banking growth, it can be said that dependent variable has significant relationship with independent variables as the sig value of all independent variables is 0.000.

C Regression Analysis

Table 4: Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.881a	.776	.774	.395

a. Predictors: (Constant), Social, Cultural and Demographic Factors, Customer Satisfaction, Service Quality/Technological Factors

Model summary tells about model fit. It tells about the strength of relationship of dependent variable and model of the study. R value depicts the strength of relationship. The greater the value of R, the stronger the relationship. R value of the model is .881 which shows that model has strong relationship with dependent variable. R square which is coefficient of determination indicates the total variation in dependent variable is explained by independent variable. Here the value of R-Square is .776 which means that 77.6% of the total variation is explained by independent variables rest is explained by unknown factors.

Table 5: ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	138.396	3	46.132	295.926	.000b
	Residual	39.908	256	.156		
	Total	178.304	259			

a. Dependent Variable: Covid19 and Online Banking Growth

b. Predictors: (Constant), Social, Cultural and Demographic Factors, Customer Satisfaction, Service Quality/Technological Factors

One of the most important table in regression analysis is ANOVA because it tells if the model is significant or not. In above table the F value of model is 295.926. As per researchers F value should be greater. The greater value of F results in more accurate result. Above table tells that our mode is significant as the sig value is 0.000.

Table 6: Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.246	.156		-1.580	.115
	Customer Satisfaction	-.016	.052	-.013	-.317	.752
	Service Quality/ Technological Factors	1.195	.076	.982	15.747	.000
	Social, Cultural and Demographic Factors	-.138	.086	-.107	-1.601	.111

a. Dependent Variable: Covid19 and Online Banking Growth

Coefficient table tells about the significance, strength and positive or negative relationship of independent variables with dependent variables. From above table it can be seen that customer satisfaction has insignificant negative relationship with Covid19 and Online Banking Growth as the sig value is greater than 0.752, and t value is -3.17. In order to accept the significance of hypothesis, Sig value should be less than 0.05 and t value should be more than 2. Therefore, it can be said that customer satisfaction doesn't have any impact on Covid19 and Online Banking Growth. Moving on to service quality/ Technological factors, the beta value is 1.195, t- value is 15.747 and sig value is 0.000, this means that service quality/ Technological factors has positive significant relationship with Covid19 and Online Banking Growth. If service quality/ Technological factors are increased this will result in online banking growth in pandemic situation. Lastly, Social, Cultural and demographic Factors has beta value -.138, t-value = -1.601 and sig value 0.111. These values tell that Social, Cultural and Demographic Factors have negative insignificant relationship with Covid19 and Online Banking Growth. Which means if Social, Cultural and Demographic Factors are increased this is result in decreased online banking growth during COVID-19 situation.

8 Conclusion and Recommendation

In view of Hansol (2020) Covid-19 badly effected on all business and consumer purchasing power and its long term risk is still continue in the world .Most of Analysis and Experts do not have complete information to covid-19 consequences on economy growth and its impact on industries .In the light of Primary source it has been observed that majority of participants use online banking on going Covid-19 pandemic ,It is found from a Secondary source(Business Recorder) the Annual Banking Channels Review 2020 the technology has given edge on banking operation during pandemic customers can effectively perform transaction through ADC channels which includes online banking and dramatically increasing day by day as compare to 2019 .On the other hand as per Sneader and Singhal (2020) consumer behavior change drastically, and they are less likely to spend on luxurious goods and more likely to essential ,similarly some customer are willing to live everyday life with pandemic situation and some are waiting for vaccine to start their life as usual in the past years .

Furthermore, as per survey results, most of the respondents agree that due to covid -19 they prefer to use online banking over the traditional banking channels. In conclusion, research suggests that online banking is choice of many customers because it appears to be favorable for them because of its worth, consumer satisfaction, advanced technology, security measures and privacy concerns. Besides these other influences on the acceptance of online banking are its advertisement that provided sufficient information to the customers through electronic and print media.

It has been discovered that around 64% people turned towards online banking. Online banking was accepted by Pakistani people owing to its usefulness and ease, still there is room for improvement as some people are unsure about online banking and its security so the benefactors need to put efforts to encourage people for online banking. Further researchers may conduct their study on online banking in relation to demographics factors including education level of the consumers. Decisions could be made on behalf of bank managers for expansion of market considering the usefulness, security and privacy in the online banking system.

References

- [1] Fatima Mazhar, S. I. (2014). An Investigation of Factors Affecting Usage and Adoption of Internet & Mobile Banking In Pakistan. *International Journal of Accounting and Financial Reporting*.
- [2] Jamil Hammoud, R. M. (2018). The Impact of E-Banking Service, Quality on Customer Satisfaction: Evidence from Lebanese Banking Sector. *SAGE Open*.
- [3] Mohammad Alafeef, D. S. (2012). The Influence of Demographic Factors and User Interface on Mobile Banking Adoption. *Journal of Applied Sciences*.
- [4] Safeena R, A. D. (2015). Case study on internet banking. *Journal of Internet Banking and Commerce*.
- [5] Stacy L Wood. (2012 Volume 78, Issue 1,). Future fantasies: a social change perspective of retailing in the 21st century. *Journal of Retailing*, 77-83.
- [6] Syed Sheheryar Ali Kazmi, M. H. (March 2015). E-Banking in Pakistan: Issues and Challenges. *International Journal of Academic Research in Business and Social Sciences*.
- [7] Business Recorder Pakistan. (2020, October 16). E-banking during pandemic. p. 1.
- [8] Silverman, D. (2017). How was it for you? The Interview Society and the irresistible rise of the (poorly analysed) interview. *Qualitative Research*, 17(2), 144-158.
- [9] Gray, C., & Malins, J. (2016). *Visualizing research: A guide to the research process in art and design*. Routledge.
- [10] Mertens, D. M. (2014). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*. Sage publications.
- [11] Sudarsono, H., Nugrohowati, R. N. I., & Tumewang, Y. K. (2020). The Effect of Covid-19 Pandemic on the Adoption of Internet Banking in Indonesia: Islamic Bank and Conventional Bank. *The Journal of Asian Finance, Economics, and Business*, 7(11), 789-800.
- [12] Hwang, H., Hur, W. M., & Shin, Y. (2020). Emotional exhaustion among the South Korean workforce before and after COVID-19. *Psychology and Psychotherapy: Theory, Research and Practice*.
- [13] Sneider, K., & Singhal, S. (2020). Beyond coronavirus: The path to the next normal. *McKinsey & Company*.
- [14] Qureshi, T. M., Zafar, M. K., & Khan, M. B. (1970). Customer acceptance of online banking in developing economies. *The Journal of Internet Banking and Commerce*, 13(1), 1-9.
- [15] Oldekop, J. A., Horner, R., Hulme, D., Adhikari, R., Agarwal, B., Alford, M., & Bebbington, A.J. (2020). COVID-19 and the case for global development. *World Development*, 134, 105000.
- [16] Vernirnrnen, P., Quiry, P., Dallochio, M., Le Fur, Y., & Salvi, A. (2014). *Cultural change and innovation*. John Wiley & Sons.

- [17] Worthington, D.L. and Bodie, G.D. eds., (2017). *The textbook of financial institutions and markets: A global perspective*. John Wiley & Sons.
- [18] Deloitte (2020) . Impact of COVID- 19 to the Banking Sector. Retrieved from <https://www2.deloitte.com/cn/en/pages/risk/articles/covid-19-impact-on-banks.html>
- [19] Dietrich, A., Keuster, K., Muller, G. J., & Schoenle, R. (2020). News and uncertainty about covid- 19: Survey evidence and short-run economic impact. Retrieved from <http://people.brandeis.edu/~schoenle/research/COVID19 March2020.pdf>
- [20] Fogel, F., & Gartner, S. (2020). The COVID-19 Pandemic and Relationship Banking in Germany: Will Regional Banks Cushion an Economic Decline or is A Banking Crisis Looming?. *Tijdschrift voor economische geografie*, 11 / (3), 416-433.

A Comparative Study of Curvature-Based and Differential Versions of Dotter Raster-stereography Techniques

Muhammad Wasim ¹S. Talha Ahsan ²Lubaid Ahmed ³

Abstract

Conventional Line-based Raster-stereography has been a popular technique for 3-D surface topography. However, in its application for human face screening, the problem of line breaking was observed. In order to resolve this problem, there came up a new technique called dotted raster-stereography. The previously reported version of dotted raster-stereography extracted the curvature features of human face. This paper presents a modified version, viz. differential dotted raster-stereography in which instead of curvature, differences in straight line distances between adjacent points are calculated. A comparative picture of the two versions of dotted raster-stereography techniques is presented. Results suggest that this new differential version of dotted raster-stereography algorithm is faster in execution due to its simpler implementation in software, though lower in accuracy, as compared with the previously reported curvature-based version of dotted raster stereography.

Keywords: Raster-stereography, dotted raster grid, differential dotted raster-stereography, face recognition

1 Introduction

Raster-stereography and moiré fringe topography have been popular techniques for the last three decades in the domain of health sciences [1-4]. Both of these techniques are very much similar in terms of 3-D surface screening [5]. Raster-stereography performs grid projection on a surface in order to extract the curvature features of that object. At initial stage of this technique, the grid comprising horizontal and vertical lines normally known as conventional raster grid was used in a number of applications. However, it was observed that lines got broken during line extraction of curved surface because of poor contrast between object's surface and black lines [6]. In order to resolve this problem, the first author of this paper introduced, the concept of dotted raster-stereography in 2013 to extract the curvature features of curved surface, for the face recognition application [7]. The dotted raster-stereography technique produced better results and efficiency in comparison with conventional line raster technique. This paper presents a modified version of dotted raster-stereography technique viz. differential dotted raster-stereography, in which instead of curvatures, differences in straight line distances between adjacent points are calculated. Results and comparative analysis of the two techniques are reported.

2 Literature Review

Raster-stereography and moiré fringe topography techniques have been used in a number of applications by the researchers. For example, the use of conventional line raster grid for

¹Usman Institute of Technology, Karachi | mwaseem@uit.edu

²Usman Institute of Technology, Karachi | stahsan@uit.edu

³Usman Institute of Technology, Karachi | lahmed@uit.edu

spinal deformity detection was reported in 2002 [2]. The work on defining the shape of spine using moiré fringe topography was reported in 2014 [8]. In another Study [9], the cogency and consistency of 4-D raster-stereography in dynamic conditions were discussed and it was recommended that raster-stereographic can be used to inspect the spinal posture with an acceptable level of accuracy. Authors in a study [10], presented a non-contact, non-invasive method for imaging and analysis of 3-D surfaces of moving boundaries along with the structure of asymmetrical formed planes. In [11], the author presented a reliable method of raster-stereography to measure the back contour of children body, which also minimized the effect of x-rays.

Human face recognition has been one of the more concerned areas of research for the last few years. Some very common face recognition techniques are Iterative Closest Point [12-17], Hidden Markov Models [18], and Principal Component Analysis [19-21]. Another technique was reported [22] that was based on co-variance-matrix, polynomial-coefficients and algorithm on common eigen-values to recognize among several human faces. In another person[23], authors presented an effective way to identify human faces using Symmetric-Local-Graph-Structure (SLGS), which was based on the concept of Local-Graph-Structure (LGS).

3 Curvature-Based Dotted Raster-stereography

In the concept of curvature-based dotted raster-stereography, a dotted grid as shown in figure 1(a) is projected on the surface of human face to extract and record the curvatures of human face [5]. In this method, human face is converted into certain number of horizontal and vertical pixels (raster image) as mentioned in figure 1(b). The geometry of curvature-based dotted raster-stereography concept is shown in figure 2. By selecting three consecutive points P , Q and R , the arc lengths PR and QR are solved. Same selection of three consecutive points is made from first pixel to last pixel of the image, as given in figure 1(b). This complete calculation of arc lengths provides the curvature feature information of human face. Using the equation of angular displacement ($\sin \frac{\alpha}{2} = \frac{d}{L}$), the arc length (s) is calculated, where radius of arc is $L = \frac{1}{\kappa}$. By putting all the values in equation (1), curvatures (κ_1 and κ_2) are calculated for both horizontal and vertical patterns, based on which, two decision parameters mean (M) and Gaussian (G) are solved.

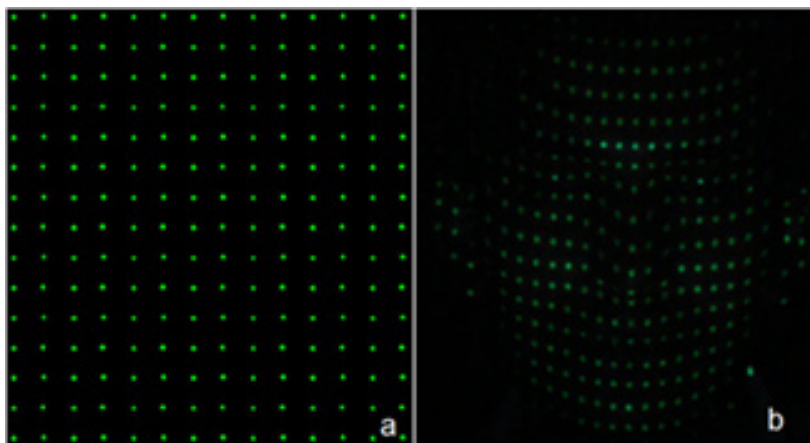


Figure 1: (a) Dotted Raster Grid (b) Curvature-Based Dotted Raster-stereography

Mathematical Formula of curvature-based dotted raster-stereography:

$$K_1=K_2=\pm \frac{1}{s} \sqrt{(24(1-d/s))} \dots \dots \dots (1)$$

K_1 =Horizontal surface curvatures

K_2 = Vertical surface curvatures

d = Linear distance PQ or QR

s = Arc length

α = Angle of tangents intersection

L = Perpendicular of triangle

Decision Parameters:

Mean= $(K_1+K_2)/2$, Gaussian= $K_1.K_2$

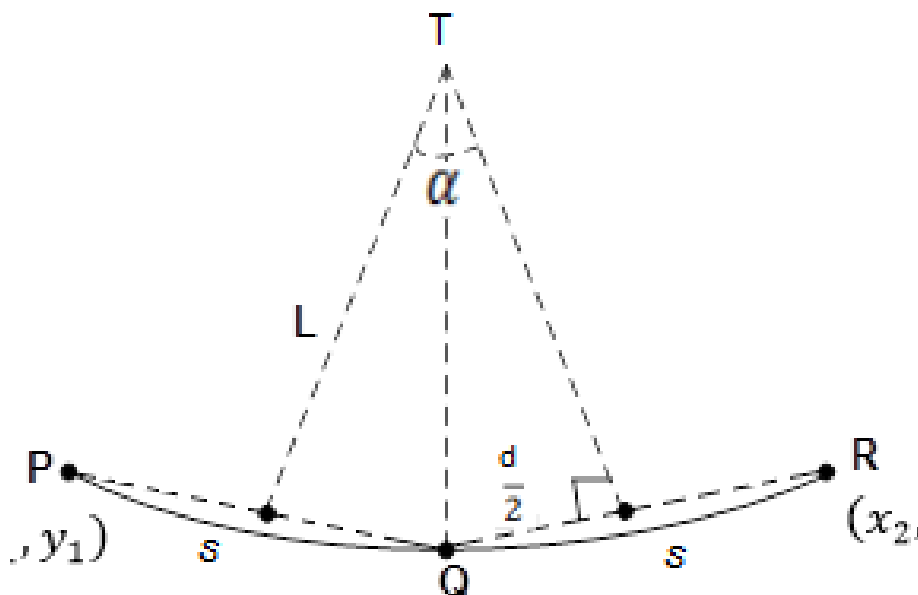


Figure 2: Geometry of Curvature-based Dotted Raster-stereography

4 Differential Dotted Raster-Stereography

This new concept of differential dotted raster-stereography is based on the linear distances between the adjacent points. In this work, ‘ d ’ is the linear distance between two adjacent points ‘ P ’ and ‘ Q ’, as mentioned in figure 3. In this method, instead of arc lengths, the linear distances between two consecutive points (for both horizontal and vertical directions) are calculated.

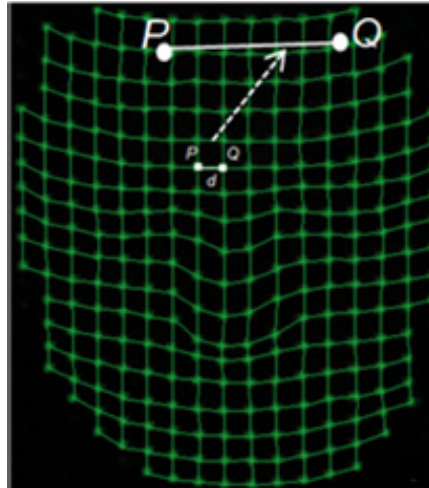


Figure 3: Differential Dotted Raster-stereography

In this proposed model, the facial features of human faces are presented on the basis of linear spacing between two consecutive points. The distance formula $d = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$ is used to calculate distances between two points as illustrated in figure 3, where (x_1, y_1) and (x_2, y_2) are the coordinates of points P and Q respectively. The decision parameter Δd for differential dotted raster-stereography is calculated using the mathematical formula

$\Delta d = \sqrt{d_x^2 + d_y^2}$, where d_x and d_y are the average distances along x- and y-axis respectively, as mentioned in equations (2) and (3).

$$d_x = \frac{(d_{x_2-d_{x_1}})+(d_{x_3-d_{x_2}})+\dots+(d_{x_n-d_{x_{n-1}}})}{n-1} \dots\dots\dots (2)$$

$$d_y = \frac{(d_{y_2-d_{y_1}})+(d_{y_3-d_{y_2}})+\dots+(d_{y_n-d_{y_{n-1}}})}{n-1} \dots\dots\dots (3)$$

The complete system of differential dotted raster-stereography comprises eight different phases i.e- capturing human face, extracting distorted grid, cropping image, finding pixel coordinates, mathematical model, calculation of decision parameters, storing in database and face identification. The detailed functional diagram along with the flow of different phases is given in figure 4.

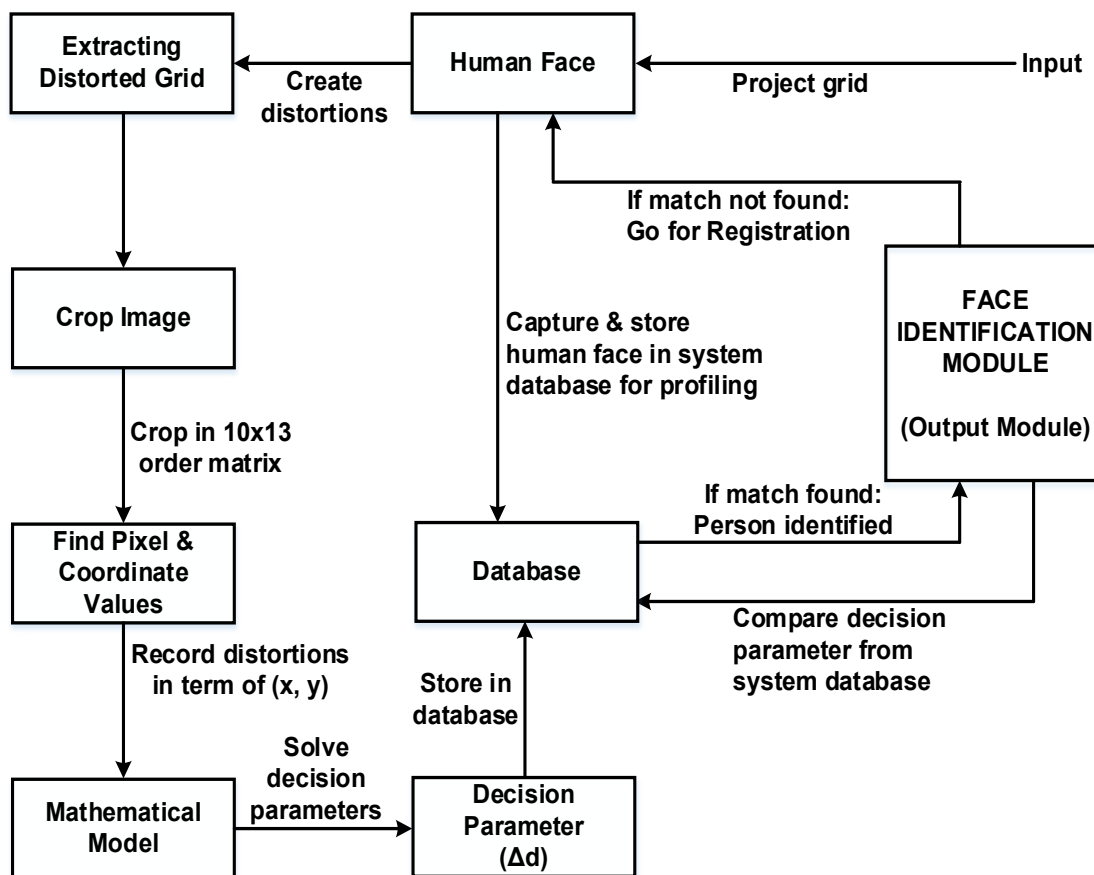


Figure 4: Functional Block Diagram of Differential Dotted Raster-stereography

5 Results and Discussion

A Measured values of curvature and distance parameters

Table 1 presents the horizontal and vertical facial deviations of first four human faces, with the following resulting values of decision-parameter ' Δd ' in differential technique:

FID-01: 46.98 cm; FID-02: 53.88 cm ; FID-03: 67.88 cm ; FID-04: 39.80 cm

Table 1: Horizontal and Vertical Facial Deviations for Face IDs 01 - 04

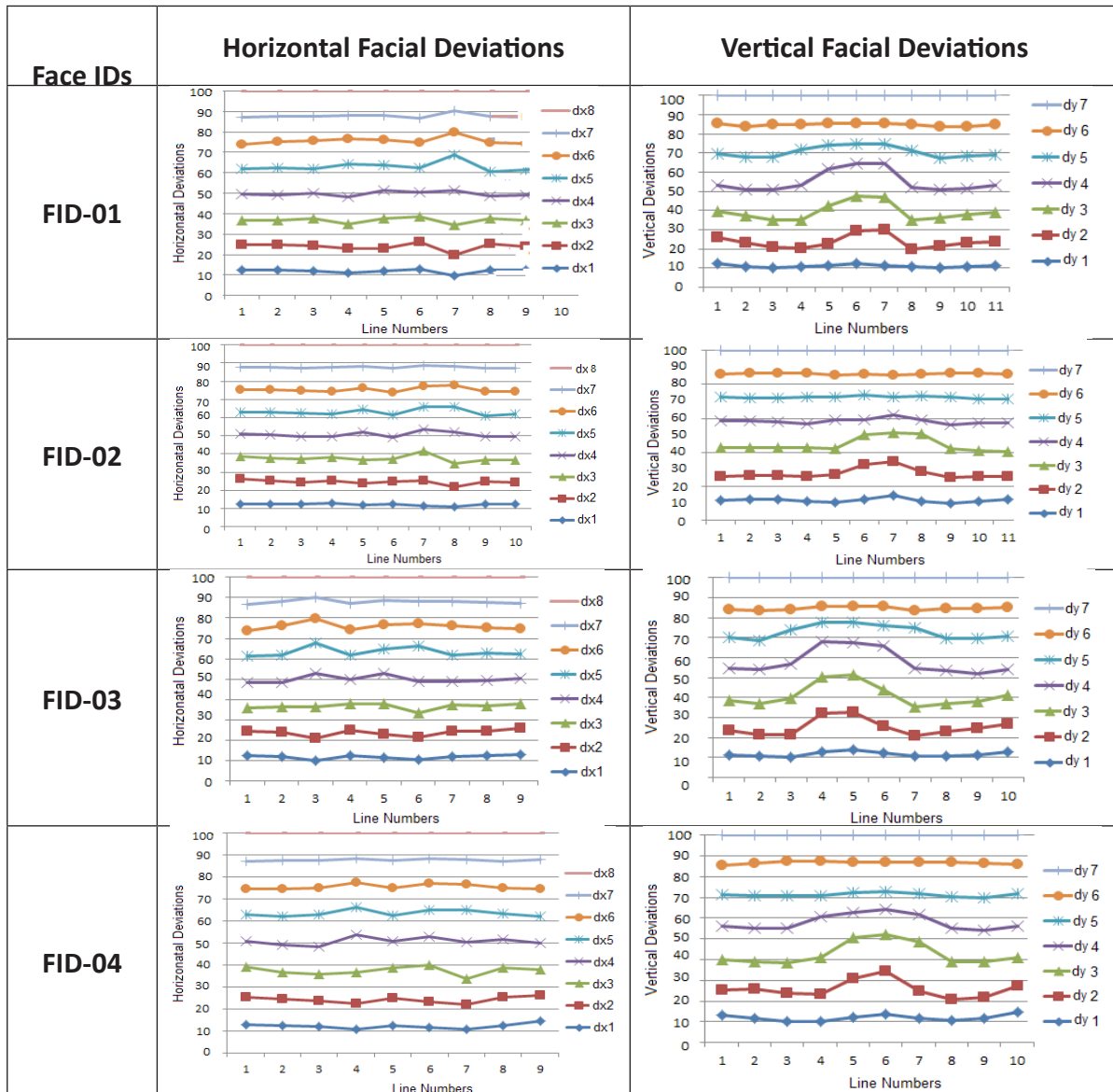

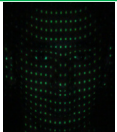

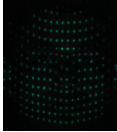



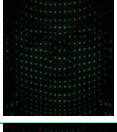

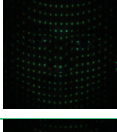

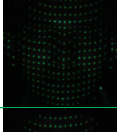
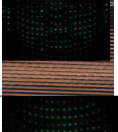
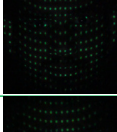
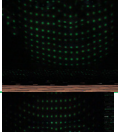
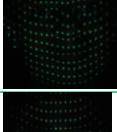
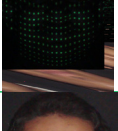
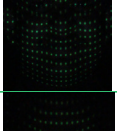

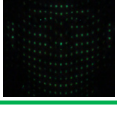


Table 2 and figure 5 show the final calculated values of decision parameters M & G , and Δd respectively, for curvature-based and differential versions of dotted raster-stereography, for faces FID-01 to FID-10. For test runs of both techniques in IPRL (image processing research lab), same set of sample faces was selected from the system database.

Table 2: Results of decision parameter values for both techniques

Face IDs	Sample Human Faces	Distorted Grids	Curvature Based Dotted Raster-stereography				Differential Dotted Raster-stereography		
			K_1 (cm^{-1})	K_2 (cm^{-1})	Mean (M) (cm^{-1})	Gaussian (G) (cm^{-1})	d_x (cm)	d_y (cm)	Δd (cm)
FID-01			14.4200	1.7300	8.0800	25.0100	6.66	46.51	46.98
FID-02			4.8541	1.2500	3.0500	6.0600	9.24	53.09	53.88
FID-03			29.3000	1.9000	15.6000	55.6000	7.66	67.45	67.88
FID-04			24.0572	02.0500	13.0500	48.1100	5.79	39.38	39.80
FID-05			22.2000	1.5000	12.0000	33.7500	9.53	49.67	50.57
FID-06			17.2367	2.1230	9.67985	36.59351	12.12	73.14	74.13
FID-07			19.1002	1.8222	10.4612	34.80438	6.31	36.15	36.69
FID-08			27.4513	1.0021	14.2267	27.50895	13.62	77.23	78.42
FID-09			24.0115	1.7698	12.89065	42.49555	11.41	68.19	69.13
FID-10			12.7687	1.8765	7.3226	23.96047	10.98	71.05	71.89

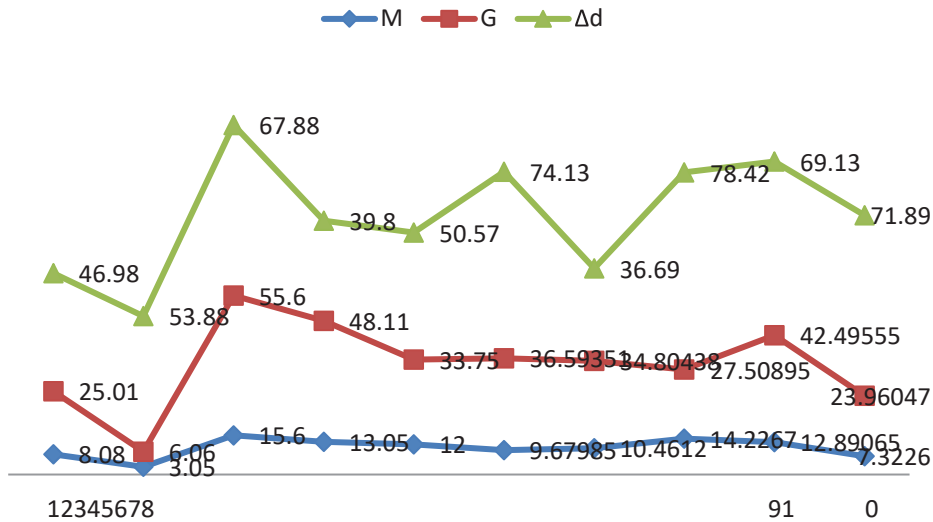


Figure 5: Decision Parameter values for Curvature-Based and Differential Versions of Dotted Raster-stereography

To get the better results in differential version of dotted raster-stereography, each test run was repeated three times. The resulting averages of d_x , d_y , and Δd values are mentioned in table 2. It was observed that generally, sufficient difference between the decision parameter (Δd) values for different faces existed that, facilitated easy recognition of different registered human faces from the system database. However, one exception was in the case of facial IDs FID-03 and FID-09, where the two decision parameter values were close i.e. having a difference of only 1.25 cm (= 69.13- 67.88), thus making unique identification of both faces rather difficult.

B Accuracy, Precision and Specificity

The accuracy, precision and specificity values of curvature-based and differential version of dotted raster-stereography techniques for the same sample of 10 faces are summarized in table 3. Each test run in IPRL was repeated three times in order to get consistent results.

Table 3: Accuracy, Precision and Specificity Values

	Curvature Based Dotted Raster-stereography	Differential Dotted Raster-stereography
Accuracy	96.30	80.00
Precision	97.44	87.50
Specificity	33.33	50.00

Out of the sample set of 10 faces, 7 were correctly identified. During test run, it was observed that face ID-09 was wrongly identified as face ID-03 due to close decision parameter values. Details of the study outcomes and the results of parameters for differential dotted raster-stereography are given in figure 6.

		Face Recognition System		
		Condition Positive	Condition Negative	
Differential Dotted Raster-stereography Outcomes	Test Outcomes Positive	True Positive TP = 07	False Positive FP = 01	Positive predictive value 87.50%
	Test Outcomes Negative	False Negative FN = 01	True Negative TN = 01	Negative predictive value 50.00%
	Accuracy 80.00%		Precision 87.50%	
	Sensitivity 87.50%		Specificity 50.00%	

Figure 6: Results of parameters for differential dotted raster-stereography

C Algorithm Execution Time

The curvature based and differential versions of dotted raster-stereography techniques were tested on the same set of 10 human faces from the database in IPRL. It was found that training and testing times for curvature-based technique were 260.5 sec and 2.1 sec respectively. In case of differential technique, training and testing times were found as 180.5 sec and 1.5 sec respectively.

D Discussion of Results

In comparison with the curvature-based version, the newly reported algorithm of differential dotted raster-stereography has lower accuracy but higher specificity. However, it is important that the differential version is faster than curvature-based version. This faster execution of algorithm is very important for saving time, if the application is to be used to recognize a very large number of human faces, such as those of employees in a big organization or of train / airline passengers. Another very important application could be security related, where access is allowed to registered persons only.

6 Conclusion

A new differential version of dotted raster-stereography technique has been reported, that performs simpler calculation as compared with the previously reported curvature-based version

of dotted raster-stereography. Consequently, the differential dotted raster-stereography is faster as compared to its curvature-based version, which is an important factor when performing face recognition operation over a very large-sized database. The accuracy of the new technique is lower, which needs to be worked on for improvement.

References

- [1] Zubairi, J. A. (2002). Applications of computer-aided rasterstereography in spinal deformity detection. *Image and Vision Computing*, 20(4), 319-324.
- [2] Hackenberg, L., Hierholzer, E., Pötzl, W., Götze, C., & Liljenqvist, U. (2003). Rasterstereographic back shape analysis in idiopathic scoliosis after anterior correction and fusion. *Clinical Biomechanics*, 18(1), 1-8.
- [3] Crawford, R. J., Price, R. I., & Singer, K. P. (2009). The effect of interspinous implant surgery on back surface shape and radiographic lumbar curvature. *Clinical Biomechanics*, 24(6), 467-472.
- [4] Melvin, M., Sylvia, M., Udo, W., Helmut, S., Paletta, J. R., & Adrian, S. (2010). Reproducibility of rasterstereography for kyphotic and lordotic angles, trunk length, and trunk inclination: a reliability study. *Spine*, 35(14), 1353-1358.
- [5] Kamal, S. A. (2010). An airport-passenger-screening system based on emitted IR and thermal radiation.
- [6] Wasim, M., Saeed, F., Aziz, A., & Siddiqui, A. A. (2018). Dotted Raster-Stereography. In *Encyclopedia of Information Science and Technology*, Fourth Edition (pp. 166-179). IGI Global.
- [7] Wasim, M., Kamal, S. A., & Shaikh, A. (2013). A security system employing edge-based rasterstereography. *International Journal of Biology and Biotechnology*, 10(4), 483-501.
- [8] Balla, P., Manhertz, G., & Antal, A. (2014). Defining of the shape of the spine using moiré method in case of patients with Scheuermann disease. *World Academy of Science, Engineering and Technology, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, 8(6), 348-353.
- [9] Betsch, M., Wild, M., Jungbluth, P., Hakimi, M., Windolf, J., Haex, B., & Rapp, W. (2011). Reliability and validity of 4D rasterstereography under dynamic conditions. *Computers in biology and medicine*, 41(6), 308-312.

- [10] Elad, D., Sahar, M., Einav, S., Avidor, J. M., Zeltser, R., & Rosenberg, N. (1989). A novel non-contact technique for measuring complex surface shapes under dynamic conditions. *Journal of Physics E: Scientific Instruments*, 22(5), 279.
- [11] Rankine, L., Liu, X. C., Tassone, C., Lyon, R., Tarima, S., & Thometz, J. (2012). Reproducibility of newly developed spinal topography measurements for scoliosis. *The open orthopaedics journal*, 6, 226.
- [12] Amor, B. B., Ardabilian, M., & Chen, L. (2006, August). New experiments on icp-based 3d face recognition and authentication. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on* (Vol. 3, pp. 1195-1199). IEEE.
- [13] Amberg, B., Knothe, R., & Vetter, T. (2008, September). Expression invariant 3D face recognition with a morphable model. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on* (pp. 1-6). IEEE.
- [14] Alyuz, N., Dibeklioglu, H., Gokberk, B., & Akarun, L. (2008, April). Part-based registration for expression resistant 3D face recognition. In *Signal Processing, Communication and Applications Conference, 2008. SIU 2008. IEEE 16th* (pp. 1-4). IEEE.
- [15] Lu, X., & Jain, A. K. (2005, January). Integrating range and texture information for 3D face recognition. In *Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on* (Vol. 1, pp. 156-163). IEEE.
- [16] Li, Y. A., Shen, Y. J., Zhang, G. D., Yuan, T., Xiao, X. J., & Xu, H. L. (2010, May). An efficient 3D face recognition method using geometric features. In *Intelligent Systems and Applications (ISA), 2010 2nd International Workshop on* (pp. 1-4). IEEE.
- [17] Smeets, D., Fabry, T., Hermans, J., Vandermeulen, D., & Suetens, P. (2010, August). Fusion of an isometric deformation modeling approach using spectral decomposition and a region-based approach using ICP for expression-invariant 3D face recognition. In *Pattern Recognition (ICPR), 2010 20th International Conference on* (pp. 1172-1175). IEEE.
- [18] Sun, Y., & Yin, L. (2008, June). 3D Spatio-Temporal face recognition using dynamic range model sequences. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on* (pp. 1-7). IEEE.
- [19] Yunqi, L., Dongjie, C., Meiling, Y., Qingmin, L., & Zhenxiang, S. (2009, December). 3D face recognition by surface classification image and PCA. In *Machine Vision, 2009. ICMV'09. Second International Conference on* (pp. 145-149). IEEE.
- [20] Mian, A. S. (2009, September). Shade face: multiple image-based 3D face recognition. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE(pp. 1833-1839).

- [21] Xue, Y., Tong, C. S., Chen, Y., & Chen, W. S. (2008). Clustering-based initialization for non-negative matrix factorization. *Applied Mathematics and Computation*, 205(2), 525-536.
- [22] Gaidhane, V. H., Hote, Y. V., & Singh, V. (2014). An efficient approach for face recognition based on common eigenvalues. *Pattern Recognition*, 47(5), 1869-1879.
- [23] Abdullah, M. F. A., Sayeed, M. S., Muthu, K. S., Bashier, H. K., Azman, A., & Ibrahim, S. Z. (2014). Face recognition with symmetric local graph structure (slgs). *Expert Systems with Applications*, 41(14), 6131-6137.

Concept Drift in Streaming Data: A Systematic Literature Review

Tariq Mahmood¹

Tatheer Fatima

Abstract

World is generating immeasurable amount of data every minute, that needs to be analyzed for better decision making. In order to fulfil this demand of faster analytics, businesses are adopting efficient stream processing and machine learning techniques. However, data streams are particularly challenging to handle. One of the prominent problems faced while dealing with streaming data is concept drift. Concept drift is described as, an unexpected change in the underlying distribution of the streaming data that can be observed as time passes. In this work, we have conducted a systematic literature review to discover several methods that deal with the problem of concept drift. Most frequently used supervised and unsupervised techniques have been reviewed and we have also surveyed commonly used publicly available artificial and real-world datasets that are used to deal with concept drift issues.

Index Terms: Concept drift, machine learning, streaming data, unlabeled streaming data.

1 Introduction

The world is evolving and getting smarter day by day, new and intelligent technologies like internet of things (IoT), Big Data and cloud computing are playing a major role in this evolution. "IoT, refers to the millions of physical devices like chips, sensors, smart phones, cameras and many more, these devices are capable of collecting and sharing data over the internet anywhere in the world". It is one of the most popular technologies of this digital era. IoT has brought the revolution in how business operates, it is now one of the major sources of the data, it produces, directs and drives data continuously. Unlike traditional batch-based flows, streaming data is more time sensitive and larger in volume, due to this more robust and efficient tools are needed to effectively analyze the ever-growing streaming data in different application domains and fields.

Machine learning has been introduced in an era of data deluge, where data is continuously being generated and stored in large volumes. To deal with this enormous amount of data, industries have moved towards machine learning techniques for data driven solutions. Many companies have adopted classification to effectually perform the task of predictive analytics and have gotten rid of cumbersome traditionally used analytical methods. The capability of generalizing and foretelling using data have made classification a desirable technique to solve many data directed business problems.

However, this generalization ability of a classifier makes a powerful presumption about the stability of the underlying distribution of the arriving data. According to which the data that is being used for training and testing the model should be Identically and Independently Distributed

¹Institute of Business Administration, Karachi | tmahmood@iba.edu.pk

(IID). Most of the machine learning algorithms assume that the relationship between the input (input features) and the output (i.e., target variable) remains static, but in real world data can change with time in unpredictable ways. This can degrade predictive performance of a model. The issue of varying the underlying relationships in the data is called concept drift. It was first proposed by (Jeffrey C. Schlimmer 1986), who aimed to point out that the same instances can be identified as noise data or non-noise information at separate times. The cause of these changes could be the variation in the hidden variables that could not be assessed directly. Areas where concept drift is involved includes recommendation systems, energy consumption, artificial intelligence systems with dynamic environment interaction, Fraud monitoring and anomaly detection systems, and biomedical signal analysis (e.g., neurogenerative diseases). For example, in cases of fraud detection, fraudsters are evolving and getting more creative with time and adopting new and advance techniques, machine learning model trained on the historical data will not be able to detect these changing patterns. The occurrence of concept drift has been perceived as the main reason of performance decaying in many data directed information systems that are being used for decision making and early warnings. In an evolving environment of big data, maintaining reliable and efficient data directed predictions and judgement facilities has become a critical problem. There is a need for detection of these changing concepts.

Researches to deal with concept drift have been increased a lot over the last decade, and many drift handling and adaptive learning techniques have been proposed. Adaptive learning means upgradation of the predictive model in an incremental manner to avoid encountering concept drift. Big data produces massive amount of data daily in a streamed fashion. But dealing with streaming data can be challenging. Machine learning can be used to perform tasks like querying, pattern recognition, real time analysis and predictive analytics. Industries have realized the potential of machine learning and are incorporating machine learning models to support business decisions. Many companies are using it to uncover hidden patterns, identify customer preferences, reveal market trends and analyze bigger and more complex data. Machine Learning models have made decision making easy and more robust. However, these learning models work in a dynamic setup, but within this era of technology things are changing at a fast pace and with this data is also changing, this can decompose the foretelling capabilities of a classifier with time and thus making it outdated.

Various researches related to drift-aware are available, however the focus is on supervised techniques. This survey aims to find out techniques for concept drift identification in not only labeled but also in unlabeled streams.

2 Background

Conventional machine learning has two main components: training/learning and prediction. Many of the algorithms assume that the data distribution stays the same over time. In other words, the examples that we see in the training set should resemble the observations we see in operation. However, streams are rarely stationary. The fast-changing surroundings of new products, new markets and new customer conduct can result in a population change or spontaneous change in the data which then changes the distribution of the data and can decrease model performance, this change in distribution is termed as Concept Drift. For an

instance, in ecommerce the customer's shopping behavior may change over time. For e.g. If we are predicting monthly sales of each product or product category, looking at the historical data such as promotion budget, the model may give good accuracy currently but it can easily decay with time as the customer psychology changes, it can be due to external factors such as trends, celebrity influence, seasons etc.

Machine learning models should identify concept drift and adjust to it so that a satisfactory model performance is retained. Learning with the concept drift has been categorized in 3 parts: concept drift detection (identify if drift has occurred), drift understanding (when, how, where it occurs) and drift adaptation (how to react to the occurring drift), this survey is focuses on the techniques to detect concept drift.

A *Concept drift definition*

In the domain of predictive analytics, an unpredictable variation in the statistical properties of the target label, that the algorithm is trying to forecast is termed as concept drift. The target variable that is being predicted is referred as the term concept. The input variables can also be termed as concepts, mostly this term is used for the target variable. Formally, concept drift is defined as follows:

At a certain time duration $[0,t]$, a group of examples is defined as $S_{0,t} = \{d_0, \dots, d_t\}$, where $d_i = (X_i, y_i)$ is a singular instance (or a data point), here X_i is denoting the input vector, y_i is the target variable, and $S_{0,t}$ follows a particular distribution $F_{0,t}(X,y)$. Concept drift occurs at timestamp $t + 1$, if there is a change in the distribution $F_{0,t}(X,y)$ that has $F_{0,t}(X,y) \neq F_{t+1,\infty}(X,y)$, denoted as $\exists t: P_t(X,y) \neq P_{t+1}(X,y)$.

Thus, the concept drift between time t and $t + 1$ is $\exists t: P_t(X,y) \neq P_{t+1}(X,y)$ where P_t is the joint distribution at time t between the input vector X and the target label y . Variation in the relationship components of input and target variables can define the variation in data.

the change in the classes' $P(y)$ prior probability,

the change in the class $P(X|y)$ conditional probabilities,

Thus, the prediction will be affected due to the change in the classes $P(y|X)$ posterior probabilities.

This gives us the concept of real and virtual concept drift. Variation in $P(y|X)$ denotes the occurrence of **Real concept**. This change can either be affected with the change in $P(X)$ or not. **Virtual drift** happens when there is change in the distribution of the input vector $P(X)$ without affecting $P(y|X)$.

B Types of concept drift

Commonly drift is categorized in four types that are shown below:

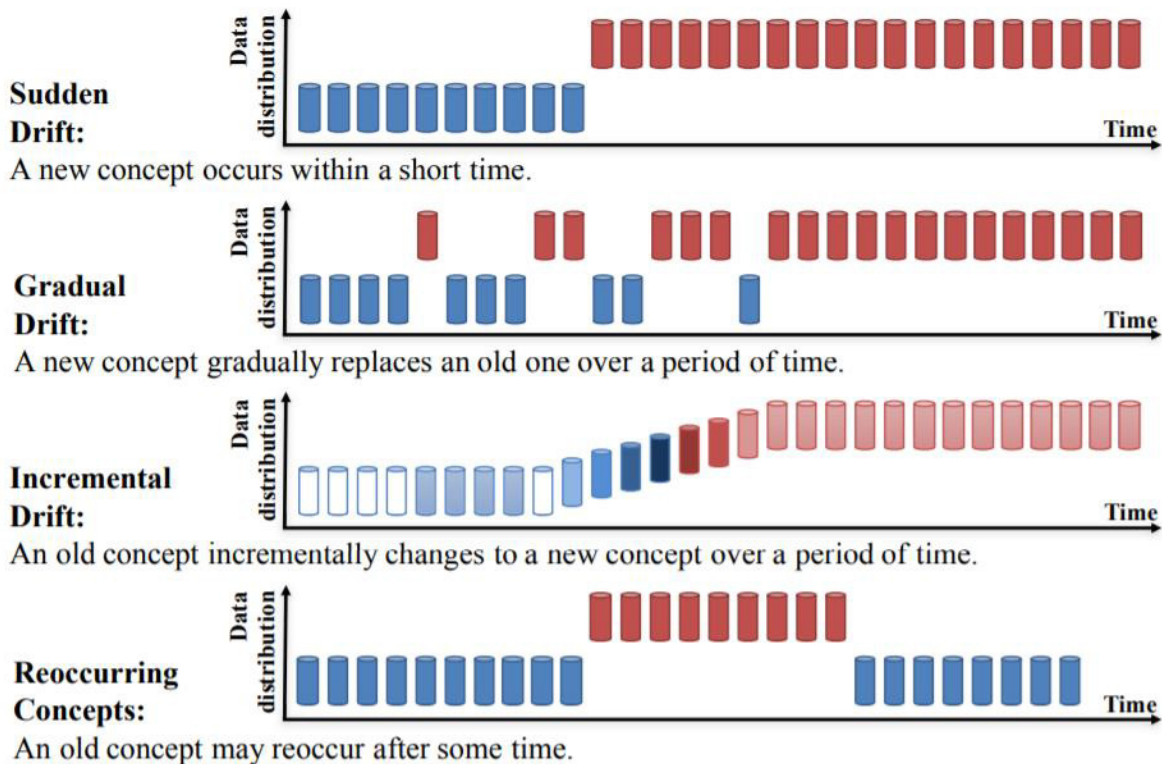


Figure 1: Categories of concept drift
Source: (Jie Lu 2018)

A sudden drift takes place when the distribution of the data is abruptly changed in a brief span of time. For example, a sudden drift may occur in a warehouse sensor data stream when an unseen abnormality is detected by it. An incremental drift occurs gradually in a longer span of time, an example of incremental drift is when the raiders change their fraudulent behaviors with time. As RFID chips are introduced, it changes patterns of raiders of committing fraud, as more customers get the new RFID card the new types of fraud will become more common until every customer has it, where the drift will stop. A gradual drift occurs when the starting distribution and the ending distribution happens simultaneously for some time, eventually transitioning to the ending distribution. An example would be customers shopping behavior, where sales figure fluctuates and shows this kind of drifts over different seasons. A recurring drift is any drift that repeats itself. An example would be energy consumption data where data fluctuates and reaches peaks during different times in a day. It is low at day and gets high at night. The focus of researches into concept drift adaptation in sudden/gradual/incremental drift is to increase the recovery rate and lower the drop-in accuracy during the concept transformation process. While, historical concepts are used in the case of recurring drift, the focus is on finding the best fitting historical concept, in the shortest span of time. The new concept may reemerge suddenly,

incrementally, or gradually. The phenomenon of “intermediate concept” is used to better exhibit the different types of concept drifts. A concept drift doesn’t have to occur at an exact timestamp, it can exist for a certain duration. Intermediate concept can be the mixture of initial and ending concepts. Thus, intermediate concepts can be used to represent the transition between initial concept and the final concept.

The above four categories are enough to define single class drift pattern, however class labels are not available in unsupervised drift detection, so there is a need to consider global data distribution change as well. To fulfill this purpose, two more concept drift types, stationary and non-stationary drifts are introduced as shown below:

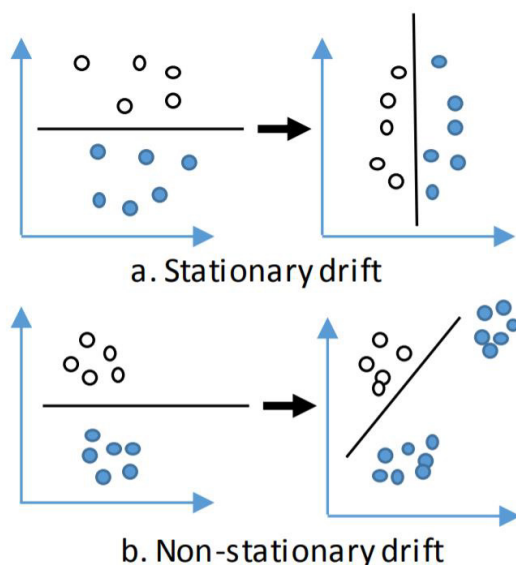


Figure 2: Stationary and Non-stationary drift
 Source: (Hanqing Hu 2018)

In stationary drift, concept drift has changed the distributions of individual classes, thus resulted in the change of data model. Nonetheless, both classes still take the similar space globally. While in the scenario of non-stationary drift, global distribution of the data has also changed along with the individual classes. A stationary drift can be sudden, incremental or gradual. There can be an abrupt change in the decision boundary forming a new hyperplane within the data space or it can slowly change over time. Likewise, in a non-stationary drift a new data space can be formed suddenly if large amounts of data appear at a new region in a short time, or it can be formed gradually by shifting the shape of existing data space slowly.

C Concept Drift Detection Algorithms

Numerous performance-based drift identification methods and algorithms have been suggested in literature. These algorithms can detect all kinds of drift but require target labels.

I) *Drift detection method (DDM)*

This method is proposed by (João Gama 2004). It is based on the detection of probability distributions of samples using errors in the learning process. It models no of errors as a binomial random variable. This approach requires class labels to monitors the increase in error in the model predictions to identify drift. The stated method does not look at the change in distribution to detect drift that's why it can identify all kinds of drift.

The error rate is the probability of misclassifying (p_i), which is calculated for every point i . The standard deviation of the error rate is $s_i = \sqrt{p_i(1 - p_i)/i}$. While training the model DDM maintains 2 global variables, p_{min} and s_{min} , whenever a new sample is treated the values of these variables are updated if the condition $p_i + s_i > p_{min} + s_{min}$ is satisfied. This method is based on an assumption that as the count of the samples rises the error rate of the algorithm (p_i) will reduce in a stationary environment, if there is a noticeable rise in the error of the algorithm it means that there is a change in class distribution. Hence the values of p_i and s_i is stored and when $p_i + s_i \geq p_{min} + 2 * s_{min}$ a warning level is triggered and the examples are then stored in anticipation of a possible drift. And when $p_i + s_i \geq p_{min} + 3 * s_{min}$ drift level is triggered, the model is re-trained using the examples that were stored after the warning level was triggered.

Similarly, EDDM (M. Baena-Garcia 2006) was proposed, it is also based on classification error monitoring, however to give better detection accuracy, EDDM uses the distance among the two errors, which makes it robust in identifying gradual drift. While DDM-OCI was proposed (S. Wang 2013) to address the limitation of DDM when data is imbalanced.

II) *Adaptive windowing (ADWIN)*

In contrast to the above techniques, which operate in an incremental fashion that takes one instance at a time, window-based approaches use a chunk based or sliding window approach over the recent samples, to detect changes. Windows of prediction errors for each partition it calculates mean error and compares the difference. (Albert Bifet 2007) presented the ADWIN2 algorithm, an improved version of ADWIN algorithm. ADWIN2 has a variable sized window: it grows or shrinks when no change or concept drift is detected, respectively.

III) *Linear four rate (LFR)*

LFR (Wang and Abraham 2015) is better for class imbalance then DDM and EDDM. The LFR strategy is straightforward that if data is stationary then the contingency table should stay constant. It uses four rates true positive rate (tpr), true negative rate (tnr), positive predictive value (ppv) and negative predictive value (npv) to identify the concept drift. During a stationary concept (i.e., $P(X_t, y_t)$ stays constant), $P_{tpr}, P_{tnr}, P_{ppv}, P_{npv}$ doesn't changed. Hence, a noticeable variation in any P_* ($* \in tpr, tnr, ppv, npv$) shows that the variation in underlying joint distribution $P(X_t, y_t)$, or concept has happened. The 4 rates can be computed as follows: $P_{tpr} = TP/(TP+FN)$, $P_{tnr} = TN/(TN+FP)$, $P_{ppv} = TP/(FP+TP)$ and $P_{npv} = TN/(TN+FN)$. All the mentioned characteristic rates in $P_* = \{P_{tpr}, P_{tnr}, P_{ppv}, P_{npv}\}$ are equal to 1, if there is no misclassification.

IV) Margin density drift detection (MD3)

(Kantardzic 2015) proposed MD3. In order to monitor the variation in the distribution of the arriving instances, MD3 traces the changes in classifier boundary. The thought behind this method is that, as the count of instances inside the margin will rise or reduce, a variation in $p(y|X)$ will be observed. A gradual drift will occur if the number of instances grows inside the margin, causing the class distribution to move closer to the boundary, as shown in (Figure. 3b). However, if an entire class distribution moves towards a different area of a feature space, a reduction in the count of instances (Figure. 3c) will be seen, this drift is said to be sudden. Further inquiry is needed in both the cases. MD3 uses unlabeled samples and is good for identifying gradual and stationary drift.

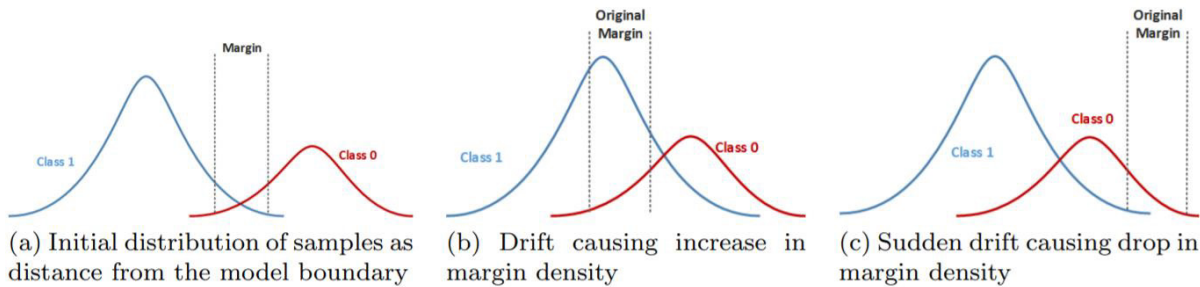


Figure 3: Drifting concepts and their effect on margin density

Source: (Kantardzic 2015)

A sliding window divides the stream in samples, the size of the window is S , which moves at a rate of S_r samples. A Decision function is used to calculate each sliding window's margin density (ρ) e.g. for SVM, considering the kernel is linear margin density is defined as,

$$\rho = \frac{\#samples\ with\ abs(w \cdot x + b) \leq 1}{\#samples}$$

A block of unlabeled instances is given to the algorithm at time t . It calculates the margin density measure of the current block and triggers an alert if $\rho_{min} - \rho_{max} > \theta_\rho$. This means a drift has occurred. The range of margin density ($\rho_{min} \rho_{max}$) since the last drift is stored as well as the threshold to signal drift (θ_ρ). The current classifier (SVM with linear kernel the classification model is given as w, b) shows the concept at $t-1$

V) Clustering based drift detection

(Joung Woo Ryu 2012) presented method. The clusters are formed with the existing data. As new data arrives, model tries to classify the new instances into existing clusters. If the number of the new instances which didn't fit into any of the existing clusters have grown enough that they can create a new cluster on their own, then a drift is said to be detected. Density-based detection is based on an assumption that the samples within the same cluster belongs to the same class. Every cluster C is determined by a radius $radc$ and a cluster density dc

$$radc = \text{distance between cluster center and the furthest sample from the center}$$

$$dc = \text{number of samples in cluster divided by } radc$$

$radc$ is calculated using Euclidean distance. If the distance between the cluster center and the new sample is less than $radc$, then it is assigned to that cluster otherwise it is viewed as an un-allocated sample and is marked as $\sim s$. As time passes and the number of $\sim s$ have increased enough to form a new cluster then it can be said that concept drift has occurred.

VI) *Grid based drift detection*

This method is proposed by (Mohammad Masud 2011). Every feature/input of the data is divided into fixed intervals to create a grid of data space. On the basis of the feature values new samples are assigned to their specific cell in the grid. If the number of samples in a cell increases above a threshold level p a potential concept drift has occurred. Grid based detection monitors the shape of the data, if the new data arrives in the current dense cells of the grid it means the existing shape of the data has not changed and if the new data arrives in the less dense cells then it can be concluded that the shape of the data is changing and thus the concept drift is identified.

D *Concept Drift Adaptation*

Concept drift adaptation refers to updating the existing models in order to handle the detected drifts. It is also called drift handling and drift learning. The two main techniques for concept drift adaptation are single learning model adaptation and ensemble learning for concept drift adaptation.

I) *Single learning model adaptation*

In single learning model adaptation only one learning model is activated at a time, to perform the classification and prediction tasks. This approach further be classified in to two categories, instance selection and incremental learning.

- *Instance selection and weighting*

The simplest way to handle concept drift is, to re-learn a fresh model with the most recent instances. A window-based approach is often adopted to store the most relevant instances with respect to the current concept and then retrain a fresh model with it to replace the outdated one. Paired Learners (Stephen H. Bach 2008) follows this method and uses two learners: the stable learner and the reactive. If the new instances are being falsely predicted by the stable learner and the same instances are being correctly predicted by the reactive learner, than the stable learner is substituted by the reactive one. Similarly, Self Adjusting Memory kNN (SAMkNN) (Viktor Losing 2016), gives weightage to the learners based on their performance on the most recent time frame.

Incremental Learning

A substitute to training an entire new model, is to build a model that incrementally updates itself and adapts to the recent changes in the data. This approach is more robust than retraining an entire new model when the drift is small or reoccurring. Majority of the algorithms in this group are based on decision trees, because trees are capable of examining each sub-region separately. Algorithms like Very Fast Decision Tree classifier (VFDT) and CVFDT (Geoff Hulten 2001) are proposed for this purpose.

II) Ensemble learning for concept drift adaptation

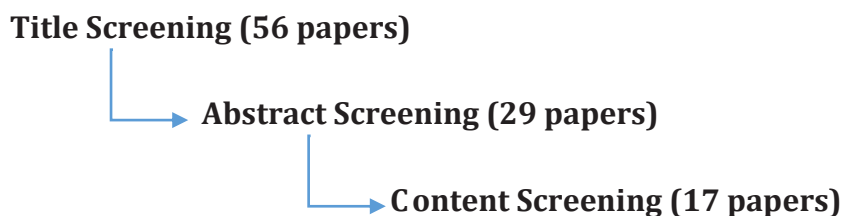
In contrast to single learning model adaptation, ensemble learning utilizes multiple models to make prediction at each point in time. The reason of using ensemble methods to deal with concept drift is that it can save the great deal of labor that is required to retrain a new model for recurring concepts. Ensemble is the most popular method being used recently. It consist of several base learners that may have distinct parameters. The final prediction is given by combining the output of each ensemble using votes of certain weights. A survey about this research topic is (Bartosz Krawczyk 2017).

3. Research Methodology

The purpose of the survey was to study the algorithms and methodologies proposed and are being used for concept drift detection. Research papers were searched on the most commonly used and popular platforms for Computer Science which include IEEE, Elsevier, Springer, Taylor and Francis, Wiley, ACM, Arxiv.org and Google Scholar using search queries:

- Concept drift
- Concept change
- Concept drift machine learning
- Concept change in machine learning
- Concept drift machine learning streaming data
- Concept change in machine learning streaming data
- Concept drift machine learning unlabeled streaming data
- Concept change machine learning unlabeled streaming data

After searching papers on mentioned platforms, the following filtration technique was applied to extract the most relevant papers.



A summary regarding all the work done is given below, Figure 4-7 represents the distribution of articles over the years and differentiated by digital sources.

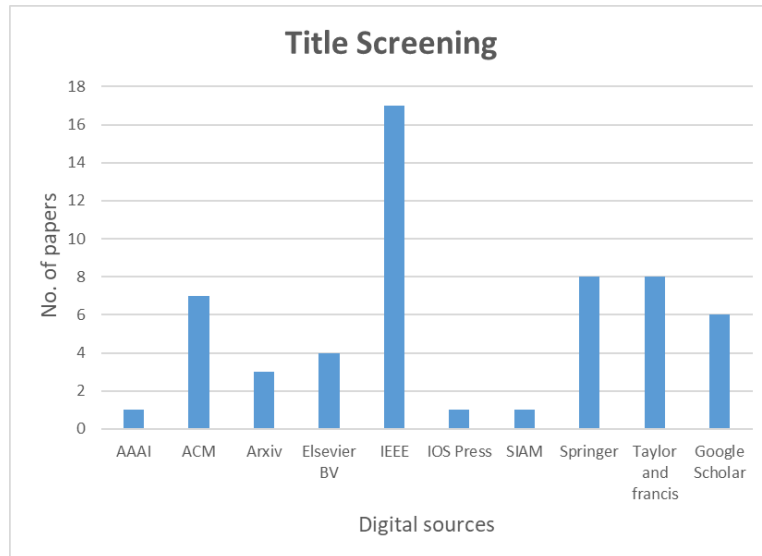


Figure 4: Distribution of 56 papers in the first stage of filtration differentiated by digital sources.

Figure. 5 shows the number of papers published by year. One can view that research in the field of concept drift has been focused even more in the recent years, hence showing that the scientific community is taking more interest in it now.

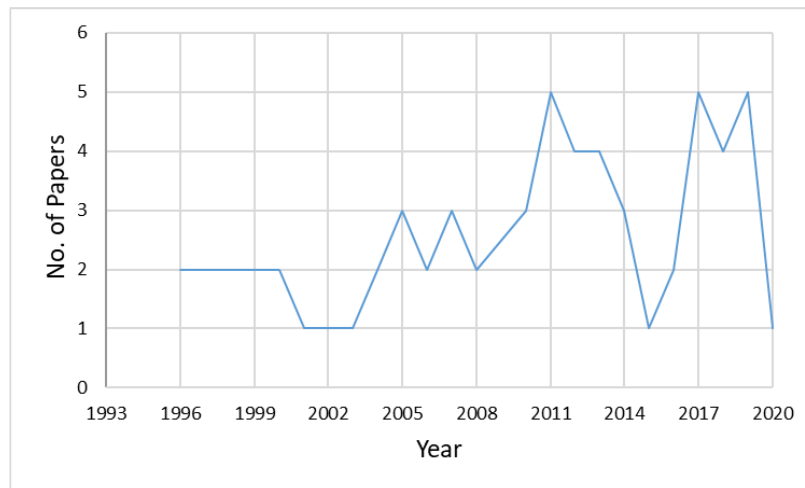


Figure 5: Papers separated by years

Figure. 6 shows the ratio of learning approaches used in the papers. It can be evidently seen, that in the area of concept drift supervised learning is by far the most used technique. The research in semi-supervised and unsupervised techniques for concept drift detection is very less as compared to supervised.

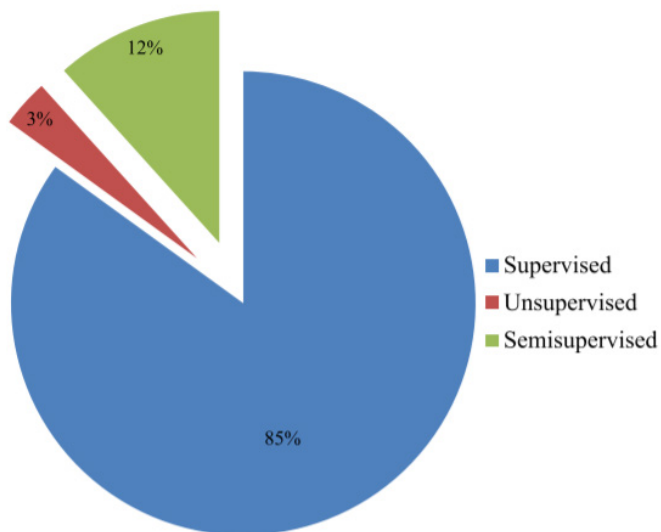


Figure 6: Ratios regarding distinct learning approaches applied for concept drift.

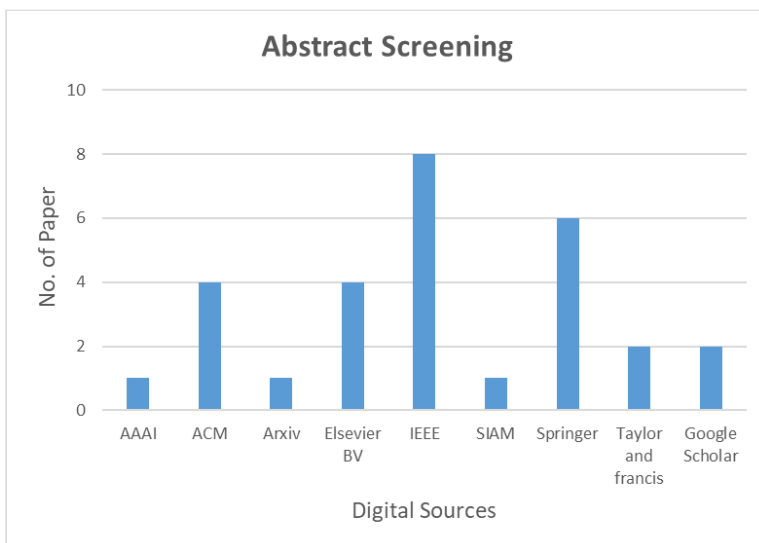


Figure 7: Distribution of 29 papers in the second stage of filtration with respect to digital sources.

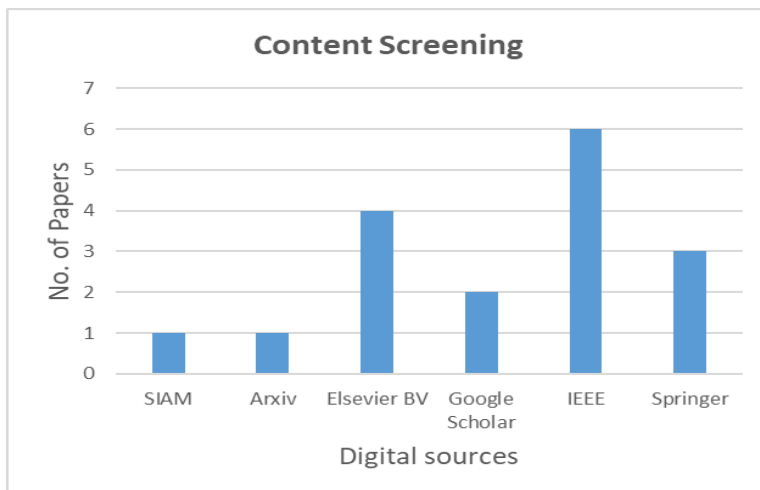


Figure 8: Distribution of 17 papers in the last stage of filtration with respect to digital sources.

4 Results

In this section, a summary of all the related articles is presented and further discussed. Table 3 contains all the basic information related to the final 17 papers that are reviewed and Table 4 contains the detailed information.

Table 1: Basic information about 17 papers

Paper	Publication Type	Venue	Publisher
(Shujian Yu 2017)	Book Chapter	Book Chapter published on 30th June 2017 in Proceedings of the 2017 SIAM International Conference on Data Mining	SIAM
(L. K. Geoffrey I. Webb 2017)	Journal article	Published in 2017	Arxiv
(João Gama 2004)	Book Chapter	Book chapter published 2004 in Advances in Artificial Intelligence – SBIA 2004 on pages 286 to 295	Springer
(M. Baena-Garcia 2006)	Article	Published in 2006	Google Scholar
(Stanley 2003)	Journal article	Published in 2003	Google Scholar
(Hanqing Hu 2018)	Conference Paper	Conference Paper Published in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)	IEEE
(Kantardzic 2015)	Journal article	Journal Article published 2015 in Procedia Computer Science volume 53 on pages 103 to 112	Elsevier BV
(Mohammad Masud 2011)	Journal article	Journal Article published Jun 2011 in IEEE Transactions on Knowledge and Data Engineering volume 23 issue 6 on pages 859 to 874	IEEE
(Joung Woo Ryu 2012)	Book Chapter	Book Chapter published 2012 in Convergence and Hybrid Information Technology on pages 533 to 540	Springer
(Maciej Jaworski 2017)	Proceedings article	Proceedings Article published Nov 2017 in 2017 IEEE Symposium Series on Computational Intelligence (SSCI)	IEEE
(Jie Lu 2018)	Journal article	Journal Article published 2018 in IEEE Transactions on Knowledge and Data Engineering	IEEE
(M. K. Tegjyot Singh Sethi 2017)	Journal article	Journal Article published Oct 2017 in Expert Systems with Applications volume 82	Elsevier BV

Paper	Publication Type	Venue	Publisher
(Xindong Wu 2012)	Journal article	Journal Article published Sep 2012 in Neurocomputing volume 92 on pages 145 to 155	Elsevier BV
(R. H. Geoffrey I. Webb 2016)	Journal article	Journal Article published Jul 2016 in Data Mining and Knowledge Discovery volume 30 issue 4 on pages 964 to 994	Springer
(M. K. Tegjyot Singh Sethi 2016)	Proceedings article	Proceedings Article published Jul 2016 in 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)	IEEE
(Adriana Sayuri Iwashita 2019)	Journal article	Journal Article published 2019 in IEEE Access volume 7 on pages 1532 to 1547	IEEE
(Niloofer Mozafari 2011)	Journal article	Journal Article published Aug 2011 in Computers & Mathematics with Applications volume 62 issue 4 on pages 1655 to 1669	Elsevier BV

Table 2: Detailed information of final 17 papers.

Paper Name	Contribution Level*	Data Type	Framework	Experiment	Algorithm
(Shujian Yu 2017)	Average	Streaming, labeled data	Yes	Yes	HLFR
(L. K. Geoffrey I. Webb 2017)	Good	-	No	No	-
(João Gama 2004)	Average	Streaming, labeled data	Yes	Yes	DDM
(M. Baena-Garcia 2006)	Average	Streaming, labeled data	Yes	Yes	EDDM
(Stanley 2003)	Average	Streaming, Unlabeled data and Labeled Data	Yes	Yes	CDC
(Hanqing Hu 2018)	Good	Streaming, Unlabeled data	Yes	Yes	EFDD
(Kantardzic 2015)	Good	Streaming, Unlabeled data	Yes	Yes	MD3
(Mohammad Masud 2011)	Average	Streaming, Unlabeled data	Yes	Yes	ECSMiner
(Joung Woo Ryu 2012)	Good	Streaming, Unlabeled data	Yes	Yes	Active Ensemble Learning

Paper Name	Contribution Level*	Data Type	Framework	Experiment	Algorithm
(Maciej Jaworski 2017)	Good	Streaming, Unlabeled data and Labeled Data	Yes	Yes	RBM
(Jie Lu 2018)	Good	-	No	No	-
(M. K. Tegjyot Singh Sethi 2017)	Good	Streaming, Unlabeled data	No	Yes	MD3
(Xindong Wu 2012)	Average	Streaming, Unlabeled data	Yes	Yes	SUN
(R. H. Geoffrey I. Webb 2016)	Good	-	No	No	-
(M. K. Tegjyot Singh Sethi 2016)	Average	Streaming, Unlabeled data and Labeled Data	Yes	Yes	-
(Adriana Sayuri Iwashita 2019)	Good	-	No	No	-
(Niloofer Mozafari 2011)	Average	Streaming, Unlabeled data	No	Yes	PSCCD

*Contribution level is based on the personal opinion, regarding the relevancy with this survey

Recently adaptive models and ensembles techniques are being researched more for concept drift adaptation while re-training of a model is being discouraged by most of the researchers. Very few unsupervised drift identification approaches have been discussed in literature. Techniques like Margin density, clustering based detection, grid-based detection and other ensemble frameworks provide satisfactory results. They give almost the same results in comparison to the benchmarking labeled techniques such as DDM. However, these techniques have been evaluated on synthetic datasets. More practical implementation of them is required to further comment on their change detection performance.

5 Conclusions and Future Work

This survey directly supports the researchers in understanding concept drift, it gives an overall view of the developments in concept drift learning. Most of the existing drift detection and adaptation techniques are proposed for labeled data streams. However, very few researches have been done to address the problem of concept drift in unsupervised or semi-supervised streams. In this survey we have tried to cover as many techniques as we can to address drift detection in unlabeled streaming data. We believe this paper provides researchers with basic understanding of concept drift and gives a know-how on applying drift identifying techniques on different domains persisting different challenges. We have tried to cover the maximum progress made in the research pertaining to this field.

Natural future research direction is obviously, experimentally comparing the above reviewed concept drift identification approaches for unlabeled data streams. Unsupervised drift detection methods need to be explored more. We have reviewed Restricted Boltzmann Machine for change detection, other areas of deep learning can also be explored for this purpose.

References

- [1] Adriana Sayuri Iwashita, Joao Paulo Papa. 2019. "An Overview on Concept Drift Learning." *IEEE Access (IEEE)* 7: 1532 to 1547.
- [2] Albert Bifet, Ricard Gavaldà. 2007. "Learning from Time-Changing Data with Adaptive Windowing." *Proceedings of the 2007 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics.*
- [3] AWS. n.d. What is Streaming Data? Accessed 2020. <https://aws.amazon.com/streaming-data>.
- [4] Bartosz Krawczyk, Leandro L. Minku, João Gama, Jerzy Stefanowski, Michał Woźniak. 2017. "Ensemble learning for data stream analysis: A survey." *Information Fusion (Elsevier BV)* 37: 132 to 156.
- [5] Geoff Hulten, Laurie Spencer, Pedro Domingos. 2001. "Mining time-changing data streams." *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01.* ACM Press.
- [6] Geoffrey I. Webb, Loong Kuan Lee, François Petitjean, Bart Goethals. 2017. "Understanding Concept Drift." *Arxiv abs/1704.00362 (Arxiv).*
- [7] Geoffrey I. Webb, Roy Hyde, Hong Cao, Hai Long Nguyen, Francois Petitjean. 2016. "Characterizing concept drift." *Data Mining and Knowledge Discovery (Springer Science and Business Media)* 30 (4): 964 to 994.
- [9] Gerhard Widmer, Miroslav Kubat. 1996. "Learning in the presence of concept drift and hidden contexts." *Machine Learning (Springer)* 23 (1): 69 to 101.
- [10] Hanqing Hu, Mehmed Kantardzic, Lingyu Lyu. 2018. "Detecting Different Types of Concept Drifts with Ensemble Framework." *17th IEEE International Conference on Machine Learning and Applications (ICMLA).* IEEE.
- [11] Jeffrey C. Schlimmer, Richard Granger. 1986. "Beyond Incremental Processing: Tracking Concept Drift." (AAAI).
- [12] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, Guangquan Zhang. 2018. "Learning under Concept Drift: A Review." *IEEE Transactions on Knowledge and Data Engineering (IEEE).*

- [13] João Gama, Pedro Medas, Gladys Castillo, Pedro Rodrigues. 2004. "Learning with Drift Detection." In *Advances in Artificial Intelligence – SBIA 2004*, 286 to 295. Springer.
- [14] Joung Woo Ryu, Mehmed M. Kantardzic and Myung-Won Kim. 2012. "Efficiently Maintaining the Performance of an Ensemble Classifier in Streaming Data." In *Convergence and Hybrid Information Technology*, 533 to 540. Springer.
- [15] Kantardzic, Tegjyot Singh Sethi and Mehmed. 2015. "Don't pay for validation: Detecting drifts from unlabeled data using margin density." *Procedia Computer Science (Elsevier BV)* 53: 103 to 112.
- [16] M. Baena-Garcia, J. del Campo-Avila, R. Fidalgo, A. Bifet, R. Gavaldá, and R. Morales-Bueno. 2006. "Early Drift Detection Method." *StreamKDD*. 77 to 86.
- [17] Maciej Jaworski, Piotr Duda, Leszek Rutkowski. 2017. "On Applying the Restricted Boltzmann Machine to Active Concept Drift Detection." *Symposium Series on Computational Intelligence (SSCI)*. IEEE.
- [18] Mohammad Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani M. Thuraisingham. 2011. "Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints." *IEEE Transactions on Knowledge and Data Engineering (IEEE)* 23 (6): 859 to 874.
- [19] Niloofar Mozafari, Sattar Hashemi, Ali Hamzeh. 2011. "A Precise Statistical approach for concept change detection in unlabeled data streams." *Computers & Mathematics with Applications (Elsevier BV)* 62 (4): 1655 to 1669.
- [20] S. Wang, L. L. Minku, D. Ghezzi, D. Caltabiano, P. Tino and X. Yao. 2013. "Concept drift detection for online class imbalance learning." *The 2013 International Joint Conference on Neural Networks (IJCNN)*. IEEE.
- [21] Shujian Yu, Zubin Abraham. 2017. "Concept Drift Detection with Hierarchical Hypothesis Testing." In *Proceedings of the 2017 SIAM International Conference on Data Mining*, 768 to 776. SIAM.
- [22] Stanley, Kenneth O. 2003. "Learning Concept Drift with a Committee of Decision Trees."
- [24] Stephen H. Bach, Marcus A. Maloof. 2008. "Paired Learners for Concept Drift." *2008 Eighth IEEE International Conference on Data Mining*. IEEE.
- [25] Tegjyot Singh Sethi, Mehmed Kantardzic. 2017. "On the reliable detection of concept drift from streaming unlabeled data." *Expert Systems with Applications (Elsevier BV)* 82: 77 - 99.

- [26] Tegjyot Singh Sethi, Mehmed Kantardzic, Elaheh Arabmakki. 2016. "Monitoring Classification Blindspots to Detect Drifts from Unlabeled Data." IEEE 17th International Conference on Information Reuse and Integration (IRI). IEEE.
- [27] Viktor Losing, Barbara Hammer, Heiko Wersing. 2016. "KNN Classifier with Self Adjusting Memory for Heterogeneous Concept Drift." 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE.
- [28] Wang, Heng, and Zubin Abraham. 2015. "Concept Drift Detection for Streaming Data." 2015 International Joint Conference on Neural Networks (IJCNN). IEEE.
- [29] Xindong Wu, Peipei Li, Xuegang Hu. 2012. "Learning from concept drifting data streams with unlabeled data." *Neurocomputing (Elsevier BV)* 92: 145 to 155.

Two-Dimensional Wavelet based Medical Videos using Hidden Markov Tree Model

Rubab Fatima Bangash¹

Imran Tauqir

Azka Maqsood

Abstract

Wavelet based statistical image denoising is a vital preprocessing technique in real world imaging. During the acquisition of most medical videos, the device introduces noise, as a result of additional image capturing techniques, resulting in the poor quality of video for examination. So, we required a trade-off between preservation and noise reduction of the actual image content to retains all the necessary information. The existing techniques are based on time-frequency domain where the wavelet coefficients need to be independent or jointly Gaussian. In denoising arena there is a need to exploit the temporal dependencies of wavelet coefficients with non-Gaussian nature. Here we present a YCbCr (Luminance-Chrominance) based denoising strategy on Hidden Markov Model (HMM) that is based on Multiresolution Analysis in the framework of Expectation-Maximization algorithm. Proposed algorithm applies the denoising technique independently on each frame of the video. It models the Non-Gaussian statistics of each wavelet coefficient and captures the statistical dependencies between coefficients. Denoised frames are restored inversely by processing the wavelet coefficients. Significant results are visualized through objectives as well as subjectives analysis. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Metric (SSIM) parameters are used for the quality assessment of proposed method in comparison with the Red Gray & Blue (RGB) scale video coefficients.

Key Words: Discrete Wavelet Transform, Expectation Maximization, Hidden Markov Tree Model, Video Denoising.

1 Introduction

Video denoising is based on time-frequency data of a video signal. As Far as Image denoising is accomplished by various methods: Time-domain, Frequency-domain, and Time-Frequency combination. Spatial domain methods do not account for the temporal correlation between the frames [1-3]. From an image Spatial noise is being successfully removed by Spatial filters resulting in achievement of high gain failed in restoring the edges particularly in less noisy areas [4]. Time domain techniques considered inter-frame correlation and performed well for motionless videos [5].

Temporal filters failed to remove the noise and produced the fewer blocking artifacts, caused blurring. On the other hand, in case of motioned videos, a temporal filter was not able to give good results in noise removing and delivered fewer blocking artifacts and caused blurring. Hence improved denoising algorithm is need of time, in order to improve the performance of image processing [6-8]

¹National University of Sciences & Technology | fatimabangash@yahoo.com

A new denoising method for medical images e.g. Ultrasound, X-Ray and Magnetic Resonance Imaging (MRI) images, was based on Daubechies Complex Wavelet Transform [9]. Statistical Data Warehouse (SDW) wavelet significantly removed the noise and preserved the details i.e. the local shifts and orientations were preserved with high computational time.

Loizou and Pattichis [10] recommended the DsFlsmv (Mean and variance local statistics despeckle filter), followed by DsFgf4d (Geometric despeckle filtering), and DsFlsminsc (Minimum Seckle Index Homogeneous Mask Despeckle Filter). The proposed method, improved the class separation between the asymptomatic and symptomatic classes. However, due to average filtering, sharp features and noisy boundaries were left unfiltered.

Methods in image denoising is primarily centered around wavelet transform. A denoising technique based on Double Density Dual Tree Complex Wavelet Transform (DDDT-CWT) [11] YCbCr and YUV space was implemented as multi-directional wavelet transform, where the edges and structural contents were restored. However, degradation in performance was seen significantly in real time scenarios.

Medical videos i.e. ultrasound, radiology and capsule endoscopy etc. are subject to noise attenuation. To address this, certain filters are used on ultrasound videos of Common Carotid Artery (CCA) were DsFkuwahara (Despeckling Kuwahara Filter), DsFhmedian (Despeckling Hybrid-Median Filtering), DsFlsmv (Mean and Variance Local Statistics Despeckle Filter) and DsFsrads (Speckle Reducing Anisotropic Diffusion) [12]. Resulting in better visual quality and improved performance in real time videos

Previously, Rabbani Gazor [13] used the Local Bessel K-Form Minimum Mean-Squared Error (BKMMSEL) and Local Bessel K-Form Maximum A-Posteriori (BKMAPL) functions on local Bessel K-Form Density (BKF), for noise free modelling of Three-Dimensional (3D) discrete complex wavelet coefficients in each sub-band. The tested video sequences were corrupted with different types of noises i.e. Additive White Gaussian Noise (AWGN), Poisson, non-stationary and speckle noise. Noise reduction was enhanced with the increased of computational outflow.

Denoising in Computerized Tomography (CT) images was done by another technique, with edge conservation in tetrolet domain (Haar-Type Wavelet Change) [14]. In this a locally adaptive shrinkage rule was applied on high frequency tetrolet coefficients to lessen the noise more viably while preserving the edges and geometrical structures. However, the updated procedure was slow for large objects.

Searching for efficient methods for image denoising is still a challenge. A comparison [15] on the productivity of wavelet based thresholding techniques with presence of speckle noise for different wavelet families i.e. Haar, Morlet, Symlet, Daubechies in de-noising for medical imaging in resonance of brain was proposed. It was found that wavelet transform was proficiently better in analyzing images at various resolutions but the edges were not restored and caused blurring.

A Hidden Markov Tree (HMT) model with 2D-Discrete Wavelet Transform (DWT) was implemented using HMT model in context with Expectation-Maximization algorithm [16] independently on each video frame. Thus, bringing about fast execution with less computational

time, while the improvement needs to be accomplished for enhanced video denoising.

Michele Claus recommended Net (ViDeNN) [17] approach for Video Denoising Thereby, combining two networks, performing first Single Frame Spatial Denoising and subsequently Temporal Denoising over a window of three frames, all in a single feed-forward process. However, the limitations of this method were additional computational power.

Here we have used the above mentioned technique which recommends a spatial-temporal filtering framework that considerably removes speckle noise from images and videos. By exploiting, the dependencies between wavelet coefficients, better performance has been accomplished. The proposed strategy manages the non-Gaussian behavior of wavelet coefficients that are frequently experienced and it gives a proficient outcome for despeckling of images The results display that the proposed strategy removed noise and it also retained almost all the structural information of each video frame.

In Section-2, the denoised wavelet model is discussed and in section-3, wavelet coefficients are modeled, elaborated by using HMT model. Later in section-4, proposed denoising model is defined. In Section-5 the results of experimentation of proposed model are discussed. In last section, this paper is concluded.

2 Statistical Video Modelling

Wavelet Transform: 2D-DWT has been utilized in numerous restorative imaging applications. Images are decomposed into three level coefficients i.e. several high frequency sub-bands and one low frequency sub-band (LL-sub band) by 2D-DWT (Figure 1). LL-sub band contains the most data in concentrated of highest level known as approximated DWT.

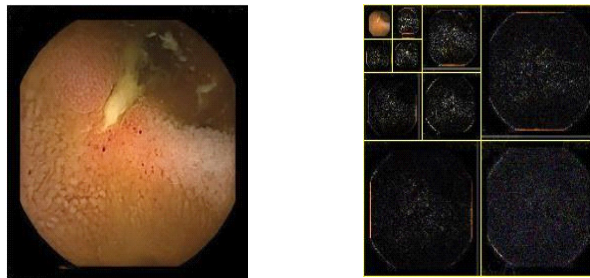


Figure 1: Three Level DWT Decomposition

Hidden Markov Tree Model: When the wavelet coefficients of the images are constructed statistically, HMM captures the Non-Gaussian statistics of these wavelets, matched as Gaussian mixture density (with observed sequence and hidden state). This, consideration of HMM can be labelled as Hidden Markov Tree Model due to the quad-tree nature of wavelets. Wavelet coefficients are linked together with a state variable i.e. every wavelet coefficient was described

by q (a-dimensional state probability) and σ (m-dimensional standard deviation vector).

$$q = (q_1, q_2, \dots, q_m) \quad (1)$$

$$\sigma = (\sigma_1, \sigma_2, \dots, \sigma_m) \quad (2)$$

A multidimensional Gaussian Mixture Model is referred as HMT. Wavelet coefficients are randomly modeled by HMT, with probability density function as a mixture of zero mean Gaussian distribution hidden state for the classification of large and small coefficients. Where pdf of C is:

$$f_c(c) = \sum_{n=1}^N p_Q(n) f_{c|Q}(c|Q=n) \quad (3)$$

Where $p_Q(n)$ is pmf, and Q is a hidden state random variable. Conditional pmf $f_{c|Q}(c|Q=n)$ is given by:

$$f_{c|Q}(c|Q=n) = \frac{2}{\sigma_n \sqrt{2\pi}} \exp\left(-\frac{(b-\mu_n)^2}{2\sigma_n^2}\right) \quad (4)$$

Here μ_n and σ_n are the mean and variance respectively.

HMT used probabilistic tree model to display the Markovian dependencies among hidden states to capture the inter-scale and intra-scale dependencies between wavelet coefficients. For the decomposition of wavelets into u scale and v sub-band, an HMT model has following parameters:

- Standard Deviation = $\sigma_{u,v}$
- Pmf for the root node $Q_i = p_{si}(n)$
- State transition probability matrix of v sub-band from scale $u-1$ to scale $u = A_{u,v}$
- Gaussian mean = $\mu_{u,v}$

The Equation (4) shows state transition matrix as parent \rightarrow children state to state links between the hidden states:

$$A_{u,v} = \begin{bmatrix} p_{u,v}^{y \rightarrow y} & p_{u,v}^{y \rightarrow z} \\ p_{u,v}^{z \rightarrow y} & p_{u,v}^{z \rightarrow z} \end{bmatrix} \quad (5)$$

where or probability of wavelet coefficients to be large or small given its parent is large or small. All these parameters are grouped θ .

$$\theta = [p(\theta_i=n), A_{u,v}, \mu_{u,v}, \sigma_{u,v}] \quad (6)$$

Every wavelet coefficient here, has different state transition probability and variances which

leads toward greater complexity in HMT model. It can be reduced by tying within scale method [19].

3 Denoising Technique

HMT denoising technique with 2D-DWT and 2D-GMM is used. Expectation-Maximization algorithm iteratively find the maximum likelihood from a data set. Our proposed strategy used the adequacy of DWT and the hierarchical relationships between sub-bands. HMT model was used to locate a parameter set θ_q .

A two state GMM was utilized to start the HMT model. At that point, to get θ_q , the inter-scale dependencies were caught by the Markov-tree and EM-algorithm.

Increase in the signal [20] variance is based on added noise while the other parameters are left unchanged. Noisy observation θ_q was extracted and then noise variance was subtracted from it:

$$\left(\sigma_{(u,v,m)n}^{(q)}\right)^2 = \left(\left(\sigma_{(u,v,m)n}^{(q')}\right)^2 - \left(\theta_{(u,v,m)n}^{(q)}\right)^2\right)_+ \quad (7)$$

where v, u and m represent v sub-band, u scale, n state, mth coefficient and

$$\begin{cases} (g)_+ = g & \text{for } g > 0 \\ (g)_+ = 0, & \text{for } g \leq 0 \end{cases}$$

A Model Training via EM Algorithm

EM (Expectation Maximization) algorithm is used for model training. The EM algorithm given in Equation (8), describes the statistical model for hidden state Q and variable C

$$E_{\theta} \left(Q_T(C, Q | C = c) \right) = E_{\theta} \mathcal{D} (c, Q) \quad (8)$$

Conditional pmf of state Q and its maximization is given as:

$$P(Q = n | C, \theta_q) = \frac{P(Q_i) g(C; O, \sigma_{u,v}^2)}{\sum_{i=0}^1 P(Q = i) g(C; O, \sigma_{u,i}^2)} \quad (9)$$

$$P(Q = n) = \frac{1}{N} \sum_{b \in Z^2} P(Q = n | C, \theta_q) \quad (10)$$

After determining θ_q and state probability through HMM, we got $q = E [q | q', \theta_q]$ (Bayes Estimator) can be used to obtain clean coefficients.

$$q = \sum_n (Q|q, \theta_q) \times \frac{\left(\theta_{(u,v,m)_n}^{(q)}\right)^2}{\left(\theta_{(u,v,m)_n}^{(q)}\right)^2 + \left(\theta_{(u,v,m)_n}^{(\epsilon)}\right)^2} q'_{u,v,m} \quad (11)$$

B Inverse Wavelet Transform

Inverse Wavelet Transform (IDWT) is applied in the end to acquire reconstructed video frames. Algorithm for proposed denoising method is summarized in Figure 2.

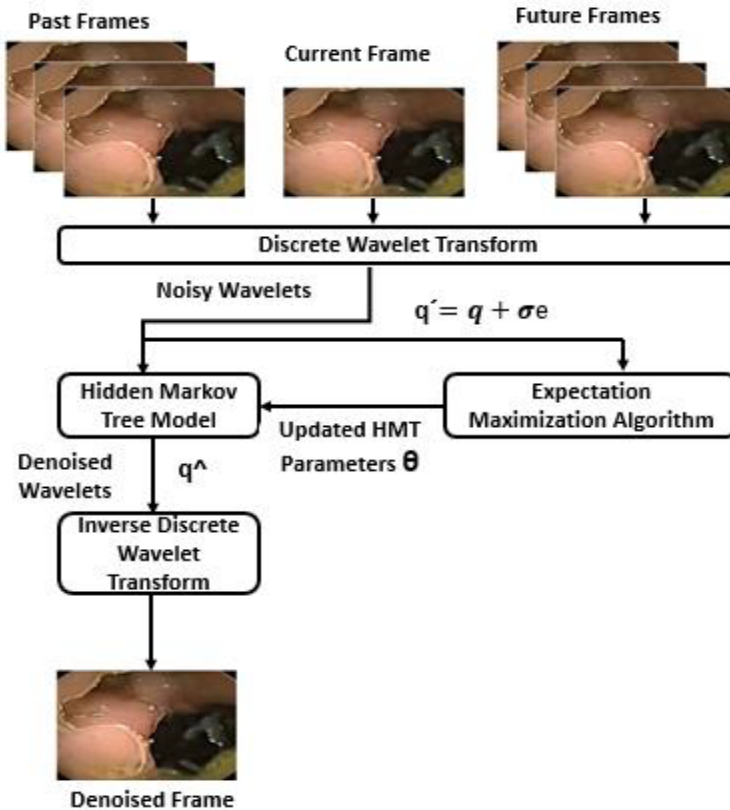


Figure 2: An Overview Proposed Denoising Process

Denoising Algorithm

- Adding noise AWGN in each frame of the video sequence.
- Then applying Daubechies-8 DWT .
- Obtaining DWT coefficients.
- Estimating the GMM parameters.
- Training of HMT model with EM-algorithm in reference to the method of tying within scale.
- Applying IDWT to get reconstructed frames.

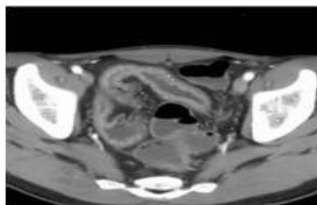
4 Simulations and Results

Each frame of the tested sequence was degraded artificially with speckle noise with noise variances i.e. 0.1, 0.2, 0.3. The sequences were tested in grayscale, RGB and proposed algorithm color space. The analysis of the measurement shown in Table 1 was made on two terms i.e. first on the quantitative performance measure PSNR, secondly on the measurement of perceived image quality with initial noise free image as reference through Structural Similarity Index Measure (SSIM).

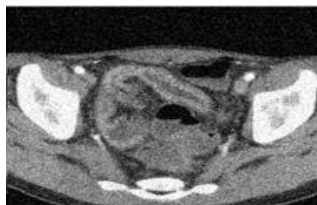
Table 1: Simulation Results

Video Sequences		Grey Scale		RGB		Proposed Algorithm	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Endoscopy	0.1	21.1167	0.881694	21.1131	0.904999	21.5982	0.884287
	0.2	15.1980	0.822676	15.2762	0.894293	15.5062	0.828196
	0.3	11.6759	0.855640	11.7776	0.909898	11.9761	0.761609
CT Scan	0.1	21.3809	0.866745	21.3847	0.866680	21.5603	0.852319
	0.2	13.9769	0.846671	13.9789	0.867870	15.4741	0.854179
	0.3	11.5437	0.826184	11.4503	0.859014	11.9199	0.837778
Mammogram	0.1	20.0024	0.836505	21.3705	0.867967	21.6039	0.856158
	0.2	13.9792	0.830533	15.1374	0.876485	15.5392	0.846681
	0.3	10.4566	0.831411	11.5517	0.853997	11.9808	0.813299
Ultrasound	0.1	21.1587	0.845638	21.2203	0.905396	21.4992	0.883864
	0.2	13.9793	0.877209	15.3598	0.885453	15.5102	0.864535
	0.3	11.6582	0.848289	11.8212	0.903502	11.9896	0.830726

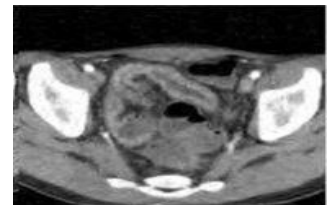
Figures 3(a, d, g) show the noiseless frames of the test sequences, which were later corrupted with speckle noise (Figs. 3(b, e, h)). These frames were then processed through denoising filters to obtain denoised frames (Figs. 3(c, f, i)). Figure 3(j) shows the graphical comparison of the test sequences.



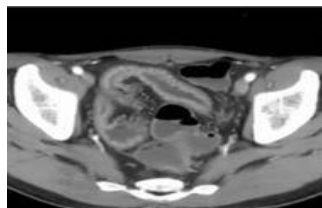
(a) Gray Frame



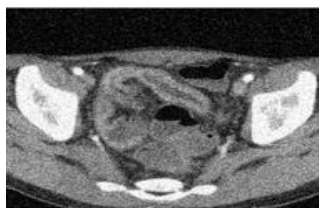
(b) Noisy Frame



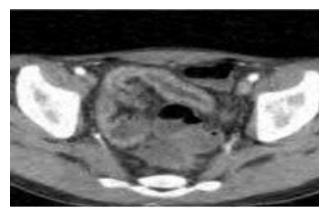
(c) Denoised Frame



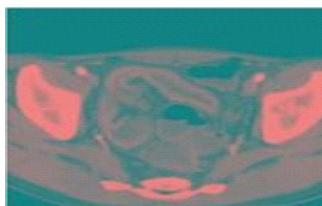
(d) RGB Frame



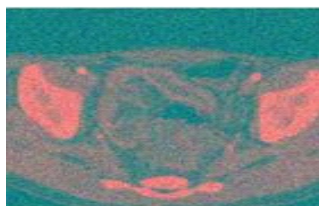
(e) Noisy Frame



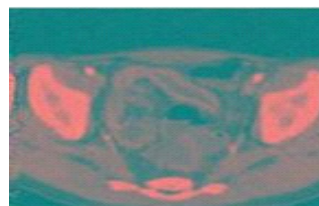
(f) Denoised Frame



(g) YCbCr Frame



(h) Noisy Frame

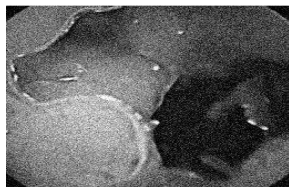


(i) Denoised Frame

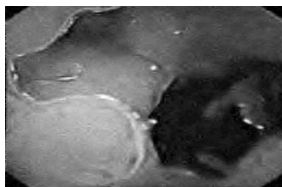


(j) Resulting Recovered Frame

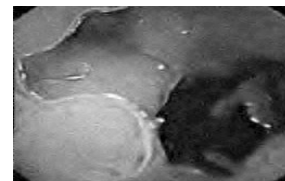
Figures 3: Qualitative Comparison Based on PSNR of CT Scan with Gray Scale (Upper One), RGB Scale (Middle One) and Proposed Algorithm (Last One) View Proposed Denoising Process



(a) Gray Frame



(b) Noisy Frame



(c) Denoised Frame



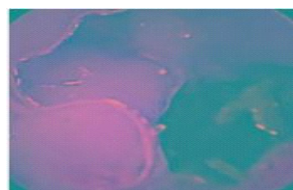
(d) RGB Frame



(e) Noisy Frame



(f) Denoised Frame



(g) YCbCr Frame



(h) Noisy Frame

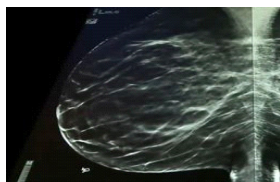


(i) Denoised Frame



(j) Resulting Recovered Frame

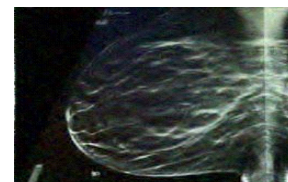
Figures 4: Qualitative Comparison Based on PSNR of Endoscopy with Gray Scale (Upper One), RGB Scale (Middle One) and Proposed Algorithm (Last One) View Proposed Denoising Process



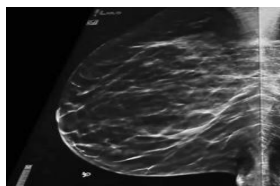
(a) Gray Frame



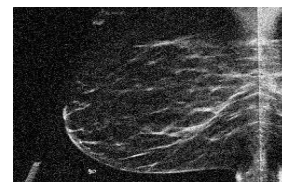
(b) Noisy Frame



(c) Denoised Frame



(d) RGB Frame



(e) Noisy Frame



(f) Denoised Frame



(j) Resulting Recovered Frame

Figure 5: Qualitative Comparison Based on PSNR of Mammogram with Gray Scale (Upper One), RGB Scale (Middle One) and Proposed Algorithm (Last One) View Proposed Denoising Process

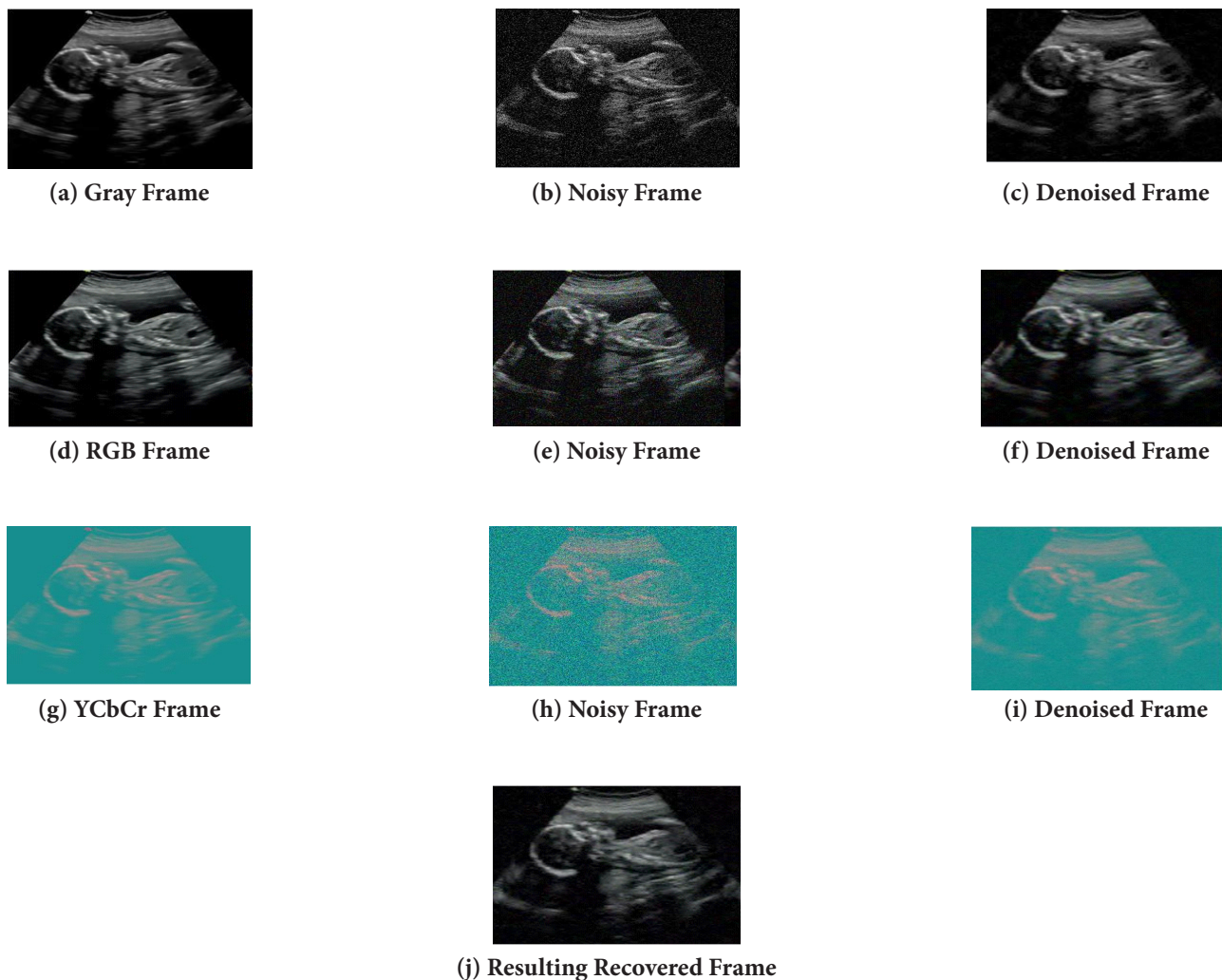
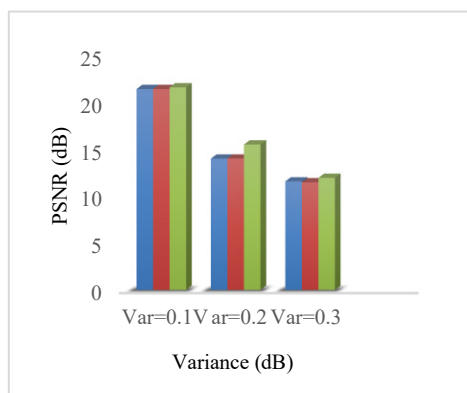
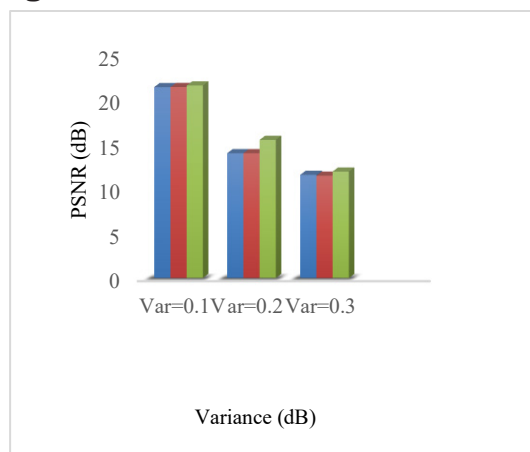


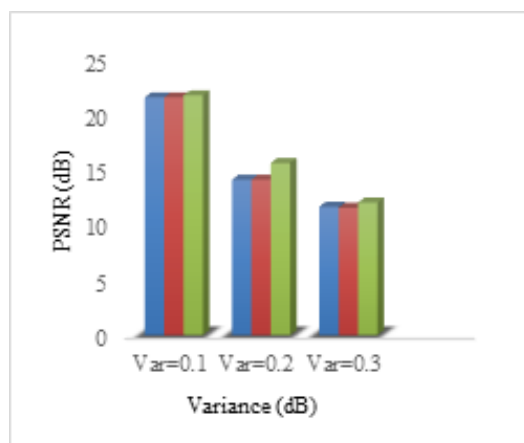
Figure 6: Qualitative Comparison Based on PSNR of ULTRASOUND with Gray Scale (Upper One), RGB Scale (Middle One) and Proposed Algorithm (Last One) View Proposed Denoising Process



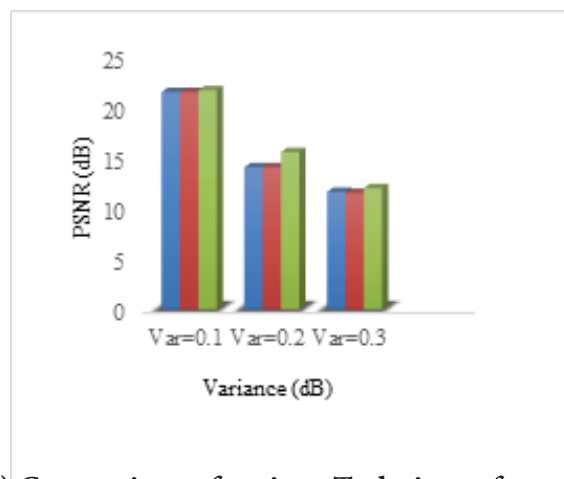
(a) Comparison of various Techniques for CT Scan



(b) Comparison of various Techniques for Endoscopy



(c) Comparison of various Techniques for Mammogram



(d) Comparison of various Techniques for Ultrasound

Figure 7: Graphical Comparison of Different Techniques through PSNR

5 Conclusion

Here we present a wavelet denoising technique based on HMT model with EM algorithm for the despeckling of video frames scale, videos and colored videos in YCbCr color space based on 2D-DWT and 2D-GMM. HMT model was trained by EM algorithm on wavelet coefficients to capture the statistical dependencies present between them. Experimental results revealed that the proposed denoising method performed better in YCbCr space with improved performance up to 0.5dB as compared to gray and RGB space. This method showed improvement in noise reduction, edge preservation and reduced computational complexity. Furthermore, increases in noise variance and using the higher order filter resulted in degraded image quality.

6 Future Work

Proposed denoising scheme can be further extended for the analysis of noises of different variances with other transforms for all color spaces. Moreover, it can be used in other transform domains like Bandelet, Contourlet, Rigdelet and Curvelet in combination with other filters i.e. Bilateral filter. Finally, this technique can be further modified with other techniques for achieving the improved performance and with reduced computational complexity and low latency.

Acknowledgment

This research has been supported by Image Processing Cell, Military College of Signals, National University of Sciences & Technology, Islamabad, Pakistan.

References

- [1] Sharma, A., and Singh, J., December 2013, "Image Denoising Using Spatial Domain Filters: A Quantitative Study", Proceedings of 6th IEEE International Congress on Image and Signal Processing, pp. 293-298, Hangzhou, China.
- [2] Narasimha, C., and Rao, N.A., January 2015, "Spatial Domain Filter for Medical Image Enhancement", Proceedings of IEEE International Conference on Signal Processing and Communication Engineering Systems, pp. 291-295, Guntur, India.
- [3] Wang, B., Xiong, Z., and Zhang, D., October 2014, "Nonlocal Image Denoising via Collaborative Spatial-domain LMMSE Estimation", IEEE International Conference on Image Processing, pp. 2714-2718, Paris, France.
- [4] Lee, J.S., March 1980, "Digital Image Enhancement and Noise Filtering by Use of Local Statistics", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 2, pp. 165-168.
- [5] Li, X., Shen, H., and Zhang, L., August 2015, "Sparse-Based Reconstruction of Missing Information in Remote Sensing Images from Spectral/Temporal Complementary Information", International Society for Photogrammetry and Remote Sensing, Volume 106, pp. 1-15.
- [6] Manjón, J.V., Coupé, P., Buades, A., Collins, D.L., and Robles, M., January 2012, "New Methods for MRI Denoising Based on Sparseness and Self-Similarity", Med Image Anal, Volume 16, pp. 18-27.
- [7] Iftikhar, M.A., Jalil, A., Rathore, S., and Hussain, M., March 2014, "Robust Brain MRI Denoising and Segmentation Using Enhanced Non-Local Means Algorithm", International Journal of Imaging System Technology, Volume 24, pp. 52-66, New York, USA.
- [8] Liu, R.W., Shi, L., Huang, W., Xu, J., Yu, C.H., and Wang, D., July 2014, "Generalized Total Variation-Vased MRI RicianDenoising Model with Spatially Adaptive Regularization Parameters", Magnetic Resonance Imaging, Volume 32, pp. 702-720.
- [9] Khare, A., and Tiwary, U.S., October 2007, "Daubechies Complex Wavelet Transform Based Technique for Denoising of Medical images", International Journal of Image and Graphics, Volume 7, pp. 663-687.
- [10] Loizou, C.P., and Pattichis, C.S., January 2008, "Despeckle Filtering Algorithms and Software for Ultrasound Imaging", Morgan & Claypool Publishers, San Rafael, CA, USA Algorithms and Software in Engineering, pp. 1-66.

- [11] Varun, P.G., and Palanisamy, P., January 2012, "Capsule Endoscopic Image Denoising Based on Double Density Dual Tree Complex Wavelet Transform", Article in International Journal of Imaging and Robotics.
- [12] Loizou, C.P, Kasparis, T, Christodoulides, P, Theofanous, C., Pantziaris, M., Kyriakou, E., and Pattichis, C.S., November 2012, "Despeckle Filtering in Ultrasound Video of the Common Carotid Artery", Proceedings of IEEE 12th International Conference on Bioinformatics & Bioengineering, Larnaca, Cyprus, pp. 721-726.
- [13] Rabbani, H., and Gazor, S., December 2012, "Video Denoising in Three-Dimensional Complex Wavelet Domain Using a Doubly Stochastic Modelling", IET Image Processing, Volume 6, pp. 1262-1274.
- [14] Kumar, M., and Diwakar, M., January 2016, "CT Image Denoising Using Locally Adaptive Shrinkage Rule in Tetrolet Domain", Babasaheb Bhimrao Ambedkar University, Lucknow, India.
- [15] Agarwal, S., Singh, O.P, and Nagaria, D., May 2017, "Analysis and Comparison of Wavelet Transforms for Denoising MRI Image", Biomedical & Pharmacology Journal, Volume 10, pp. 831-836.
- [16] Maqsood, A., Tauqir, I., Siddique, A.M., and Haider, M., January 2019, "Wavelet Based Video Denoising using Mehran University Probabilistic Models", Research Journal of Engineering & Technology, 38(1), pp. 17-30, Jamshoro, Pakistan.
- [17] Claus, M., and Gemert, J.V., April, 2019, "ViDeNN: Deep Blind Video Denoising", CVPR Workshops on Computer Science.
- [18] Dass, R., June 2018, "Reduction of Ultrasound Images Using BFO Cascaded with Wiener Filter and Discrete Wavelet Transform in Homomorphic Region", International Conference on Computational Intelligence and Data Science, pp. 1-1866.
- [19] Malfait, M., and Roose, D., April 1997, "Wavelet-Based Image Denoising Using a Markov Random Field a Prior Model", IEEE Transactions on Image Processing, Volume 6, pp. 549-565.
- [20] Crouse, M.S., Nowak, R.D., and Baraniuk, R.G., April 1998, "Wavelet-Based Statistical Signal Processing using Hidden Markov Models", IEEE Transactions on Signal Processing, Volume 46, pp. 886-902.
- [21] Golshan, H.M., Hasanzadeh, R.P., and Yousefzadeh, S.C., September 2013, "An MRI Denoising Method Using Image Data Redundancy and Local SNR Estimation", Magnetic Resonance Imaging, Volume 31, pp. 1206-1217.

Keystroke dynamics Based Technique to Enhance the Security in Smart Devices

Farman Pirzado¹ Shahzad Memon² Lachman Das Dhomeja³ Awais Ahmed⁴

Abstract—Nowadays, smart devices have become a part of our lives, hold our data, and are used for sensitive transactions like internet banking, mobile banking, etc. Therefore, it is crucial to secure the data in these smart devices from theft or misplacement. The majority of the devices are secured with password/PINbased user authentication methods, which are already proved a less secure or easily guessable user authentication method. An alternative technique for securing smart devices is keystroke dynamics. Keystroke dynamics (KSD) is behavioral biometrics, which uses a natural typing pattern unique in every individual and difficult to fake or replicates that pattern. This paper proposes a user authentication model based on KSD as an additional security method for increasing the smart devices' security level. In order to analyze the proposed model, an android-based application has been implemented for collecting data from fake and genuine users. Six machine learning algorithms have been tested on the collected data set to study their suitability for use in the keystroke dynamics-based authentication model. **Index Terms**—Keystroke dynamics; Smart Devices; user authentication.

I INTRODUCTION

Now a day's mobile phones are one of the most important things for people. Mobile phones are not only used for calling or sending text messages. They are also used in many confidential transactions such as (E-Banking, E-Commerce, and social networking). Therefore, it is becoming more important to secure mobile phones. As far as the security of smart devices is concerned nowadays, the majority of devices use integrally weak authentication mechanisms based on usually passwords and PINs. This method is based on the user name and password secrecy. The password usually consists of common words and phrases associated with a particular user. That is universally considered weak because it can be easily hacked by different password hacking methods like guessing, phishing, etc. [1] Another user authentication is a pattern lock, which is a graphical password. It includes 3 x 3 grids of small dots. On that small grid, the user is required to draw a graphical pattern with his finger for authentication, and it can also be easily broken by a computer attack named smudge attack. An increase in the deficits of password-based authentication still the majority of smart devices use weak authentication mechanisms based on PIN or password, which do not ensure an appropriate security level for access to the stored information and to the available services. It is important to implement a strong authentication system for smart devices. As compared to physiological biometric systems, behavioral biometrics systems are considered more secure and unique as it is not possible to copy the behavior of the user to the system. The keystroke dynamics is one of the types of behavioral biometric; it monitors the behavior of a user by analyzing the user behavior based

¹Mohammad Ali Jinnah University Karachi, Pakistan | farman.ali@jinnah.edu

²University of Sindh Jamshoro, Pakistan | Shahzad.memon@usindh.edu.pk

³University of Sindh Jamshoro, Pakistan | lachman@usindh.edu.pk

⁴Mohammad Ali Jinnah University Karachi, Pakistan | awais.ahmed@jinnah.edu

on the typing patterns of the user. Keystroke dynamics can be used to improve the security level of pin-based user authentication in smart devices. Keystroke dynamics is the behavioral biometric, with dynamic keystroke user can be easily authenticated via his key typing unique feature such as key press time, the difference between two keys pressed and over all typing speed of the user. With the help of keystroke dynamics, we can strengthen the security of smart devices; even if a hacker hacks your password, he needs to know how to type that password [2]. However, KSD undergoes many implementation challenges such as low accuracy, low permanence, user situation, and emotion because of which, as per our literature survey, research on KSD is very challenging and is in its infancy.

2 BACKGROUND AND RELATED WORK

In the literature, Significant research efforts have been conducted over the years in an attempt to improve the quality of keystroke dynamics as an additional authentication method for smart devices.

Nowadays, smart devices are an important part of our daily life; they are not only used for general communication but also for private communication and important financial transactions. This makes security issue one of the important concerns in smart devices. There exist a number of user authentication methods, including Password/PIN, pattern lock, etc. All of these suffer from various drawbacks and limitations, as discussed before [3].

A biometric identification trait is unique for every single user, and so biometrics can provide secure means for user authentication in smart devices. The term biometric comes from two Greek words 'bio' and 'metrics'; the former means 'life,' and the latter means "measurement." Therefore, biometric authentication has to do with measuring and analyzing the biological characteristics of an individual. Biometrics is classified as being physiological biometrics and behavioral biometrics. The former involves physical characteristics of the user (Thumb, Iris, Retina, etc.) and the latter behavioral characteristic of the user (Keystroke dynamics, signature, etc.) [4], as shown in the figure 1 below

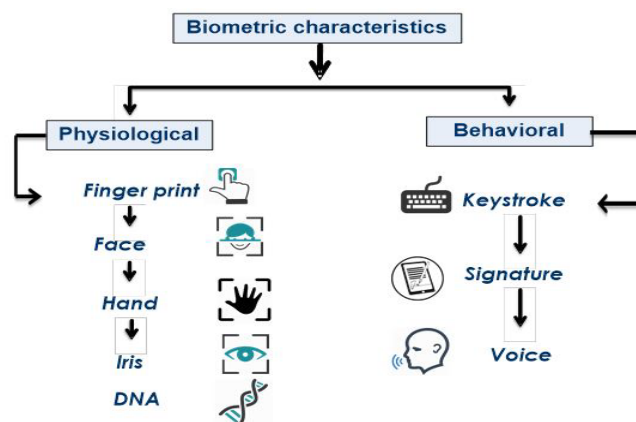


Figure 1: Categories of Biometrics

However, KSD undergoes many implementation challenges such as low accuracy, low permanence, user situation, and emotions, which are discussed below [5]

- ***Low accuracy***

One of the significant issues faced by keystroke applications in its implementation is low accuracy. This issue is fundamental stroke dynamics (KSD) authentication is caused by the large variation in typing style caused by many external factors such as injury, fatigue, or any other distraction. KSD undergoes many issues due to these reasons and challenges, but research has not been stopped; it is being carried out, and one can hope for improved and better results in the future.

- ***Lower permanence***

While researchers have put forward many methods, KSD suffers from lower permanency as compared to all other biometric systems. It is because of a human's typing pattern that may regularly change following the customization towards a password.

- ***User's situation & emotions***

Under such a different user situation like walking, driving, etc. can also cause apprehension in user's major dynamicity in keystroke value. They were thus contributing to the implementation challenges of keystroke dynamics. Such situations and user's happy or angry emotions will play a significant role in the typing behavior of a user. To solve the above problem and challenges in KSD, different models have been introduced in literature based on different methods such as machine learning classifiers, neural networks, and statistical. However, KSD is still in its early stages in mobile devices, and much research needs to be done to make it useful and accurate. In our proposed, model six different classification and regression algorithms of machine learning are applied to the data set of keystroke values, and six different results are generated and compared.

Dr.T.Pandikumar et al.] have proposed the model based on KSD; they have only used a random forest classifier algorithm to achieve the better performance of the proposed system. While in our proposed model, we have applied six different classifiers. To test the performance of these models on the data set we have collected from the users. To evaluate the performance of the system, the authors have used two parameters such as false acceptance rate (FAR) and false rejection rate (FRR), Whereas in our proposed work, four different performance parameters are used to evaluate the performance of the proposed model, including accuracy level and classification error.

In 2020 [6], to improve the authentication accuracy of Keystroke dynamics, researchers assess the feasibility of KSD based biometric verification from a model that is based on sequences of typed characteristics with a Gaussian mixture model (GMM).. At the same time [7] suggested KSD based authentications multi-factored with PIN-based authentication using the Novel feature-scoring method. Another article published in 2020 that focuses on user touch time and force features extracted from the piezoelectric force-touch panel of smart devices; in this works,

researchers used a Support vector machine, Artificial neural network, and Random forest in order to judge error rate in Keystroke dynamics using different classifiers. [8]. Apart from applying different techniques to judge the accuracy of the Keystroke dynamics, there is also a work that focuses on the variation of Keystroke value according to user position; considering it as a research problem, authors presented a three-step authentication model based on three user position, i.e., sitting, walking and relaxing in. [9]. Later [10], Researchers also focus on designing a data mining system by Applying the dimension reduction technique on KSD data and applied two data mining algorithms on data to find out the accuracy, i.e., K - nearest neighbor, Bay classifier, and Decision tree. Currently, a neural network is used as a good problem-solving approach in research; therefore, authors represent a novel analysis for the KSD authentication using timing and no timing feature using neuralnetworks. Researchers also focus on designing a data mining system by Applying the dimension reduction technique on KSD data and applied two data mining algorithms on data to find out the accuracy, i.e., K - nearest neighbor, Bay classifier, and Decision tree. [3]. In addition, Neural networks also help researchers to explore the effectiveness of employing KSD to differentiate between authentic and fake users of mobile using deep learning techniques based on conventional neural networks. [11]. Authors also focus on comprehensive analysis using KSD using Neural network; different neural network classifiers are used in this work in order to find different performance parameters like false acceptance rate, low error rate, and equal error rate [12].

Another research forwarded in [13] by [Dr T.Pandikumar, Abraheem fekde].In their proposed research, they used artificial neural networks for training and classification purposes. Working on the same, they developed an artificial keyboard for collecting the keystroke values from smart devices as well. Whereas it is our proposed research, we have developed an android application to collect the keystroke values from the users. In international journal of soft computing and engineering, M.Karnan and N.Krishnaraaj, 2012 [14], They brought another security model for smart devices. These both the researchers, for user authentication, introduced a hybrid model that is based on the fusion of six biometrics, including keystroke, fingerprint, and palm print. In this model, data from all six biometrics are captured in order to develop a template, which is later used to authenticate genuine users. In our proposed research, we introduced a hybrid model based on the fusion of two user authentication approaches, including password/PIN with keystroke dynamics. On the other hand, as in work, a shorter text has been used by the Authors. In this, only once a password, along with an extended length of text for continuous verification of the user using the vector machine (SVM), was used. In addition, the results of this showed that they could validate authentic users while rejecting imposters with a slight error rate [13]. The user's prediction was calculated by using "SVM." Based on this prediction, a number of sham and genuine users were selected. All 34 users who were selected in the experiment were identified using two-keystroke dynamics features flight and dwell time of 12 most common digraphs. The above researchers applied only one classification algorithm support vector machine, whereas in the proposed research, six different classification algorithms are applied, and the performance of the system is evaluated. Moreover, in our proposed work, for experimental setup we have collected 50 positive samples and 50 negative passwords in order to train classifiers. Apart from the difference in the number of users, they have developed a desktop application to collect keystroke values, while in our proposed work, we developed an android application for data collection. As far as the performance of the proposed model is concerned, a threshold value was set to detect the genuine and fake user, while we extracted

four different performance parameters to evaluate the performance of the proposed model.

3 PROPOSED MODEL

The proposed model show in figure 2 indicates that the user will enter his/her user name and password in a smart device, the sensors are monitored, and keystroke features are collected in the data acquisition phase. After collection of keystroke features such as key holders, time features are normalized, and a dataset is designed for the classification model.

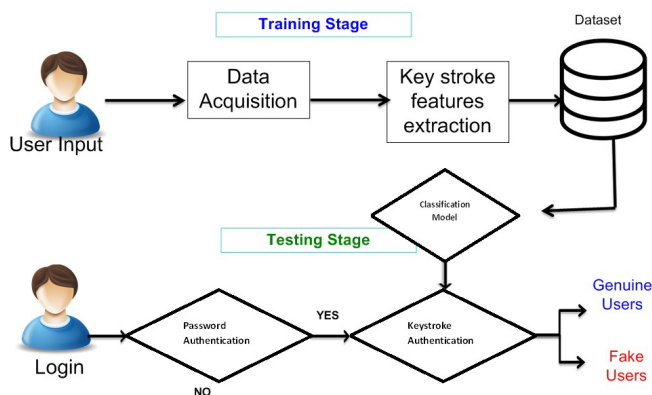


Figure 2: Proposed System

The classification model will accept the dataset as input for authentication, and it will apply a classification algorithm to the given data set, which will provide the two results.

According to the classification algorithm applied in the classification model on a given dataset, if keystroke values don't match with the features stored in the classification model, the user is rejected if the keystroke values match with the data set in the classification model, the user is authenticated.

• **User Registration:** The registration phase asks a user to type his/her user name and password. Simply, in this, a user requires only to log in to the system, then the android application will automatically extract the keystroke features such as (Key hold time (KHD), Key up downtime (KUD), and Total typing speed (TTS) of the user. Dealing with hardware keyboards is very different from dealing with touch screens. There are many features of keystroke dynamic in our proposed system following features have been extracted:

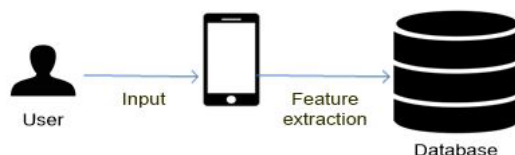


Figure 3: User Registration

- Key hold time (KHT). Time between key down and key up for single character.
- Key Up down Time (KUD). Time between key up of a character and key-down of the next character.
- Total typing speed (TTS). Total time to press a single character.

• **Data acquisition:** For data acquisition, an android application is developed using android development toll Eclipse, which runs as a stand-alone application on an android smart device, asking users to type their password. Here touch sensor of the smart device is used to collect keystroke features of users.

4 PERFORMANCE PARAMETER

The performance of any biometric system, including KSD, is measured in a false acceptance rate (FAR) and false rejection rate. The false acceptance rate (FAR) is defined as “the percentage of invalid inputs which are incorrectly accepted,” while false rejection rate (FAR) is defined “as a percentage of valid inputs which are incorrectly rejected.” [15]. The above algorithms generate different FAR and FRR, which determine the accuracy and performance of the proposed model. Since different values in the operating threshold may result in variation in values of FRR and FAR, the receiver operating system characteristic ROC curve will show the trades off between the FAR and FRR. In addition to FAR and FRR, two other performance parameters, i.e., accuracy level and classification error, are generated by the algorithms being applied.

5 DATA ANALYSIS

Over the last six decades, many classification methods have been applied in keystroke dynamics study. Keystroke dynamics can be perceived as a pattern recognition problem, and the most commonly deployed methods can be broadly categorized as statistical and machine learning approaches [12].

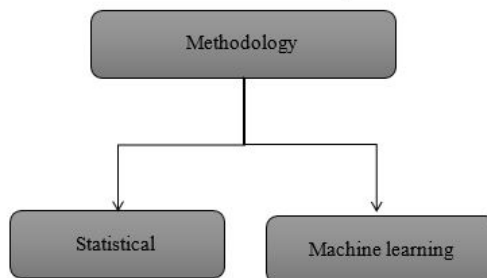


Figure 4: Data Analysis Approaches

Statistical methods are commonly used for data analysis based on different statistical parameters. This approach elaborates in carrying out a study including planning, designing, and analyzing, which results in expressive understanding and reporting of research conclusions. There are many generic statistical measures available such as mean, median, and standard deviation. Machine learning is the branch of artificial intelligence, which is based on the idea that a system can learn from data, can identify patterns, and can make decisions and predictions. MATLAB

also provides immediate access to prebuilt functions, toolboxes, and special applications for classification, regression, and clustering. It is an acronym for matrix laboratory used to solve different mathematical and machine learning problems. It is used as a data analysis tool in research worldwide. For a large amount of data, machine learning is used to find patterns from data and build models. That model can predict future outcomes based on data sets; MATLAB also provides immediate access to prebuilt functions, toolboxes, and special applications for classification, regression, and clustering. Data is acquired using a self-designed android application, which will ask users to type their password. Data has been collected from users 50 positive samples and 50 negative passwords to train classifiers using a self-designed android application. The data set of genuine and fake users has been developed in Ms. Excel according to the format that classification algorithms support for further data analysis in MATLAB.

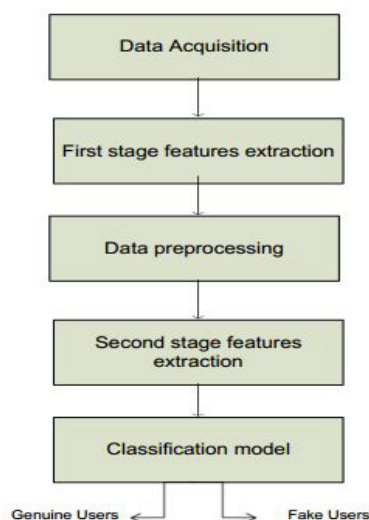


Figure 5: Data Analysis Process

As the range of values from the first stage of feature extraction is different, therefore; normalization will be applied to transform all the values in the range of 0 - 1.

- The first stage features include the numeric value of keystroke features such as key hold time, key up-down time, and full typing speed.
- MATLAB converts these first stage features into second stage features, such as statistical, frequency domain, and non-linear features, to accurately predict results.
- A statistical feature includes the first difference of mean values, Standard deviation, the first difference of standard deviation values, mode, median, root mean square, etc.
- Frequency domain features include spectral density, Fourier coefficients, magnitude, and peak value.
- Non-linear features include point-care geometry features, which will be extracted after transforming the first stage features into given vectors.
- Based on the second stage features, the feature space will be constructed, and data set will be labeled with the ground truth such as the genuine user (1) or fake user (0).
- The system will then be trained using different classification algorithms, such as decision tree, linear discriminant functions, logistic regression, support vector machine, Knearest neighbor, random forest, and ensemble methods.

There are many machine-learning algorithms used to identify and classify patterns and make correct decisions based on data provided. The classification algorithms, which are used in this research, are described below in Table I.

Table 1: Machine-learning algorithms [16] & [17]

S.No.	Classification Algorithm	Description
1	Support Vector Machine (SVM)	A discriminative classifier formally defined by a separating hyper plane.
2	Random Forests	This algorithm creates forests with number of trees. The more the trees in forests the high the accuracy.
3	Decision tree	It creates a training model which is used to predict the values from a given data set.
4	K-nearest	It classifies a new class based on similarity measures. Used for statistical and pattern recognition.
5	Logistic Regression	Used to predict a binary outcome 0/1 from a given set of independent values.
6	Adaptive Boosting	Also known as AdaBoost used for classification or regression. Often sensitive to noisy data outliers.

6 RESULTS

In this section all the results are discussed and presented in graph. Total number of passwords in data set is 10 and for each password we have collected 50 positive samples and 50 negative passwords in order to train classifiers. As a result the data set contains 1000 of 500 of positive and 500 of negative.

A Support Vector Machine

The first classification algorithm that is applied to the data set is the support vector machine. As shown in the graph, SVM generated 89% accuracy level when applied on the collected data set, whereas the true positive rate is 86% and the false positive rate is 7%.

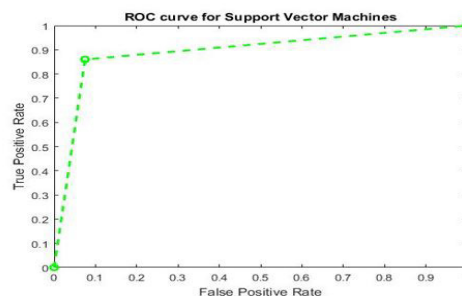


Figure 6: Support Vector Machine

B *Random Forest*

As shown in graph in figure 4.3, Random Forests algorithm generated 97% accuracy level when applied on the collected dataset, whereas the true positive rate is 98% and false positive rate is 3%.

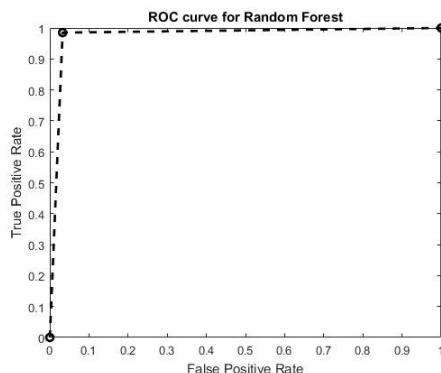


Figure 7: Random Forest

C *Decision tree*

As shown in figure 4.4, decision tree generated 97% accuracy level when applied on collected data set, where as true positive rate is 97% and false positive rate is 3% only.

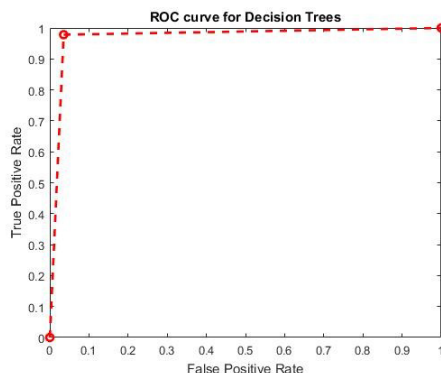


Figure 8: Decision tree

D *Adaptive boosting (Ad boost)*

Adaptive boosting is also known as adaboost: used for both classification and regression. After applying this algorithm on dataset 77% of accuracy level is accomplished with 2% of classification error and 16% of false positive rate and 71% of true positive rate as show in the graph below.

E *k-Nearest neighbor*

This algorithm classifies a new class based on similarity measures is used for statistical and pattern recognition. By applying K- nearest algorithm the accuracy level that is achieved is 93% with classification error of 23%. This algorithm generates the true positive rate of 88% and 1% of false positive rate.

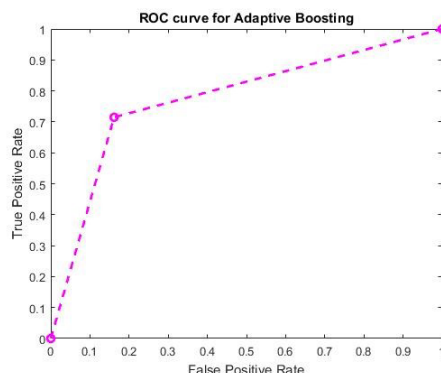


Figure 9: Adaptive boosting (Ad boost)

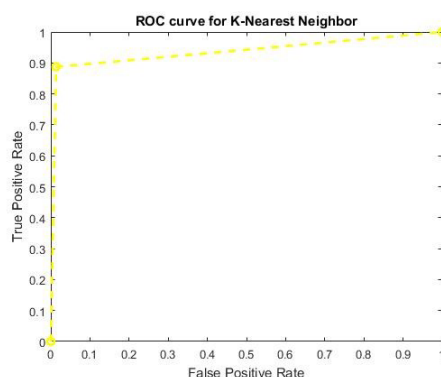


Figure 10: k-Nearest neighbor

F Logistic regression

Results of this algorithm are shown in the graph given below. It generates the accuracy level of 66% with classification error rate of 13%. It gives the 35% of false positive rate and 69% of true positive rate.

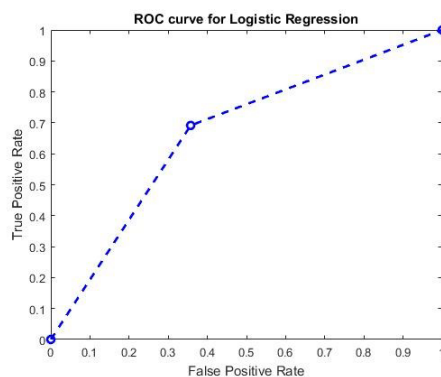
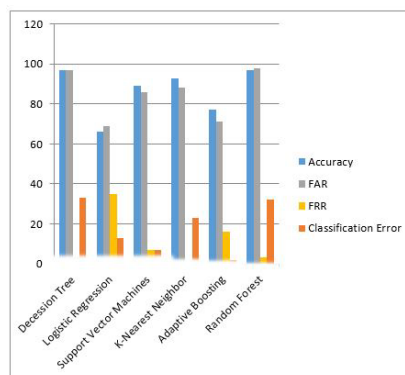


Figure 11: Logistic regression

Six different classification algorithms, when applied on collected data set, produced the six different results as shown in the table II.

The proposed work is explored to design and develop a model based on Keystroke dynamics for user authentication in smart devices. As shown below in table II, the two algorithms Decision Tree and Decision tree provide maximum accuracy level of 97% and SVM provides 80% of accuracy level. Classification error is the proportion of classification a classifier gets wrong therefore if classification rate decrease the false acceptance rate also decreases and performance of a system increases.

**Figure 12: Comparative Analysis S.No Algorithm Accuracy False****Table 2: Results**

S.No	Algorithm	Accuracy	False Positive Rate	True Positive Rate	Classification error
1	DT	97%	3%	97%	33%
2	LR	66%	35%	69%	13%
3	SVM	89%	7%	86%	7%
4	K-NN	93%	1%	88%	23%
5	AB	77%	16%	71%	2%
6	RF	97%	3%	98%	32%

The results suggest that using machine-learning algorithms can play a vital role in order to implement keystroke dynamics with traditional PIN based authentication in smart devices.

7 FINDINGS

In order to find out the performance of the proposed model, different classification machine learning algorithms were applied to the collected data set. The findings are mentioned below:

- Results suggest that the use of machine learning algorithms in keystroke dynamics based authentication models is a promising technique as they provide a good accuracy level as shown in results that two algorithms, Decision Tree and Random Forests, provide a maximum accuracy level of 97% and has a minimum false positive rate of 3%.
- Another advantage of implementing KSD in the smart device is that it is easily implementable since it can be implemented at the software level without requiring any additional hardware and hence a cost-effective method for user authentication.
- KSD can prove itself a good additional user authentication method along with password in smart devices.

8 CONCLUSION

The main objective of this research is to propose keystroke dynamics- based technique to enhance the security in smart de-vices. For achieving our goals, an android application has been developed, which collects keystroke features of authorized and unauthorized users. After collecting data (from genuine and fake users), the data set is developed for further analysis and results. Later on, different classification algorithms of machine learning are applied to data set using the tool of MATLAB. Results show that implementing machine learning algorithms provide acceptable levels in performance measures, as a good factor of authentication. Six different classification algorithms were applied to the same data set using MATLAB, and six different results were generated. The results show that implementing KSD using machine learning algorithms as an additional security method, along with a password, can enhance the security of smart devices. Since the proposed model uses keystrokes dynamics along with the password and that keystroke values are unique for individual users, it provides a strong security model. In case of the password being hacked, the hacker still has to know the typing behavior. Moreover, unlike other authentication methods (fingerprints, face recognition, etc.) that require dedicated hardware for implementation, keystroke dynamics are easily implementable since they can be implemented at the software level. Therefore, it has proven itself capable of providing additional security with other authentication methods such as user name and password. In this research, we have proposed and designed a user authentication model for smart devices based on keystroke dynamics and username and password.

9 FUTURE WORK

As the research in the field of keystroke dynamics suggests that it is a promising technique to be used as an additional security method with other authentication mechanisms such as user name, etc., it is still in its infancy and research phase. The proposed keystroke-based user authentication model can further be investigated with respect to the following directions:

- Currently, four keystroke features (i.e., key hold time, key up downtime, key down-down time, and total typing speed) have been used for user authentication in the proposed model. However, it may be interesting to investigate the accuracy of the proposed model by exploiting more keystroke features such as finger pressure, latency, etc.

- Neural network-based techniques and other data analysis techniques may be applied for further investigation of the proposed model.
- The proposed model may be tested under other operating systems such as AppleIOS, Blackberry, and Symbian, to see how the proposed model performs on these platforms.

REFERENCES

- [1] M Karnan and N Krishnaraj. A model to secure mobile devices using keystroke dynamics through soft computing techniques. *International Journal of Soft Computing and Engineering (IJSCE) ISSN*, pages 2231–2307, 2012.
- [2] Mohd Anwar and Ashiq Imran. A comparative study of graphical and alphanumeric passwords for mobile device authentication. In *MAICS*, pages 13–18, 2015.
- [3] Asma Salem, Ahmad Sharieh, Azzam Sleit, and Riad Jabri. Enhanced authentication system performance based on keystroke dynamics using classification algorithms. *KSII Transactions on Internet & Information Systems*, 13(8), 2019.
- [4] Anil K Jain, Karthik Nandakumar, and Abhishek Nagar. Biometric template security. *EURASIP Journal on advances in signal processing*, 2008:1–17, 2008.
- [5] Yu Zhong and Yunbin Deng. A survey on keystroke dynamics biometrics: approaches, advances, and evaluations. *Recent Advances in User Authentication Using Keystroke Dynamics Biometrics*, pages 1–22, 2015.
- [6] Himanka Kalita, Emanuele Maiorana, and Patrizio Campisi. Keystroke dynamics for biometric recognition in handheld devices. In *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*, pages 410–416. IEEE, 2020.
- [7] Dong In Kim, Shincheol Lee, and Ji Sun Shin. A new feature scoring method in keystroke dynamics-based user authentications. *IEEE Access*, 8:27901–27914, 2020.
- [8] Anbiao Huang, Shuo Gao, Junliang Chen, Lijun Xu, and Arokia Nathan. High security user authentication enabled by piezoelectric keystroke dynamics and machine learning. *IEEE Sensors Journal*, 2020.
- [9] Baljit Singh Saini, Parminder Singh, Anand Nayyar, Navdeep Kaur, Kamaljit Singh Bhatia, Shaker El-Sappagh, and Jong-Wan Hu. A three-step authentication model for mobile phone user using keystroke dynamics. *IEEE Access*, 8:125909–125922, 2020.

- [10] Shri Kant, Alok Katiyar, and Shubhii Shuklla. Smart mobile device authentication using keystroke dynamics based behavior classification. [11] Emanuele Maiorana, Himanka Kalita, and Patrizio Campisi. Deepkey: Keystroke dynamics and cnn for biometric recognition on mobile devices. In 2019 8th European Workshop on Visual Information Processing (EUVIP), pages 181–186. IEEE, 2019.
- [12] Asma Salem and Mohammad S Obaidat. A novel security scheme for behavioral authentication systems based on keystroke dynamics. *Security and Privacy*, 2(2):e64, 2019.
- [13] Hayreddin C, eker and Shambhu Upadhyaya. User authentication with keystroke dynamics in long-text data. In 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), pages 1–6. IEEE, 2016.
- [14] Ricardo N Rodrigues, Glauco FG Yared, Carlos R do N Costa, Joao BT ~ Yabu-Uti, Fabio Violaro, and Lee Luan Ling. Biometric access control through numerical keyboards based on keystroke dynamics. In *International Conference on Biometrics*, pages 640–646. Springer, 2006.
- [15] Naveed Riaz, Ayesha Riaz, and Sajid Ali Khan. Biometric template security: an overview. *Sensor Review*, 2018.
- [16] Taiwo Oladipupo Ayodele. Types of machine learning algorithms. *New advances in machine learning*, 3:19–48, 2010.
- [17] Ayon Dey. Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies*, 7(3):1174–1179, 2016.
- [18] T Pandikumar, Abraham Fekede, and Capt Zinabu Haile. Enhancing performance and usability of keystroke dynamics authentication on mobile touchscreen devices using features extraction scheme. *International Journal of Engineering Science*, 13415, 2017.
- [19] Amund Tveit, Magnus Lie Hetland, and Haavard Engum. Incremental and decremental proximal support vector classification using decay coefficients. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 422–429. Springer, 2003

Call for Papers/Authors Guideline

KIET Journal of Computing and Information Sciences (KJCIS) is the bi-annual, multi-disciplinary research journal published by the College of Computing & Information Sciences (CoCIS) at Karachi Institute of Economics and Technology (KIET), Karachi, Pakistan. KJCIS is a HEC recognized “Y” category journal which is published in January and July every year.

KJCIS Academic Editorial Board and Advisory Board consists of prominent and scholarly academicians across the globe. Our reviewers committee also consist of senior academicians who are responsible to ensure the quality of papers published in KJCIS.

KJCIS aims to provide a panoramic view of the state of the art development in the field of computing and information sciences at a global level. It provides a premier interdisciplinary platform to researchers, scientists and practitioners from the field of computing and information sciences to share their findings and contribute to the knowledge domain at a global level. The journal also fills the gap between an academician and industrial research communities.

KJCIS is a multi-disciplinary journal covering viewpoints/ researches / opinions relevant to the non exhaustive list of the topics including data mining, big data, machine learning, artificial intelligence, mobile applications, computer networks, cryptography & information security, mobile and wireless communication, adhoc & body area networks, software engineering, speech & pattern recognition, evolutionary computation, semantic web & its application, data base technologies & its applications, Internet of Things (IoT), computer vision, distributed computing, grid and cloud computing.

The authors may submit manuscripts abiding to following rules:-

- Certify that the paper is original and is not under consideration for publication in any other journal. Please mention so, in case it has been submitted elsewhere.
- Adhere to normal rules of business or research writing. Font style be 12 points and the length of the paper can vary between 3000 to 5000 words.
- Illustrations/tables or figures should be numbered consecutively in Arabic numerals and should be inserted appropriately within the text.
- The title page of the manuscript should contain the Title, the Name(s), email address and institutional affiliation, an abstract of not more than 200 words should be included. A footnote on the same sheet should give a short profile of the author(s).
- Full reference and /or websites link, should be given in accordance with the APA citation style. These will be listed as separate section at the end of the paper in bibliographic style. References should be justified.
- All manuscripts would be subjected to tests of plagiarism before being peer reviewed.
- All manuscripts go through double blind peer review process.

- Please submit the manuscript through the KJCIS website (www.kjcis.pafkiet.edu.pk) by registering and logging into the system to upload the manuscript.
- Please do not email the manuscripts; however for all other queries, please email us at kjcis@pafkiet.edu.pk.
- The Journal does not have any article processing and publication charges.

Submission is voluntary and all contributors will find a respectable acknowledgment on their opinion and effort from our team of editors. Submission of a paper will be held to imply that it contains original unpublished work. In case the paper has been forwarded for publication elsewhere, kindly apprise in time if the paper has been accepted elsewhere. Manuscripts may be submitted before September and May to get published in Jan & July issues respectively.



Karachi Institute of Economics and Technology

Korangi Creek, Karachi-75190, Pakistan

Tel: (9221) 3509114-7, 34532182, 34543280 Fax: (92221) 35009118

Email: kjcis@pafkiet.edu.pk

<http://kjcis.pafkiet.edu.pk>