

Anomalous Behaviour Detection using Heterogeneous Data



Azliza Mohd Ali

This dissertation is submitted for the degree of Doctor of

Philosophy

March 2018

School of Computing and Communications

*Dedicated to my husband, Hasrin and my sons: Zaim, Wafi and Aqil, whose always gives
unconditional loves and support;
to my parents, without whom I would not be where I am today optional*

It always seems impossible, until it is done - Nelson Mandela

Declaration

I certify that work presented in this thesis has not been submitted in support of an application for another degree at this or any other university. It is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated. Many of the ideas in this thesis were the product of discussion with my supervisor Prof. Plamen Angelov. Wherever contributions of others are involved, I have taken extra care to indicate this clearly, with due reference to the literature.

Azliza Mohd Ali

April 2018

List of Publications

1. Mohd Ali, A., & Angelov, P. Computational Intelligence Techniques for Biometrics and Forensics: A Survey. In UKCI 2015
2. Gu, X., Angelov, P., Mohd Ali, A., & Gruver, W. A. (2014). Online Evolving Fuzzy Rule-Based Prediction Model for High-Frequency Trading Financial Data Stream. In *Evolving and Adaptive Intelligent Systems (EAIS)*, 2016 (pp. 1–9). IEEE. <http://doi.org/10.1109/EAIS.2016.7502509>
3. Mohd Ali, A., Angelov, P., & Gu, X. (2016). Detecting Anomalous Behaviour Using Heterogeneous Data. In *Advances in Computational Intelligence Systems: Contributions Presented at the 16th UK Workshop on Computational Intelligence, September 7–9, 2016, Lancaster, UK* (pp. 253–273). Springer. <http://doi.org/10.1007/978-3-319-46562-3>
4. Mohd Ali, A., & Angelov, P. (2017). Applying Computational Intelligence to Community Policing and Forensic Investigations. In *Community Policing - A European Perspective* (pp. 1–16). <http://doi.org/10.1007/978-3-319-53396-4>
5. Wang, X., Mohd Ali, A., & Angelov, P. (2017). Gender and Age Classification of Human Faces for Automatic Detection of Anomalous Human Behaviour. In *International Conference on Cybernetics (CYBCONF 2017)* (pp. 1–6). IEEE. <http://doi.org/10.1109/CYBConf.2017.7985780>
6. Mohd Ali, A. & Angelov, P. (2018) Anomalous Behaviour Detection Based on Heterogeneous Data and Data Fusion. In *Soft Computing: A Fusion of Foundations, Methodologies and Applications*. (pp 1-15). Springer. <https://doi.org/10.1007/s00500-017-2989-5>

Abstract

Anomaly detection is one of the most important methods to process and find abnormal data, as this method can distinguish between normal and abnormal behaviour. Anomaly detection has been applied in many areas such as the medical sector, fraud detection in finance, fault detection in machines, intrusion detection in networks, surveillance systems for security, as well as forensic investigations. Abnormal behaviour can give information or answer questions when an investigator is performing an investigation. Anomaly detection is one way to simplify big data by focusing on data that have been grouped or clustered by the anomaly detection method. Forensic data usually consists of heterogeneous data which have several data forms or types such as qualitative or quantitative, structured or unstructured, and primary or secondary. For example, when a crime takes place, the evidence can be in the form of various types of data. The combination of all the data types can produce rich information insights. Nowadays, data has become ‘big’ because it is generated every second of every day and processing has become time-consuming and tedious. Therefore, in this study, a new method to detect abnormal behaviour is proposed using heterogeneous data and combining the data using data fusion technique. Vast challenge data and image data are applied to demonstrate the heterogeneous data. The first contribution in this study is applying the heterogeneous data to detect an anomaly. The recently introduced anomaly detection technique which is known as Empirical Data Analytics (EDA) is applied to detect the abnormal behaviour based on the data sets. Standardised eccentricity (a newly introduced within EDA measure offering a new simplified form of the well-known Chebyshev Inequality) can be applied to any data distribution. Then, the second contribution is applying image data. The image data is processed using pre-trained deep learning network, and classification is done using a support vector machine (SVM). After that, the last contribution is combining anomaly result from heterogeneous data and image recognition using new data

fusion technique. There are five types of data with three different modalities and different dimensionalities. The data cannot be simply combined and integrated. Therefore, the new data fusion technique first analyses the abnormality in each data type separately and determines the degree of suspicious between 0 and 1 and sums up all the degrees of suspicion data afterwards. This method is not intended to be a fully automatic system that resolves investigations, which would likely be unacceptable in any case. The aim is rather to simplify the role of the humans so that they can focus on a small number of cases to be looked in more detail. The proposed approach does simplify the processing of such huge amounts of data. Later, this method can assist human experts in their investigations and making final decisions.

Acknowledgements

Alhamdulillah and thank you, Allah. All praise to the most Gracious and Merciful Almighty who makes this journey possible, without whom nothing is possible. I would like to thank my supervisor Prof. Plamen Angelov. Thank you for your patience, guidance, encouragement, inspiration and advice throughout my time as your student. I am fortunate to have a supervisor who always cared about my work, and who responded to my questions and queries so promptly. Thank you also to my lab mate, Xiaowei Gu, who is always help me in any problem that I had. I will always remember your kindness and your support during my study. I would like to thank all the members of School of Computing and Communications for assisting throughout my study here.

To my beloved husband, Hasrin Abu Hassan, thank you for your endless love, sacrifice and support especially before I began my study and along my PhD journey. Thank you for believing me that I can do it. My sons, Hassan Zaim, Hassan Wafi and Hassan Aqil, I always love you. To my parents, Mohd Ali and Azaiyah, my mother in law, Mariah, my brothers, Azlizam and Azlimin, and my in-law, Shuhada, thank you for prayer, love, and everything.

To my Pandora Chics, Juliana, Irni, Liyana and Anis thank you for always be a good listener and never-ending support. Thank you also to all my Malaysian friends for being like my family here. Thank you to my sponsor, Ministry of Higher Education and Universiti Teknologi MARA for giving the opportunity to study abroad further and giving financial support. Lastly, to all, who involved in this journey, thank you for your support and prayers. Thank you, Allah.

Table of Contents

Dedication.....	i
Declaration.....	ii
List of publications.....	iii
Abstract.....	iv
Acknowledgement.....	vi
List of Tables.....	x
List of Figures.....	xi
List of Abbreviations.....	xiii
1 INTRODUCTION.....	1
1.1 Problem definition.....	1
1.2 Aims and Objectives	3
1.3 Contributions of the Thesis	3
1.4 Thesis Structure.....	4
2 LITERATURE REVIEW	7
2.1 Digital Forensics	7
2.2 Biometrics	14
2.3 Heterogeneous Nature of the Human Behaviour Data.....	20
2.4 Anomaly Detection	23
2.5 Data Fusion	32
2.6 Summary	43
3 RESEARCH METHODOLOGY.....	45
3.1 Challenges.....	45
3.2 Proposed Approach.....	47
3.3 Summary	51
4 DETECTING ANOMALOUS BEHAVIOUR.....	53
4.1 Introduction.....	53
4.2 Anomaly Detection	54

4.3 Summary	68
5 GENDER AND AGE CLASSIFICATION BASED ON IMAGES FOR DETECTING ANOMALOUS BEHAVIOUR.....	69
5.1 Introduction	69
5.2 Gender and age classifications	70
5.3 Deep learning	70
5.4 Summary	74
6 DATA FUSION OF HETEROGENEOUS DATA FOR DECISION MAKING. 75	
6.1 Introduction	75
6.2 Data Fusion	75
6.3 Data transformation.....	76
6.4 Fusion technique	82
6.5 Summary	84
7 EXPERIMENTAL DATA	85
7.1 Heterogeneous Data	85
7.2 VAST Challenge 2014	86
7.3 Face Image Dataset - GAFace Dataset.....	88
7.5 Summary	95
8 RESULTS AND ANALYSIS.....	96
8.1 Introduction	96
8.2 Anomaly Detection Results.....	96
8.3 Image Data Results	109
8.4 Data Fusion Results.....	114
8.5 Summary	119
9 EVALUATION	120
9.1 Introduction	120
9.2 Comparison of anomaly detection	120
9.3 Summary	126
10 CONCLUSION	128

11 REFERENCES 133

List of Tables

Table 2.1 Categories And Comparison of Biometric Systems (High, Medium and Low are denoted by H, M, and L, respectively [68])	15
Table 6.1: Sample data after calculating the degree of suspicion in terms of the distance	82
Table 7.1: Description of Datasets	87
Table 7.2: Sample of GPS Data.....	87
Table 7.3: Sample of Car Assignment Data	87
Table 7.4: Sample of Credit Card Data	88
Table 7.5: Sample of Loyalty Card Data.....	88
Table 7.6: Descriptions of datasets.....	89
Table 7.7: Samples of images from GAFace dataset	89
Table 7.8: Sample images of each class	89
Table 8.1: Location legend.....	103
Table 8.2: Description on abnormal trajectory.....	108
Table 8.3: Haar-like features and classification SVM rates (accuracy)	112
Table 8.4: Pre-trained net results.....	112
Table 8.5: Confusion matrix of age classification.....	112
Table 8.6: Confusion matrix of gender classification	113
Table 8.8: Results of data fusion	115
Table 8.9: Example of several cases of different gender.....	118
Table 8.10: Weighted Total.....	118
Table 9.1: Labels in the comparison table.....	123
Table 9.2: Comparison of anomalies detected using credit card data	123
Table 9.3: Comparison of anomalies detected using credit card data (without 10000 spending data)	124
Table 9.4: Comparison of anomalies detected using loyalty card data	125
Table 9.5: Comparison of anomalies detected using time, distance and ratio of trajectory angle	126

List of Figures

Figure 2.1: Example of point anomaly	27
Figure 2.2: Durrant -Whyte’s Classification Based on the Relations between the Data Sources [130]	35
Figure 2.3: Darasathy’s Classification [130].....	36
Figure 2.4: JDL Data Fusion [130].....	38
Figure 2.5: Classification based on the type of architecture [130].....	40
Figure 3.1: Research Framework	47
Figure 4.1: Normal Distribution	55
Figure 4.2: Arbitrary Distribution	55
Figure 4.3: Histogram plot of temperature in Manchester from 2010 to 2012	56
Figure 4.4: Histogram plot of temperature in Manchester in 2010	57
Figure 4.5: Histogram plot of temperature in Manchester in 2011	57
Figure 4.6 Histogram plot of temperature in Manchester in 2012	58
Figure 4.7: Plotting the temperatures and days from 2010 to 2012	58
Figure 4.8: Plotting the temperatures and days in 2010	59
Figure 4.9: Plotting the temperatures and days in 2011	59
Figure 4.10: Plotting the temperatures and days in 2012	60
Figure 4.11: Eccentricity of temperature from 2010 to 2012.....	66
Figure 4.12: Eccentricity of temperature in 2010.....	66
Figure 4.13: Eccentricity of temperature in 2011.....	67
Figure 4.14: Eccentricity of temperature in 2012.....	67
Figure 5.1: Difference between machine learning approaches and knowledge transfer approaches [Ling Shoa 2015].....	72
Figure 5.2: AlexNet Structure	73
Figure 5.3: Pre-trained net application structure	74
Figure 6.1: Degree of suspicion – credit card.....	78
Figure 6.2 : Degree of suspicion – Loyalty card	79
Figure 6.3: 50 metres distance from the car park to the shop location.....	80
Figure 6.4: 555 metres distance from the car park to the shop location.....	81
Figure 6.5: 700 metres distance from the car park to the shop location.....	81
Figure 8.1: RDE on credit card data – 6σ	97

Figure 8.2: RDE on loyalty card data – 6σ	97
Figure 8.3: RDE on loyalty card data – 3σ	98
Figure 8.4: RDE on credit card data – 3σ	98
Figure 8.5: Anomaly on credit card usage.....	100
Figure 8.6: Anomaly on money spend, day and staff ID.....	101
Figure 8.7: Anomalies based on the credit card transaction data after removing the first anomaly	101
Figure 8.8: Anomalies based on the money spend, day and staff ID after removing the first anomaly	102
Figure 8.9: Anomalies by the loyalty card data.....	102
Figure 8.10: Anomalies by the loyalty card data, day and staff ID.....	103
Figure 8.11: Comparison of the total spending using credit card and loyalty card in different locations.....	103
Figure 8.12: Total spending per person using credit card data.....	104
Figure 8.13: Total spending per person using loyalty card data.....	104
Figure 8.14: Total spending per day using credit card data- Staff Member no.31	104
Figure 8.15: Total spending per day using loyalty card data – Staff Member no.31	105
Figure 8.16: Total spending using credit card data per day at Frydos Autosupply.....	105
Figure 8.17: Total spending using loyalty card data per day at Frydos Autosupply.....	105
Figure 8.18: Eccentricity of travel time, distance and trajectory angle ratio and normal and abnormal behaviour.....	107
Figure 8.19: Example of abnormal and normal trajectory	107
Figure 8.20: Anomalies detected based on the travel time using 5σ	108
Figure 8.21: Anomalies detected based on distance using 5σ	108
Figure 8.22: Anomalies detected based on the ratio of the trajectory angle using 5σ	108
Figure 8.23: Comparison of the trajectory for staff member no.31 (left) and staff member no.41 (right).....	109
Figure 8.24: Example rectangle features shown relative to the enclosing detection window. Figure (A) and (B) show two-rectangle features. The three-rectangle feature shown in figure (C), and (D) shown a four-rectangle feature.....	111
Figure 8.25: Details on data no.1.....	117
Figure 8.26: Details on data no.2.....	117
Figure 8.27: Details on data no.3.....	117

Figure 8.28: Comparison between original data and fused data 119

Figure 9.1: Distribution of UK income from 2012 to 2013 [156]..... 122

List of Abbreviations

CSI	:	Crime Scene Investigation
DNA	:	Deoxyribonucleic Acid
GPS	:	Global Positioning System
EDA	:	Empirical Data Analytics
DFRWS	:	Digital Forensics Research Workshop
ANN	:	Artificial Neural Network
SVM	:	Support Vector Machine
k-NN	:	k-Nearest Neighbour
GAs	:	Genetic Algorithms
FL	:	Fuzzy Logic
ML	:	Machine Learning
ALS	:	Autonomous Learning Systems
RDE	:	Recursive Density Estimation
CNN	:	Convolutional Neural Network
PCA	:	Principle Component Analysis
CCTV	:	Closed-circuit Television
ECG	:	Electrocardiogram
NLP	:	Natural Language Processing
IDS	:	Intrusion Detection Systems
TPM	:	Topology Preserving Mapping
JDL	:	Joint Directors of Laboratory
KF	:	Kalman Filter
HCI	:	Human-Computer Interaction
DoD	:	Department of Defense
PDA	:	Probabilistic Data Association
PIR	:	Passive Infrared
MRI	:	Magnetic Resonance Imaging
fMRI	:	functional Magnetic Resonance Imaging
sMRI	:	Structural Magnetic Resonance Imaging
EEG	:	Electroencephalography
jICA	:	joint Independent Component Analysis

tIVA	:	transposed Independent Vector Analysis
DCxWT	:	Daubechies Complex Wavelet Transform
DTCWT	:	Dual-Tree Complex Wavelet Transform
LWT	:	Lifting Wavelet Transform
MWT	:	Multiwavelet Transform
SWT	:	Stationary Wavelet Transform
CT	:	Contourlet Transform
NSCT	:	Nonsampled Contourlet Transform
WSNs	:	Wireless Sensor Networks
QoS	:	Quality of Service
BA	:	Blowfish Algorithm
GM-KRLS	:	Grey Model – Kernel Recursive Least Squares
ReLU	:	Rectified Linear Unit
NLP	:	Natural Language Processing
VAST	:	Visual Analytics Science and Technology
ROC	:	Receiving Operating Characteristics
PR	:	Precision-Recall

Introduction

1.1 Problem definition

Forensics is a branch of science used to solve questions related to crimes. It has been investigated since the nineteenth century. Forensic processes include forensic investigation, evaluation, and intelligence, automated surveillance, and forensic identity management [1]. Forensic science became especially popular among the public after the success of the Crime Scene Investigation (CSI) television show [2]. Even though there are many inaccurate processes in that show, it has still given the public an idea of how the investigation process is being done [2]. Forensic investigation is a detailed, time-consuming and laborious process incorporating both scientific knowledge and technical methods to solve crime cases. Digital forensics is a combination of computer science and law. Knowledge of both is vital. Forensic investigations were previously conducted using anthropological measurements based on photographic documentation [3]. Biometric features such as fingerprint and deoxyribonucleic acid (DNA) identification are more accurate and reliable [2]. After the forensic process, evidence collected from a crime scene must be reliable to be presented to the court. The advancement of technologies in digital forensics has helped to solve many cases in the justice system.

Human behaviour can be identified from everyday routine. Every day, waking up, taking a shower, having breakfast, then going to work is an example of a normal routine which we can consider normal human behaviour. This routine human behaviour can be tracked digitally; for example, if we have surveillance cameras at home, all activity there can be recorded. When going to work by car, if a Global Positioning System (GPS) is installed, then trajectory data can be traced. If we buy a coffee at the café or groceries and pay using a debit or credit card, data about financial and transaction can be retrieved. Data also can be created from mobile applications such as WhatsApp, Telegram, or email. All information about communication in the applications can be processed to produce a pattern of communication. This is a simple example of data on human behaviour that can be easily retrieved. People nowadays carry electronic devices everywhere they go. They may connect to the Wi-Fi to access the internet. Details of each person can be easily retrieved in real-time from interconnected devices. All of this data contains information about human behaviour.

Anomaly detection is one of the most important methods to find and process abnormal data [4], as this method can distinguish between normal and abnormal behaviour. Anomaly detection has been applied in many areas, such as the medical sector [5], fraud detection in finance [6], fault detection in machines [7], intrusion detection in networks [8], surveillance systems for security [9], as well as forensic investigations [6], [10]. Abnormal behaviour can give information or provide answers during an investigation. Anomaly detection is one way to simplify the big data. We can focus on the data that has been clustered by the anomaly detection method. In this study, a new method to detect abnormal behaviour is proposed in different datasets, combining anomalies and evidence from images automatically.

Data fusion is used to combine data or information from multiple sensors, reports, or databases. Data fusion can enhance the decision-making because it combines data from many sources. Therefore, it has been widely used in multi-sensor environments [11]. Advantages of

data fusion include addressing data ambiguity, as data fusion can improve detection, confidence, and reliability [11]; and the use of multiple sensors, which can improve the result of estimation or velocity (e.g., in tracking moving the object by radar) [12]. Data fusion is a challenging task. According to Lahat et al. [13], there are three challenges in data fusion:

- a) data produced from very complex systems such as biological, environmental, sociological and psychological systems;
- b) increased diversity, including the number, type, and scope of the data;
- c) working with heterogeneous datasets, meaning that the respective advantages of each dataset are maximally exploited and drawbacks suppressed.

In this study, data fusion is applied to challenge problem in heterogeneous data. This study applies many types of data and integrates all using data fusion techniques.

1.2 Aims and Objectives

- a) Develop a new framework and methods for detecting anomalies in heterogeneous data.
- b) Develop new algorithm which combines anomaly detection result using data fusion technique and constructing diagram and sequence of events.
- c) Improve the computational efficiency and assist the human expert in processing huge amount of data proficiently.

1.3 Contributions of the Thesis

The contributions of this thesis are as follows:

a) Heterogeneous data

This study applied different data modalities. In a real situation, data may come from different modalities. Therefore, this study applied data on numerical (financial data), signal data (GPS) and images (face images).

b) Anomaly detection on heterogeneous data

Anomaly detection was applied to find possible anomalies in the original data. This method is novel because it is applied to different data modalities (financial data and GPS signals), is based on new technique recursive density estimation (RDE) and does take context into account.

c) Data fusion for detecting anomalous behaviour

This is a new technique and algorithm for data fusion by integrating heterogeneous data to produce well-organised and summarised data. Data fusion helps in making overall decision based on heterogeneous types of data.

1.4 Thesis Structure

This thesis consists of ten chapters, including this introductory chapter. This section presents an overview of each chapter throughout the research journey of the author.

Chapter 2: Literature Review

Chapter 2 reviews the literature relevant to the study. It begins with an overview of digital forensics, biometrics, and the heterogeneous nature of human behaviour, anomaly detection and data fusion.

Chapter 3: Research Framework

Chapter 3 present the challenges of the study and overall view of research framework implemented in this study.

Chapter 4: Detecting Anomalous Behaviour

This chapter discusses how to detect anomalous behaviour using heterogeneous data. The new methodology of empirical data analytics (EDA) is employed to detect possible anomalies. There are two publications generated from this study. [10], [14]

Chapter 5: Gender and Age Classification for Detecting Anomalous Behaviour

In this chapter, the study uses face images to classify gender and age. Transfer learning is applied based on the deep convolutional neural network to extract the features. Then they are classified using support vector machine classifier. There is one publication generated from this study. [15]

Chapter 6: Data Fusion of Heterogeneous Data for Making the Overall Decision

This chapter describes the new fusion technique introduced to integrate different data modalities of the heterogeneous data. Several algorithms are developed based on the various data types. After all the data is integrated using this new fusion technique, the output can be utilised to make the final decision. One journal paper was published from this study [16].

Chapter 7: Experimental Data

This chapter details the dataset that has been used in this study as a proof of concept. There are two main datasets: the one coming from the VAST Challenge 2014 [17] and one called GAFace.

Chapter 8: Result and Analysis

This chapter discusses the results of the experiments with the data. Then, the analysis of the results is discussed.

Chapter 9: Evaluation

This chapter compares the proposed method with other comparative methods and discusses the outcome from this comparison.

Chapter 10: Conclusion

Finally, the conclusion chapter concludes up the entire journey of this research. It discusses the research summary, limitations and challenges, as well as outlines the future works.

Literature Review

2.1 Digital Forensics

Forensics is a science or body of knowledge using technical methods to solve questions related to criminal activities [1]. It has been developed since the nineteenth century. The forensic process begins after a crime takes place, and can be called a post-event activity [18]. There are six forensic processes: forensic investigation, forensic evaluation, forensic intelligence, automated surveillance and forensic identity management [1]. Forensic science became well known to the public after the success and the popularity of the Crime Scene Investigation (CSI) television show [2]. Even though there are many inaccurate processes in that show, it still gives the public an idea of how the investigation process is conducted [2]. Forensic investigation is a detailed and crucial task and is time-consuming. It uses scientific knowledge and technical methods to solve crime cases. Previously, a forensic investigation was performed using anthropological measurements based on photographic documentation [3]. They used biometric features such as fingerprint and DNA identification which is more accurate and reliable [2]. The evidence collected must be reliable for presentation to the court after the forensic process. The advancement of technologies in digital forensics helped to solve many cases in the justice system.

A definition of digital forensics was provided by the Digital Forensics Research Workshop (DFRWS) in 2001. The participants of the DFRWS workshop suggested a definition of digital forensic as follows:

Digital Forensics is the use of scientifically derived and proven methods towards the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations.

This definition concerns the methods used in digital forensics, which is transitioning from the traditional methods in forensic science [19]. According to Pratama et al. [2], digital forensic science is a method to preserve, collect, validate, identify, analyse, interpret, document and present digital evidence derived from digital sources. Digital forensic principles, procedures and methods shared by practitioners such as defence consultants, corporate investigators, criminal prosecution and compliance officers make the process in digital forensics becomes advanced [19]. Recently, the number and size of digital sources such as in the hard disks, smartphones, tablets and cloud services have grown exponentially. Crimes can involve a network. For instance, fraud often occurs in e-commerce or e-business. Digital evidence such as video or photo images can be used in the court to solve crime cases. Quick & Choo [20] found that a major challenge in the digital forensic analysis is the growth of storage technology and increased storage capacity in consumer devices or cloud storage devices. They also identified the research gap in digital forensic data, including data reduction techniques, data mining, intelligence analysis, and the use of an open and closed source of information. This also includes the application of real-world data and the acceptance of the courts.

Between 1994 and 2014, the literature demonstrated eleven research trends in digital forensics. The top three domains are:

- a) Image integrity analysis (with 137,518 publications)
- b) Surveillance object tracking (32,823 publications)
- c) Fingerprint analysis (29,485 publications)

Research into shoeprint analysis (175), bloodstain analysis (2,008) and surveillance image restoration (3,872) have the smaller number of publication for the same period [2]. Moreover, there are many challenges in digital forensics. The major challenge is the evolution of storage technology [21], [20], and ‘big data’ [19]. Another challenge is system reliability and accuracy, response time, and protecting user privacy [22]. In cybersecurity, real-time response is essential in digital forensics, and providing results in real-time is not as important. The data to be investigated is more significant [21].

2.1.1 Machine learning for forensics

Machine learning is part of computer science which examines learning from data. It is one of the methods employed in data analytics. Machine learning has been applied in many applications such as self-driving Google Car, online recommendations on Amazon and Netflix, linguistic rule creation which can predict what people say about us on Twitter, fraud detection and other forensic applications. Machine learning has evolved from pattern recognition. Most forensics applications use images and require pattern recognition to be processed. Three broad categories of machine learning are supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is learning with a target output while unsupervised learning is no label in the learning algorithm. In reinforcement learning, the agent collects information by interacting with the environment and not passively receives a labelled data set. Two types of information provided by the agent are their current state in the environment, and a real-valued reward (specific task and its corresponding goal).

Machine learning is closely related to statistics and data mining and can be used in making predictions. Approaches to machine learning include the decision tree, artificial neural network, support vector machine, clustering and genetic algorithm. Forensic applications apply these approaches to solving a problem. Mangesh et al. [23] and Navega et al. [24] applied machine learning in forensic anthropology to detect the sex of skeletons. Mangesh et al. combined neural network and support vector machine in the cranial metric dataset and achieved 80% accuracy in classifying male and female skulls. Navega et al. [24] achieved an 88.3% success rate in classifying the sex of the skeleton using a Portuguese sample with a statistical and machine learning method.

Cybercrime is a cybersecurity problem which can apply machine learning to assist in digital forensics. Cyberstalking is a cybercrime in which the suspect or cyberstalker may send email, SMS or instant messages to harass, threaten or insult victims. Zinnar et al. [25] created a framework on cyberstalking to help in finding the identity of the cyberstalker. They discussed a few machine learning methods such as an artificial neural network (ANN) and support vector machine (SVM) to be applied to categorise text to identify a cyberstalker. Reynolds et al. [26] applied a decision tree and instant based learning techniques to determine cyberbullying in the social network with 78.5% accuracy. Al-gharadi et al. [27] applied Twitter network data to detect cyberbullying using machine learning techniques such as Naive Bayes, SVM, Random Forest and k-nearest neighbour (KNN). Then, they proposed a model that provides a feasible solution to detecting cyberbullying.

Machine learning has been applied in forensic image analysis such as in image forgery. Bunk et al. [28] and Bayar and Stamm [29] applied deep learning methods to detect image forgery. Bunk et al. [28] detected and localised images with 94.86% classification accuracy. Bayar and Stamm [29] achieved higher accuracy at 99.10% in detecting the manipulation of an image. Monson and Kumar [30] proposed a method for image forgery detection and

applied Behaviour-Knowledge Space method to classify the image and later point out the area of forgery.

2.1.2 Computational Intelligence for forensics

Computational Intelligence is a fast-growing research field and widely used in the digital forensics. It has many real-world applications such as biometric authentication and identification [31] [32]. Applications include face detection, iris recognition, speech recognition, handwriting and signature recognition [32]. Computational intelligence can be applied to simplify the processes of investigation and analysis of criminal cases and in forensics. It can substitute or augment the human intelligence. The use of computational intelligence is hoped to be better than humans in performing various forensic activities [2]. Forensics is a post-event activity [18]. Forensic investigation begins after a crime takes place and is a crucial task. It requires skills and time to solve the cases. Computational intelligence can be used in assisting the process of investigation. There are many techniques which involve computational intelligence, such as knowledge representation, automated reasoning, inference, pattern recognition techniques and machine learning that can be applied to digital forensic investigation and analysis [10]. Computational intelligence has great potential in finding patterns and solving issues in cyber-security and digital forensics [21]. Computational intelligence is based on biologically inspired computation such as artificial neural network (ANN), genetic algorithms (GAs), and fuzzy logic (FL), and is included in computational intelligence such as machine learning (ML) and autonomous learning systems (ALS).

ANN is a computational model that simulates the nervous system and brain function [33]. Artificial neurons are created from a group of non-linear interconnected elements. They then convert some weight from an activation function to produce some desired output [34].

There are many types of ANN, including feedforward neural network, multi-layer perceptron, radial basis function, and Kohonen self-organizing network [35]. ANN training can be divided into supervised and unsupervised learning. In supervised learning, a neural network obtains different input patterns and determines important features in these input patterns, categorising the data into the right categories. However, unsupervised learning only has input patterns without output data. It aims to learn more quickly than supervised learning. Therefore, it can be used in real time [36].

GAs are inspired by Darwinian evolution [37]. The major role in GAs is selection, mutation, and crossover. The algorithm is based on fitness function to optimise problems [34]. The GAs process starts with generating the population of chromosomes randomly. It represents all possible solutions to a problem. Different chromosome positions (referred as genes) are set as bits, characters and numbers. The fitness function is used to calculate each of chromosomes based on the desired solution [38]. Genetic operators simulate the random variation of a chromosome which modify and combine chromosomes [33]. For example, the crossover is used to exchange part between two single chromosomes to produce offspring [39], and mutation will randomly change individual genes [33].

Fuzzy logic was introduced by Lofti Zadeh in 1965 [40]. It reflects how people think and attempts to model decision making and common sense [36]. Fuzzy logic is based on degrees of membership ranging between 0 and 1 [34], [36]. There is a difference between the mathematical and mental representation in set theory. Mathematical representation based on binary logic is too rigid for gradual membership. In contrast, mental representation is based on natural language terms such as fast speed, tall man and beautiful lady which are diffused with vagueness [39]. The meaning of sign language is concerned by vagueness. Vagueness is not fixed by boundaries. Fuzzy sets allow partial element that belongs to a set. Thus, results

are not limited to true or false but in some degree of true or false. Fuzzy logic can use vagueness to get a better answer [39].

Fuzzy logic is a technique that can be applied in forensic investigations. For example, Al Amro et al. [41] combined fuzzy inference with linguistic variables to detect the security threats. Fuzzy systems are more efficient when detecting anomalous behaviour [41]. Classification and clustering using fuzzy techniques produce an effective result, such as when dealing with malicious attacks [41]. Also, fuzzy logic can be instrumental in forensic investigations, because it is a human-intelligible and transparent form of expression, has excellent robustness and scalability and can deal efficiently with the subjective type of uncertainty and ambiguous cases. Many forensic applications are based on ANN techniques [42]. For example, Boo and Ahalakoon [43] applied Growing Self-Organizing Maps (GSOM) to perform crime profiling computationally, mimicking the human profiling processes. Rodriguez et al. [44] applied Growing Neural GAs (GNG) in surveillance systems to detect and interpret the action in video sequences. GNG is a kind of self-organising network introduced in the early 1990s. Beebe and Liu [45] compared clustering algorithms for digital forensic text string search output. Research showed that the Latent Dirichlet Allocation and k-means method produces best results.

Autonomous learning systems (ALS) are part of machine intelligence and closely related to artificial intelligence and cybernetics. They are used in advanced industries as well as in defence and aerospace, but also in many new products [46]. An autonomous system can self-learn and self-develop its internal structure and representation of the environment that surrounds it. Such systems are capable of learning rapidly and flexibly. Many applications use autonomous learning systems in the chemical and petrochemical industries [47], mobile robotics [48][49], object detection and tracking [50][50], [51], fault detection [52][7] and modelling evolving user behaviour [53][54].

Specific techniques and methods that fall into the category of ALS include, but are not limited to:

- a) Evolving Clustering - eClustering [55], Evolving Local Means [56], Dynamically Evolving Clustering [57], Data Density Based Clustering [58], and Evolving Clustering Method [59]
- b) Evolving Classifier - eClass [60][61], AutoClassify [46] and TEDAClass [62]
- c) Evolving Predictors - eTS [63], neuro-fuzzy systems with nonparametric antecedents called AnYa [64]; evolving controllers [65][66], self-calibrating sensors and eSensor [60]
- d) Anomaly detection method within EDA – RDE [46] and standardised eccentricity

The pillar of ALS is their ability to develop an internal structure in response to changing data patterns based on the computationally efficient recursive density estimation (RDE) method [46]. In addition to that, the evolving Takagi-Sugeno fuzzy system forms human-intelligible linguistic descriptions which are easy to understand and interpret. Indeed, humans solve problems based on rules. Therefore, fuzzy rule-based systems can help forensic investigators in decision-making by producing sets of rules to solve the cases. For example, AnYa is a type of fuzzy rule-based systems with simplified linguistic expressions in comparison with the Mamdani and Takagi-Sugeno types [46]. Therefore, such a rule-based system can be more efficient in extracting knowledge from data as well as to formalise the existing expert knowledge. Hence, it can assist forensic investigators in solving cases.

2.2 Biometrics

Biometrics are popular in security systems [67]. Such systems can be divided into two categories; physiological and behavioural (Table 2.1). Forensics and Biometrics complement

each other. Biometrics concern a pre-event activity while forensics is the post-event activity [18]. Biometrics is used to identify and verify a person's identity. Identification of a person using biometric techniques is reliable, unlike presenting an ID card, for example. Biometric system properties include the following [68]:

- a) Universality
- b) Uniqueness
- c) Permanence
- d) Measurability
- e) Performance
- f) Acceptability
- g) Circumvention

Table 2.1 Categories And Comparison of Biometric Systems (High, Medium and Low are denoted by H, M, and L, respectively [69])

Biometric Category	a	b	c	d	e	f	g
Physiological							
• Face	H	L	M	H	L	H	H
• Fingerprint	M	H	H	M	H	M	M
• DNA	H	H	H	L	H	L	L
• Palmprint	M	H	H	M	H	M	M
• Iris	H	H	H	M	H	L	L
Behavioural							
• Keystroke	L	L	L	M	L	M	M
• Signature	L	L	L	H	L	H	H
• Voice	M	L	L	M	L	H	H
• Gait	M	L	L	H	L	H	M

Biometrics is an alternative in the forensic science, and have a wide application outside of forensics as well. They are a class of emerging technology that provides the identification and verification of people. Identification based on “who you are” is more recognised rather than “what you possess” (for instance, Smart Card) or “what you have” (for example password)

[68]. The identification process uses biometrics because of uniqueness. People cannot borrow, steal, or forget biometric traits. Biometric authentication is an automatic method used to identify and verify users using unique physiological or behavioural traits. Physiological traits are biometric features that are owned by the people. Examples of physiological traits include fingerprint, iris, ear and facial characteristics. The leading biometrics widely used in forensic applications include fingerprints, facial characteristics, and iris patterns [22].

Fingerprint identification has been used since the eighteenth century [3]. In 1970, law enforcement agencies such as Interpol, FBI, Home Office in the UK, and Police Department in Paris, started developing Automatic Fingerprint Identification Systems (AFIS) to improve the efficiency and accuracy of fingerprint matching [70]. Automatic fingerprint technology has been widely used in many applications such as PC logins, electronic commerce, ATMs and mobile phones [68]. Research on fingerprint identification systems still faces challenges related to system performance. The current challenge is the identification time [22]. Another challenge is related to latent fingerprint [71]. A latent fingerprint is a significant evidence taken from a crime scene. Its resolution, therefore, may differ from live-scanned fingerprints. It also needs special care to enhance the quality of feature extraction and matching operations.

The face is one biometric trait that can be easily captured from the surveillance systems. There are three main issues in capturing facial images in forensic investigation. These are:

- a) facial image immutability over time
- b) facial image retrieval
- c) matching forensic (sketched) images with real captured facial images

The effect of ageing is also a great challenge for the facial recognition development. Little research has focused on addressing the ageing effect problem for the reliable recognition system. Another issue is the face occlusion problem [22]. From the face image, age and gender

can be detected. Since 2001, there has been a great deal of research on detecting age and gender from face images [72][73][74].

The iris allows for high accuracy in identifying a person. Iris is better than other types of biometrics because of its uniqueness and is stable for extended periods of time [75], [67]. According to Awad and Hassanien [22], one issue regarding iris is related to capturing the iris images. User misbehaviour can lead to problems when capturing irises. Problems include blurred irises and irises occluded by eyelids and eyelashes. The new direction of individual identification and authentication is to use saccadic eye movement [22]. Additionally, it requires very close physical positioning of the person.

Behavioural traits concern the behaviour of individuals that are acquired and learned over time. Gait, signature, voice, handwriting and keystroke dynamic are examples of behavioural traits. Gait is an evolving biometric and attracts many researchers and industry [76]. Gait is defined as the way of walking and can be recognised at a distance. Thus, gait can be captured from a surveillance video. However, choice of footwear, walking surface, nature of clothing, and so on can affect the individual's gait [70]. A signature is a behavioural form of biometric influenced by physical and emotional factors and may change over time [70]. Authentication using signature biometrics does not require special devices. It only uses a pen and a piece of paper. Signature played a significant role for centuries in document authentication, for example, in banking, government documents, forum applications, and so on. Therefore, signature biometrics is widely used in forensic applications. The aim of using signature biometrics in forensics is to prevent crime from happening [77].

Keystroke dynamics refers to a pattern of typing, typically how users type their passwords [78]. In [79], [80] the authors claim that keystroke dynamics are a behavioural measurement in biometrics that helps to identify users based on their typing rhythm. The significance of keystroke dynamics is that they help increase the security of a system. Nowadays, there are

many activities on the Internet such as email, online banking and online shopping. If an authentication system is not safe, private profiles can easily be accessed. Pisani and Lorena [81] stated that passwords are a method of authentication but can easily be copied and guessed. An access card can be used as authentication but can be stolen. There are many identity thefts because of the weaknesses of the passwords. Keystroke dynamics are a form of biometrics that can be used without additional cost because they incorporate the common keyboard. Compared to a traditional signature, the system still needs to define which users are legal and which are illegal.

2.2.1 Machine learning for biometrics

Machine learning is a technique that has been successfully applied in biometrics. Images or signals are common data needed in biometric systems. Face recognition is a biometric system that has been under research since the 1960s. The convolutional neural network (CNN) is one of the latest technique that been implemented in face recognition. Nakada et al. [82] proposed an active face recognition system to identify human behaviours in three common scenarios: changing viewpoint, greeting, and ignoring. For classification, K-NN produces 90% results compared with SVM and linear regression at only 20%. Wen et al. [83] also applied CNN to recognise the face. Ucar et al. [84] studied face expression using a new technique with a curvelet transform and online sequential extreme learning machine, then spherical clustering to find the optimal hidden node.

Gait recognition is used to recognise human behaviour. Charalambous and Bharath [85] found confounding factors that can reduce the usefulness of gait in the biometric systems. Due to this, they conducted a simulation-based methodology to generate synthetic video frames for data augmentation gait recognition. This method uses deep supervised learning algorithms to

the augmented labelled training data. Reynard and Terrier [86] and Shirakawa et al. [87] used gait to determine healthy adults by looking their gait on the treadmill. Reynard and Terrier [86] applied random forest and multiple adaptive regression splines to access the preferred walking style for an adult who can influence their gait. However, Shirakawa et al. [87] clustered the walking pattern of adults. They applied principle component analysis (PCA) and hierarchical clustering.

The EEG signal is fast becoming one of the most popular modalities for biometric authentications. EEG signals can be used in identification as an alternative to password, PIN or card which can be forgotten or stolen. Bashar et al. [88] applied SVM and get 94.44% accuracy in identifying human using EEG signal. Johannesen et al. [89] also applied SVM with EEG data but to classify working memory performance in healthy human and schizophrenia patient. They achieved 87% of accuracy on distinguishing individuals with schizophrenia. However, most methods for EEG signal acquisition are intrusive and require a special electrode head-set to be able to work which makes them inconvenient.

2.2.2 Computational Intelligence for biometrics

Many computational intelligence techniques have been applied to real-world applications, including biometrics. Most biometric applications use images, text and signal as sources of information. Images are usually captured from cameras or closed-circuit television (CCTV). Image data (fingerprint, face, iris or gait) can be processed using image processing techniques. Techniques such as ANN, GAs, support vector machine (SVM) are used to reduce the processing time and increase the system accuracy [22]. Computational intelligence techniques are also used for pre-processing raw data, identification or verification of data for biometrics. Many types of research have applied hybrid systems to obtain better results [90]. Both,

Sánchez et al. [90] and Sánchez and Melin [91] applied modular granular neural networks and genetic algorithm for human recognition based on iris and ear. They reported good results in human recognition based on both, iris and ear of individuals. Mai et al. [92] applied a multilayer perceptron and radial basis function type ANNs to identify electrocardiogram (ECG) signal of individuals. Beltrán et al. [93] combined ANNs and fuzzy logic methods to find patterns in the fingerprint. GAs is used to optimise the fuzzy rules and then to combine these with ANNs. This research obtains good results especially for blurred patterns in fingerprint analysis. Tiwari et al. [94] researched retrieving face images using Self-Organizing Map type of Neural Networks. They proposed an algorithm that can retrieve an image in one second. Research is done in [95] applied kernel-based machine learning to recognise face and iris. There is research on human classification using gait features. They applied SVM and achieve higher results [96], [97]. Boyadzieva and Gluhchev [98] classified forgeries into random and skilled categories by verifying the online signature using K-nearest neighbour (KNN) method combined with ANNs. These are only some of the many reported types of research applying computational intelligence to biometrics to produce better results.

2.3 Heterogeneous Nature of the Human Behaviour Data

Data is a raw material which available both offline and online. In the 1970s to 1980s, most data were scattered in files, reports or books. In that time, data were hard to process and time-consuming. More recently, most data are digital and easy to access. Data become “big data” because it is generated every day and growing exponentially. According to International Data Corporation UK, [99], data will reach 16 zettabytes in 2017. Data is created from sensors to gather climate information, social media such as Facebook or Twitter, smartphone applications such as GPS signal, digital picture and video, and purchase transaction records.

Digital data such as text and images from social media are an example of unstructured data which have different modalities and are sometimes too short and contain noise. Due to the enormous amount of data, data scientist processed all of this data to produce beneficial results which can assist the expert in making decisions. Data science is a new area of study which helps to process and extracting knowledge and information in various forms [100]. For example, when people go shopping and buy grocery, the supermarket gives customers a loyalty card. Customers give personal information to the supermarket, and they can establish spending behaviour from shopping lists using data mining techniques. Special offers will be given on selected items a particular customer purchases, and this will make the customer happy and loyal to the supermarket. Moreover, there are many ways to detect human behaviour which later can help an investigator if sometimes abnormal happens.

Digital data is generated every day in exponentially growing quantity, and it has become really “big data”. More than 2.5 exabytes of data is being created every day, and the amount doubles every few years [101]. In [99], the authors forecast that digital data will reach 16 zettabytes in 2017. Data can be seen as a raw material in various forms such as text, numbers, images, video or signals. This diversity of the data leads to heterogeneous data which may be structured or unstructured. Structured data is data in the same format and is easy to organise, while unstructured data does not have a common structure; for example, emails, images, and so forth. It is hard to combine and computationally analyse such data. In the age of big data, one of the challenges in the data analysis is how to process and integrate heterogeneous, unstructured data such as social media data, images and streaming data [102]. Extracting knowledge from text and images, social media such as email, Twitter, and Facebook has become an issue because the data is in different modalities and is sometimes too short or noisy (e.g. text messaging and WhatsApp) [103]–[106].

The data can be seen as a raw material (e.g. facts, numbers, letters and symbols) that can be extracted from observations, experiments, computation and record keeping [107]. Until the 1980s, most data were scattered. Text data can be found in documents such as letters, reports, books or journals and processed as image data (analogue photography, radio, and telephony; images were produced from negative film or drawing, and signal data can be recorded using vinyl records or compact cassette and transmitted through analogue communications). Text data is unstructured and requires the natural language processing (NLP) technique to pre-process the data. Features such as keywords, topics, and so forth need to be extracted from the data. Then, the text data can be processed to produce information. Image data differs from text data. Digital images consist of binary representations. Many formats are available, such as jpg, bmp, gif and png. The size of the digital images is based on the number of pixels. There are several steps in pre-processing digital images such as image resampling, segmentation, grey scale and noise removal. Previously, all of this data had to be processed manually before becoming useful information. The process is difficult and time-consuming.

The first modern computer was introduced by John Vincent Atanasoff [108]. After this invention, many data storage technologies were developed, and the size of storage become huge [109]. For example, in 1956 a computer hard disk size was only 5 megabytes; however, in 2016 the size may be ten terabytes. Following the Internet revolution in the early 1990s, more data has been created from email, file sharing (FTP), and telephony (voice, fax, SMS, voice messaging). Since the mid-2000s, when the iPhone and Android were introduced, there have been many applications developed which generate more and more data every day, especially in social media applications. Now, there are many digital devices in homes, workplaces and public places, including mobile, distributed, and cloud computing, social media and the Internet of Things. These platforms are important to all aspects of the everyday life such as work, communication, travel and leisure. All of these generate data signatures

[110]. With digitalisation, traditional databases have moved to network data infrastructure and more data is publicly available, leading to data revolution which brought to life the term “big data”. In business, for example, big data providing new resources for company activities and can leverage additional profit by enhancing productivity, competitiveness and market knowledge.

Heterogeneous data includes data forms or types such as qualitative or quantitative, structured or unstructured, and primary or secondary. In reality, most situations or applications will produce heterogeneous data such as when the crime takes place; evidence can be in the form of many types of data. The combination of all the data types can produce rich information insights. Nowadays, data becomes ‘big’ because it is generated every second of every day. The processing part is time-consuming and tedious. Also, different data must be processed differently based on data features. Features extraction need to be carried out to select the relevant features. This phase is a pre-processing phase. After processing data, anomaly/outliers can be detected, and in statistics, outliers should be removed to prevent errors in results [111]. However, in some situations such as forensic applications, anomalies or outliers can represent significant information.

2.4 Anomaly Detection

Abnormal data can be detected using anomaly detection techniques [4]. Anomaly detection is one of the methods for data analytics which aims to identify the data samples that “stand out”, are “untypical”, differ from normality significantly. It can also differentiate between normal and abnormal behaviour. There are many types of problems related to anomaly detection. These include the nature of the input data, types of anomalies (points, contextual or collective anomalies), data labels (supervised, semi-supervised or unsupervised)

and outputs of the anomaly detection [112]. Anomaly detection is crucial in the analysis of fraud detection, drift detection in data streams [113], clustering, outlier detection and autonomous video analytics [4]. The result of such detection is used in many applications such as intrusion detection in cybersecurity [114], fraud detection [6], surveillance system [115] and military surveillance of enemy activities [116].

Traditionally, anomaly detection is addressed using statistical methods where frequentistic techniques representing probabilities are applied, and *prior* assumptions must be made [4]. The main decision is traditionally made using threshold values. These thresholds are based on the normal distribution of random variables (usually assuming Gaussians) while for arbitrary distributions, they are based on the well-known Chebyshev inequality [4]. These approaches have the following disadvantages [4]:

- a) they require strict *prior* assumptions;
- b) they relax the conditions too much to avoid false positives to the level where it misses many true positives (the 3σ rule sometimes fails to detect some obvious outliers);
- c) Many data samples are required;
- d) a single data sample is compared with the average, instead of comparing pairs of data samples; therefore, the information is blurred and is no longer point-wise and local.

According to [4], eccentricity can be applied to avoid the disadvantages of the traditional statistical method. This approach does not require any prior assumption, and a σ_{gap} can be formulated between the eccentricities of the data samples with the larger eccentricity [4].

2.4.1 Types of Anomaly

Anomalies can be classified into three types: point, contextual, or collective [112]. The point anomaly is the simplest type of anomaly, in which any single data point has different

attributes value from the rest of the data. This type of anomaly can be identified in appropriate visualisation [117]. The simplest plot is two-dimensional. For example, figure 2.1 shows points A1, A2 and A3 as anomalies since they are far away from the other points and have different features. An example of real-life application is an individual's credit card transactions. If the amount spent is very high compared to another spending, this is known as a point anomaly.

Contextual anomalies refer to a specific context [112]. For example, in the same credit card transaction, if we take the time and the amount spent, then we can relate why there is too much spending in December, because of so many things to buy during Christmas and can be considered normal. However, it will be an anomaly if the spending is too high in May. Contextual anomalies are mostly applied to time series data. Another example of time series data is the temperature in climate data. During the summer season, it is normal to have a temperature between 15° to 30° during summer time in England. If the temperature is below 5° during summer, then it can be considered an anomaly. It is not easy to define the context and to depend on data and applications.

Collective Anomaly is a collection of data instances which is anomalous concerning entire data set [112]. Collection of anomaly is not anomalies when looking into individual data, but when the data is together as a collection is anomalous [117]. The way to detect collective anomalies is different from the point and contextual anomaly. An anomaly in sequential data is an example of a collective anomaly. One example of a collective anomaly is a sequence of events (buffer-overflow, FTP) which corresponds to a typical web-based attack. The attack is using remote machine and copying data from the host computer to a remote destination. The event can be called a collective anomaly. However, individual events are not anomalies when they occur in other locations in the sequence [117].

2.4.2 Categories of anomaly detection

Anomaly detection can be categories into three main types[117]:

- a) supervised
- b) semi-supervised
- c) unsupervised.

Supervised anomaly detection has normal and anomalous data in the training set, and the data has a label. Then, this data usually can be trained by supervised machine learning to find the anomaly. The difference between supervised machine learning and supervised anomaly detection is that the training data is unbalanced. The probability of anomalies is usually below 5% (0.05) [117]. This anomaly detection can be categorised as extremely rare because anomalies are by nature often new and occur for the first time. Due to this, anomaly detection is also known as novelty detection. The axioms of anomalies also state that anomalies differ in their features and are rare [117].

Semi-supervised anomaly detection looks similar to supervised anomaly detection. However, in semi-supervised anomaly detection, normal data are used in the training data, and the testing phase will be used to find the anomalies [117]. The anomalous class or target does not need to be known in advanced. The traditional method of semi-supervised anomaly detection is using clustering algorithms [117]. Normal data will be clustered during the training phase, then in the testing phase, the distance to the cluster centre is computed. If the data has a large distance to the centre, it may be considered anomalous.

In unsupervised anomaly detection, data contains anomalies and normal instances without any label [117]. Unsupervised anomaly detection cannot be separate from training and testing, and the algorithm is based on distance information. This anomaly detection is most powerful and flexible because it can distinguish without any label [117]. It is necessary

for unsupervised anomaly detection that anomalies need to be rare, and the features must be different from the normal behaviour. The algorithm normally performs density estimation such as done by [46].

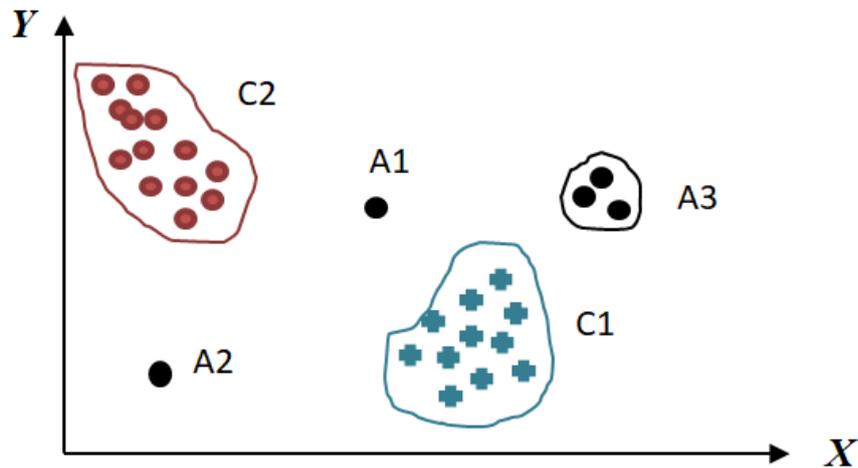


Figure 2.1: Example of point anomaly

2.4.3 Chebyshev Inequality

Chebyshev inequality is a theorem in probability theory introduced by a Russian mathematician Pafnuty Chebyshev in 1867. Chebyshev inequality normally used when data distribution is unknown. It is calculated as follows [118]:

$$P(|X - \mu| \leq k\sigma) \geq (1 - \frac{1}{k^2}) \quad (2.1)$$

Where X denotes the data, μ is represent the mean of data, σ is the standard deviation of the data and k denotes the number of standard deviations from the mean. The observation is

independent of each other because there is no assumption on data distribution. At least 75% of the data would fall within two standard deviations ($k = 2$) from the mean [118].

If the data is normally distributed (based on Gaussian) and represents a significant amount of data sample, the vast majority of the data is 99.7% is considered normal (if use 3σ) [4]. The probability of the data is anomalous is 0.3%. In order cases, where the data distribution is unknown, Chebyshev inequality is applied when no more than $\frac{1}{n^2}$ of the data samples are more than $n\sigma$ (where σ denotes the standard deviation) the away from the mean [4]. The probability of data point distant from the mean is greater: for example, 3σ is $< \frac{1}{9}$ (11%). A large amount of data can be considered anomalous. Therefore, for this case (unknown distribution), many researchers use 6σ (to get 3% or $< \frac{1}{36}$) or even higher if the data is declared anomalous [4].

2.4.4 Thresholds based (subjective) method

The simplest statistical anomaly detector is the simple threshold method. This method is commonly used in IT analytic applications. If the data falls inside or outside a specific range of values then the simple threshold method flag data as anomalous [119]. This method is very fast to compute and required virtually no CPU or memory to compute and suitable for thousands of metrics data for the need for straightforward detection. Simple threshold methods have the following disadvantages:

- a) The threshold value is normally set by the user. It will be challenging and time-consuming when there are many data streams. The threshold also needs to be adjusted based on environment changes. Thus, manual real-time is required.

- b) A simple threshold can generate many false positives because it cannot learn new changes or behaviour. For example, if IT staff increases memory on a server, the percentage of memory used may fall below a threshold.
- c) The threshold does not consider the temporal sequences. Therefore, the simple threshold cannot identify a pattern within the range.

The threshold in anomaly detection also called $n\sigma$ principles. In the normal distributed random variable, for example, Gaussian, if used 3σ , >99.7% will be considered “normal” and probability of abnormal data is <0.3%. Another example is if used 2σ , the percentage of normal is 95.45% and abnormal 4.55% [4]. In any distribution of data, Chebyshev inequality is applied which state that no more than $\frac{1}{n^2}$ of the data points are more than $n\sigma$ away from the mean (where σ denotes the standard deviation) [4].

2.4.5 How to define “normality.”

Before detecting an anomaly in data set, normal data should be defined first. It is not easy to define normality. For example, in a medical application, small deviation might be abnormal such as fluctuation in body temperature, though similar deviation in the stock market might be considered as normal [112]. Therefore, it is not straightforward to apply any anomaly detection technique one application to another. Most of the anomaly detection is a problem specifically because it is not easy to solve anomaly detection problem in general. Various factors are involved such as nature of data, availability of labelled data, types of anomalies to be detected [112]. Many studies applied statistics, machine learning or data mining techniques to a specific problem formulation [112].

2.4.6 Applications of Anomaly Detection

Anomaly detection is a method or process to identify data samples which is different from normal behaviour in datasets. In statistics, anomaly or outliers will be removed to get a better result of analysis [120]. Anomalies or outliers are important in some applications. Anomalous data can help solving a problem and giving a better solution for a certain problem such as in forensic applications, medical diagnosis, and surveillance systems. This method has been successfully applied in:

a) Intrusion detection systems (IDS) – IDS can monitor cyber-attacks or cyber threats in the network systems and server applications. Anomaly detection evaluates monitoring data against normal baseline and will issue an alert if there is an occurrence of the abnormal behaviour. Challenges in IDS are the big heterogeneous data which need to be processed in real-time [8]. Zuech, Khoshgoftaar, and Wald [8] stated that correlating security events from various heterogeneous sources such as network and server could enhance the cyber threat analysis and cyber intelligence. Machine learning can be used to learn the nature of the normal traffic behaviour autonomously, which can adapt normal structure and recognise suspicious or anomalous events [121]. This technique can detect unknown attacks and not be restricted to any specific environment.

b) Fraud detection – Anomaly detection has been successfully applied in financial application to detect fraud [6]. Fraud detection is important, especially, in the era of electronic commerce which involves electronic payment systems [122]. Digital transactions always carry a risk, and the scammers are hard to be identified. Behavioural profiling method can model each behavioural pattern and detect the abnormal behaviour in a transaction [123]. Credit card fraud detection also applies the same technique, which is after the cardholder profile is

created; the system can analyse spending behaviour patterns of the user. If an inconsistency appears throughout the transaction, then the suspicious activity can be detected [124].

c) Medical applications – Patient monitoring such as electrocardiography (ECG) is an example of a medical application that utilises anomaly detection. ECG is a test that produces a signal from the heartbeat to monitor possible heart problems [125]. The reported application also develops a new algorithm based on time series to detect an abnormality in the ECG signal. However, the algorithm cannot be applied to large datasets in real-time interaction. Meanwhile, Salem, Guerassimov, Marcus, and Furht, [52] also detected anomalies in ECG using a combination of various patient metrics such as blood pressure, body temperature, respiration rate, and blood glucose level. This study develops a wireless sensor network to detect an anomaly in the patient's body and achieved high detection accuracy.

d) Surveillance systems – Surveillance systems use closed-circuit television (CCTV) and are installed in buildings for security purposes. The system analyses suspicious movement recorded in the video and categorised them as an abnormal activity. Examples of surveillance system applications have been reported in [115], [126]. A surveillance system applied to the train platform to monitor people jumping or falling off train platform is reported in [115], which can be considered as an abnormal activity, while Li, Mahadevan, and Vasconcelos [126] created a model of the crowded scene and applied benchmark dataset to detect an anomaly in the pedestrian walkway. These two applications were utilised in crowd environment and Delgado, Tahboub, and Delp [115] claim their system to be capable of achieving 90% accuracy in detecting an anomaly. Li et al. [126] achieved a better result as compared with other techniques.

e) Maritime surveillance systems– Anomaly detection was applied in maritime surveillance systems to assist in finding abnormal behaviour in the trajectory of vessels. Maritime surveillance systems used conventional sensors such as radar, video camera, or

aircraft to observe enemy activities. In [116] authors apply topology preserving mapping (TPM) and Pallotta, Vespe, and Bryan [127] applied rule-based and log-likelihood methods to detect an anomaly. These are unsupervised methods. TPM can capture and visualise vessels' behaviour, and have a probability estimator which later can evaluate likelihoods and detect anomalies;

2.5 Data Fusion

Data fusion is used to enhance the decision-making because it combines data from many sources. It is widely deployed in multisensor environments [22]. The goal of data fusion is to combine and aggregate data derived from several sensors. These techniques can be applied to the text processing domain as well [23]. There are many definitions of data fusion in the literature. The Joint Directors of Laboratory (JDL), known as Data Fusion Group, began collecting terminology related to data fusion in 1986. JDL defines data fusion as the multilevel and multifaceted association, correlation, estimation, and combination of data and information from single and multiple sources [128]. Hugh Durrant-Whyte defines data fusion as the integration of a multi-sensor system problem into single best-estimate of the state of the environment [129]. Data fusion can be applied to several fields including signal processing, statistical estimation, inference, information theory and artificial intelligence. Advantages of data fusion include enhancement in data authenticity or availability. Data fusion can improve detection, confidence, and reliability, as well as reduction in data ambiguity and benefits for some application contexts such as wireless sensor networks composed of a large number of sensor nodes, which face new scalability challenges caused by potential collisions and transmission of redundant data. Communication should be reduced to increase the lifetime of the sensor nodes due to energy restrictions, when data fusion is

performed during the routing process, sensor data is fused and only the results are forwarded. Thus, the number of messages is reduced, collisions are avoided, and energy is saved.

2.5.1 Challenging in Data Fusion

The most challenging areas of data fusion are data imperfection, data inconsistency, data correlation and data disparateness [11]. Sometimes sensor data may have uncertainty, ambiguity, vagueness, incompleteness or granularity data. A mathematical theory can represent data imperfection. For example uncertainty can be addressed using probability, ambiguity using Dempster-Shafer, vagueness using fuzzy set theory, incompleteness using possibility theory, and granularity using rough set. Data fusion approaches such as the Kalman filter (KF) require independent or prior knowledge of the cross variance of data to produce consistent results. Fusion data are correlated with potentially unknown cross variance in many applications. Data correlation can lead to biased estimation, for example, artificially high confidence value. Inconsistent data fusion occurs because of spurious data points or outliers, disorder and conflicting data. Spurious data maybe acquire because of an unexpected situation such as permanent failures. If spurious data fused with correct data, the result may lead to dangerous, inaccurate estimation. For example, if exposed to outliers, the Kalman filter algorithm can easily break down. Conflicting data must be handled specially because it can produce counter-intuitive results. The fusion system may receive input data from a variety of sources such as sensors, humans, or archived sensors. It is a difficult task for a fusion system to build a coherent and accurate global view using disparate data. However, there are some fusion applications such as human-computer interaction (HCI), in which it is necessary to have a natural interaction with humans, who generate soft data as opposed to the hard data of sensors. There is very limited work on data fusion from human

and non-human sensors [11]. Hall et al. [130] briefly discuss dynamic fusion of soft and hard data. They also present the motivation and challenges also advantages and requirements for data fusion of soft and hard data.

2.5.2 Data Fusion Classification Techniques

Data fusion is a combination of data and related information to improve information and produce a better decision. Data fusion has been employed in multisensor environments to achieve lower detection error probability and higher reliability based on multiple sources of data. Based on a review by [131], there are three nonexclusive categories of data fusion techniques: data association, state estimation and decision fusion. Data fusion is difficult to reach a clear and strict classification because it involves several fields in multidisciplinary areas. Data fusion techniques can be classified into the following [131]:

a) Classification based on the relation between the data sources

This classification was introduced by Durrant-Whyte (figure 2.2):

- i. Complementary – The input is the different part of information and then combined to obtain complete information. For example, two cameras provide information on the same target but with different fields of view;
- ii. Redundant - Two or more inputs provide the same information and maybe overlapped. For example, input data are from overlapped areas in visual sensor network;
- iii. Cooperative – Provided information is combined with new information which is more complex than the original information. For example, multimodal data fusion includes audio and video.

b) Dasarathy's Classification

Dasarathy's classification is one of the well-known data fusion classification systems and there are five categories in this classification (figure 2.3):

- i. data in-data out – The most basic data fusion method. Data fusion process input and output raw data and the results are more accurate or reliable. Data fusion is processed immediately after received data from the sensors. Signal and image processing are the algorithms that will be applied at this level;
- ii. data in-feature out – Data fusion is generated raw data from the sensor and extracts the features that explain in the environment.
- iii. feature in-feature out – Features in the data fusion process are both the input and output. Therefore, the features in the data fusion a process are addressed to improve, refine, or obtain new features.
- iv. feature in-decision out – Set of features is obtained as input, and set of decisions becomes an output.
- v. decision in-decision out – This classification is decision fusion, which fuses the input decisions to acquire better decisions.

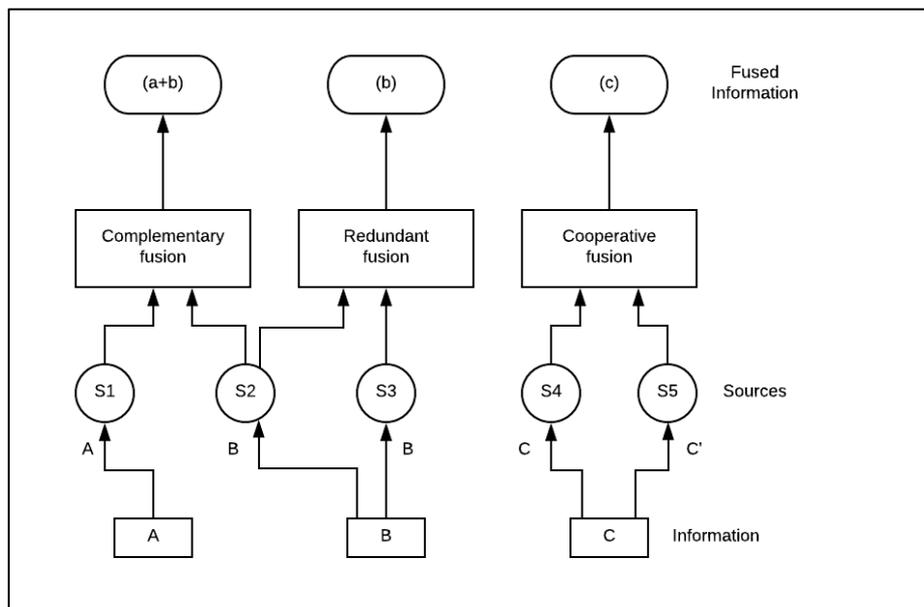


Figure 2.2: Durrant -Whyte's Classification Based on the Relations between the Data Sources [131]

c) Classification based on the abstraction levels [132]

There are four abstraction levels: signal level, in which the signal from the sensor is directly addressed; pixel level, in which an image level is improved for image processing tasks; characteristic level, in which features are extracted from the image and signal; and symbol level, in which information is represented as symbols, known as the decision level.

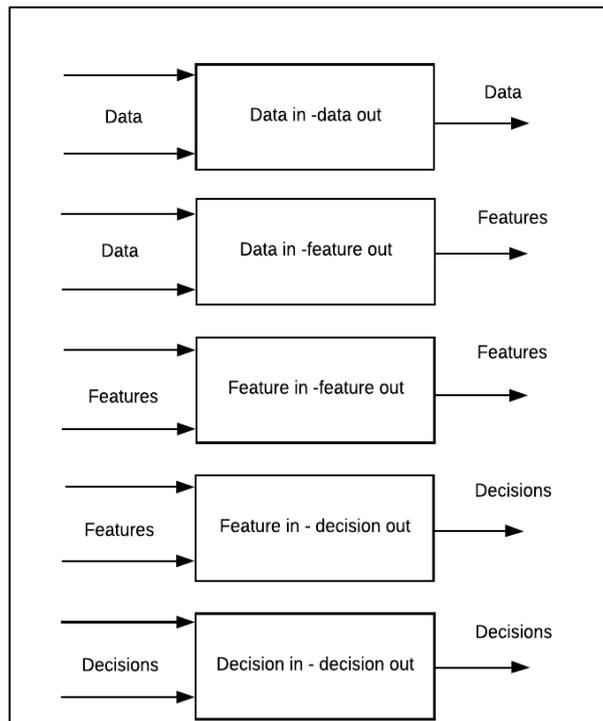


Figure 2.3: Darasathy's Classification [131]

d) JDL Data Fusion Classification

The most popular conceptual model of data fusion was proposed by JDL and the American Department of Defense (DoD). They classified the data fusion process into five processing levels.

- i. level 0 – source pre-processing – The lowest level in data fusion, which includes fusion of signal and pixels level. This level maintains important information and reduces the number of data for high-level processes;

- ii. level 1 – object refinement – Received data from level 0. Common procedures include spatiotemporal alignment, correlation, clustering, association, identify fusion and combine the features that have been extracted from images. Output from this level includes object discrimination (classification and identification) and object tracking (state of the object and orientation);
- iii. level 2 – situation assessment – Higher level inference than level 1. This level purposes to identify the possible situation from the observed events and acquired data. It creates links between objects (e.g. proximity, communication). Performing high-level inferences and identifying significant events are also the aims of this level. The output from this level includes a set of high level inferences;
- iv. level 3 – impact assessment – The impact of detected activities in level 2 is evaluated to acquire a proper perspective on this level. This level will evaluate the possible risks, vulnerabilities, and operational opportunities and then predicted the logical outcome;
- v. level 4 – process refinement – Improvement of processes from levels 0 to 3 is performed at this level to offer management of resources and sensors. The goal in this level is to achieve efficient resource management such as task priorities, scheduling, and the control of resources.

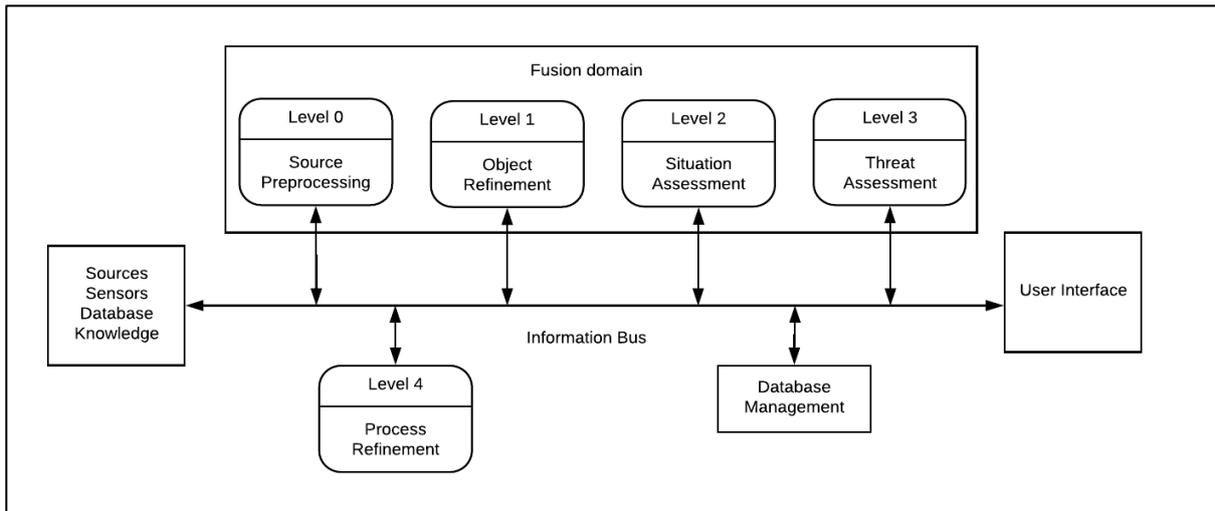


Figure 2.4: JDL Data Fusion [131]

e) Classification based on the type of architecture

When designing a data fusion system, a key issue is where the data fusion process will be implemented. The following are data fusion classifications based on the type of architecture:

- i. Centralised architecture – The fusion node resides in the processor which in the centre of the architecture and all input sources received information in the processor;
- ii. Decentralised architecture – Every network node has its processor, and there is no single point of data fusion. Data fusion is processed autonomously where every node fuses its local information with the information received from its peer;
- iii. Distributed architecture – Each source node is processed independently. Then, the information is sent to the fusion node. Before the information is communicated with the fusion node, data association and state estimation are performed in the source node;
- iv. Hierarchical architecture – This architecture is the combination of decentralized and distributed nodes. Data fusion is processed at a different level in the hierarchy.

2.5.3 Data Fusion Algorithms

Many algorithms have been developed to apply data fusion. Algorithms can be divided into data association, state estimation and decision fusion. Data association is the process of computing measurements that correspond to each other [131]. Nearest neighbours, k-means, probabilistic data association (PDA), joint probabilistic data association, distributed joint probabilistic data association, distributed multiple hypothesis test and graphical models are examples of data association algorithm. The basic algorithm to perform data association is nearest neighbours and k-means. These algorithms are clustering algorithms which calculate distances between instances and then cluster the instances.

State estimation technique is a tracking technique used to determine the state of the target based on observation and measurements [131]. The most common estimation methods are maximum likelihood and maximum posterior, Kalman filter, particle filter, distributed Kalman filter, distributed particle filter and covariance consistency methods. Decision fusion is a high-level inference about event and activities that are produced from the detected target (normally use symbolic information). Examples of algorithms to apply decision fusion include Bayesian methods, Dempster-Shafer inference, Abductive reasoning and semantic methods

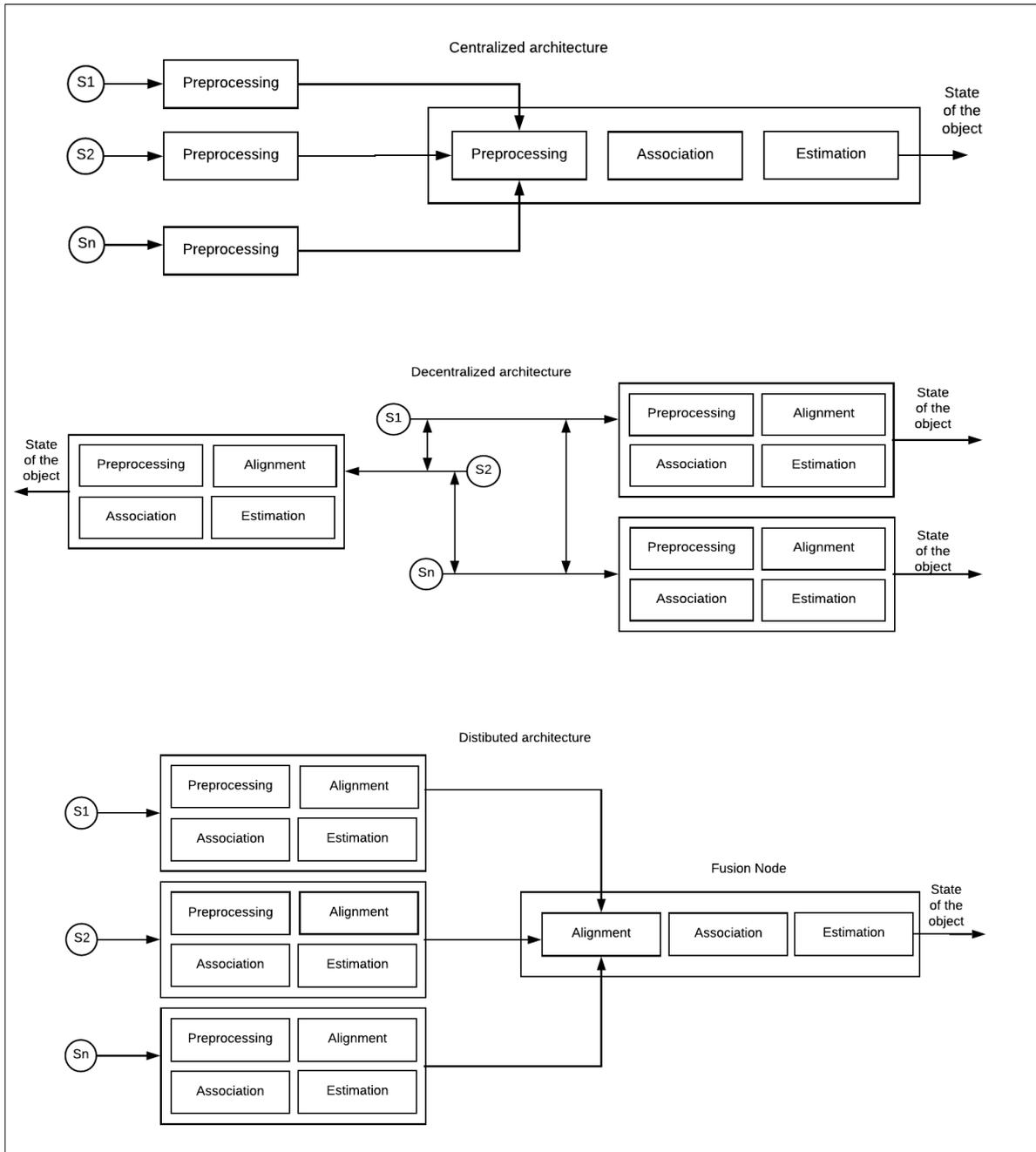


Figure 2.5: Classification based on the type of architecture [131]

2.5.4 Data Fusion Applications

Recently, there have been many applications developed for applied data fusion to produce better results and decisions. Data may be produced from sensors or other sources. A combination of different sources of data will create better results or decisions. Data fusion is

one way to combine data from many sources and provide decision or solution. Examples of applications that have applied data fusion include human detection, networking, transportation, medical and fraud system. These are some examples of successful applications applying data fusion.

a) Human Presence Detection

Human presence detection is one of the challenging research areas in the smart environment. Human presence detection can help in improving the security system such as detecting if there is an intruder in the property. Sonia et al. [133] performed research on human presence detection and applying sensor fusion techniques. This research combines passive infrared (PIR) sensor and ultrasonic sensors using voting based approach. This approach is used to classify signals received from human or non-human objects; therefore, it can identify human presence. The results of the experiment show that voting based sensor approaches perform better than using only individual sensors. Fotiadis [134] proposed research on human detection systems in autonomous surveillance that can be employed on a mobile robotic platform. They applied two detection modules, one for laser and another for vision data and fused them. Vision data is a combination of the histogram of oriented gradients descriptor and the linear support vector machine classifier. After the experiments, the results accurately detected human presence using the robotic mobile platform. The mobile robotic platform received information from both laser and vision sensors, improving performance over single sensor detection. From these two applications, it shows that having more sensors will give better results.

b) Medical imaging

Medical imaging refers to capturing images of the body for clinical and medical use. Examples of medical imaging include ultrasound image, x-ray image, radiography image and magnetic resonance imaging (MRI). Image modalities can be multiplied to improve the quality of imaging. Medical fusion imaging can be applied to get a better medical or clinical result. Adali et al. [135] used medical imaging from three modalities, namely functional magnetic resonance imaging (fMRI), structural MRI (sMRI), and electroencephalography (EEG). They are applying joint independent component analysis (jICA) and the transposed independent vector analysis (tIVA) models to enable fusion of data from the three image modalities. The results show that jICA produces a more desirable solution within the limited number of subjects. Singh and Khare [136] proposed a multimodal medical image fusion using Daubechies complex wavelet transform (DCxWT). They also compared the proposed fusion method with wavelet domain (Dual tree complex wavelet transform (DTCWT), Lifting wavelet transform (LWT), Multiwavelet transform (MWT), Stationary wavelet transform (SWT), spatial domain (Principal component analysis (PCA), linear and sharp), Contourlet transform (CT), and Nonsubsampled contourlet transform (NSCT) image fusion methods. The results show that the proposed method DCxWT produce better performance than another fusion method. These two examples of medical image fusion show that fusion techniques give significant results and can help in making decisions using multimodalities images.

a) Wireless Sensor Networks

Wireless sensor networks (WSNs) consist of a large number of sensor nodes for collecting data, processing and communicate with the sensors. WSNs have been applied to various applications such as target tracking and event detection, and the fundamental issues are how

the applications can increase confidence in the sensor measurements and minimize energy consumption [137]. Quality of service (QoS) is an important issue in data accuracy, delay in data aggregations and maximisation the lifetime network. Izadi [137] proposed research about the fuzzy-based approach for WSN to increase the QoS and reduce energy consumption. The proposed data fusion approach is compared with two baseline approaches in term of data collection, the data packet and energy consumption. Their research achieves better results compared to baseline approaches. Luo et al. [138] studied a prediction-based data sensing and fusion technique in WSN to reduce data transmission and maintain the coverage of sensor levels. The proposed technique is grey model – kernel recursive least squares (GM-KRLS), and Blowfish algorithm (BA). GM will predict data of the next period with a small number of the data item and KRLS will predict the value approximate the true value with high accuracy during data sensing and fusion. This proposed technique achieves high accuracy prediction, good scalability, and low communication and confidentiality. Besides the two examples of applications, there are much more research applying data fusion and received better results. Therefore, it can be concluded that applying data fusion in multisensor offers many benefits compared to a single sensor.

2.6 Summary

This chapter presented an overview of digital forensics, biometrics, heterogeneous data, anomaly detection and data fusion which related to this study. Applications of digital forensics and biometrics in machine learning and computational intelligence are also discussed in this chapter. Then, heterogeneous data is defined and presented. After that, the important part of this study which is anomaly detection is discussed. The fundamental theory, types of anomaly which is a global and local anomaly, the category of anomaly detection and

applications of anomaly detection also presented. There are disadvantages of Chebyshev inequality and threshold-based anomaly which later this study proposed the recent technique to detect an anomaly. The last section of this chapter explained on data fusion techniques, algorithms, applications and challenging in data fusion. Some challenges in data fusion are overcome in this study using the proposed data fusion technique. Next chapter will discuss the research methodology of this study.

Research Methodology

3.1 Challenges

Human behaviour is both physically and emotionally based and changes as people get older. These changes mean that it is not easy to trace or track human behaviour. Recently, with the advent of the digital era, many devices have become available which can be used to monitor humans, but it is not 100% guaranteed. Due to the ubiquitous nature of digital applications, people leave many digital traces without their notice. For example, if people use mobile applications which tag their location, for instance in social media applications, then anybody can detect and trace them. In another example, when you sign up your profile in online shopping websites such as Amazon or eBay, the company has data about your preferences when you continue shopping through the website. These examples show that any digital activity can be used to track behaviour.

From one source, human behaviour cannot be simply detected, such as when using a shopping website or location tag. Other data have to be considered and combined with all the data sources, such as whom one called or perhaps what one spent. This is known as heterogeneous data. Heterogeneous data can give more confidence in detecting human behaviour because it is derived from multiple sources of data. In a forensic investigation, heterogeneous data is crucial to producing more accurate results and decisions. Digital

forensic data is heterogeneous and may be structured, unstructured, or semi-structured [19]. Processing heterogeneous data is not as easy as single data points, especially during a forensic investigation. The volume of digital forensic data is increasing and becoming a big data issue [31]. Big data includes telephone calls, trajectory data, email and text messaging. Anomaly detection can be one way to simplify investigating and can detect abnormal behaviour in a data set.

Additionally, image data is also important to recognise humans. Image data can be acquired from photos or videos. Based on the forensic situation, surveillance camera significant in producing any video data and may be recorded the crime scene. From this video data, image data can be retrieved and maybe can give an answer or hint to find the suspect. However, it is not easy to look at the image and determine identity. From an image, it is easy to identify gender with the naked eye, but not age. Age is difficult to guess. However, age can be categorised into groups, for example, baby, youth, and adult. Thus, a combination of these two data points can help detecting age and gender in finding the identity of the suspect or at least shorten the time taken to locate the suspect.

Heterogeneous data is significant when it can be combined or integrated to produce information or decisions. It is difficult to combine different data types. According to Kaisler [32], the integration of various databases is one of the big data issues. However, data fusion is a one of the techniques to combine the data. There are many research using data fusion in multi-sensor environment [13], [128], [131]. Nevertheless, data fusion is always applied in applications such as remote sensing and meteorological monitoring in environmental studies, biomedical and health (medical diagnosis and smart patient monitoring), and multisensory systems (audio-visual interaction and human-machine interaction) [13]. In real situations, heterogeneous data consists of many types of data. For example, when a crime takes place, data can be derived from the crime scene, or from personal data such as email, phone calls or

bank accounts. When the investigators want to make a decision or solution to the case, all the data must be linked or integrated. Data fusion is one method to combine heterogeneous data.

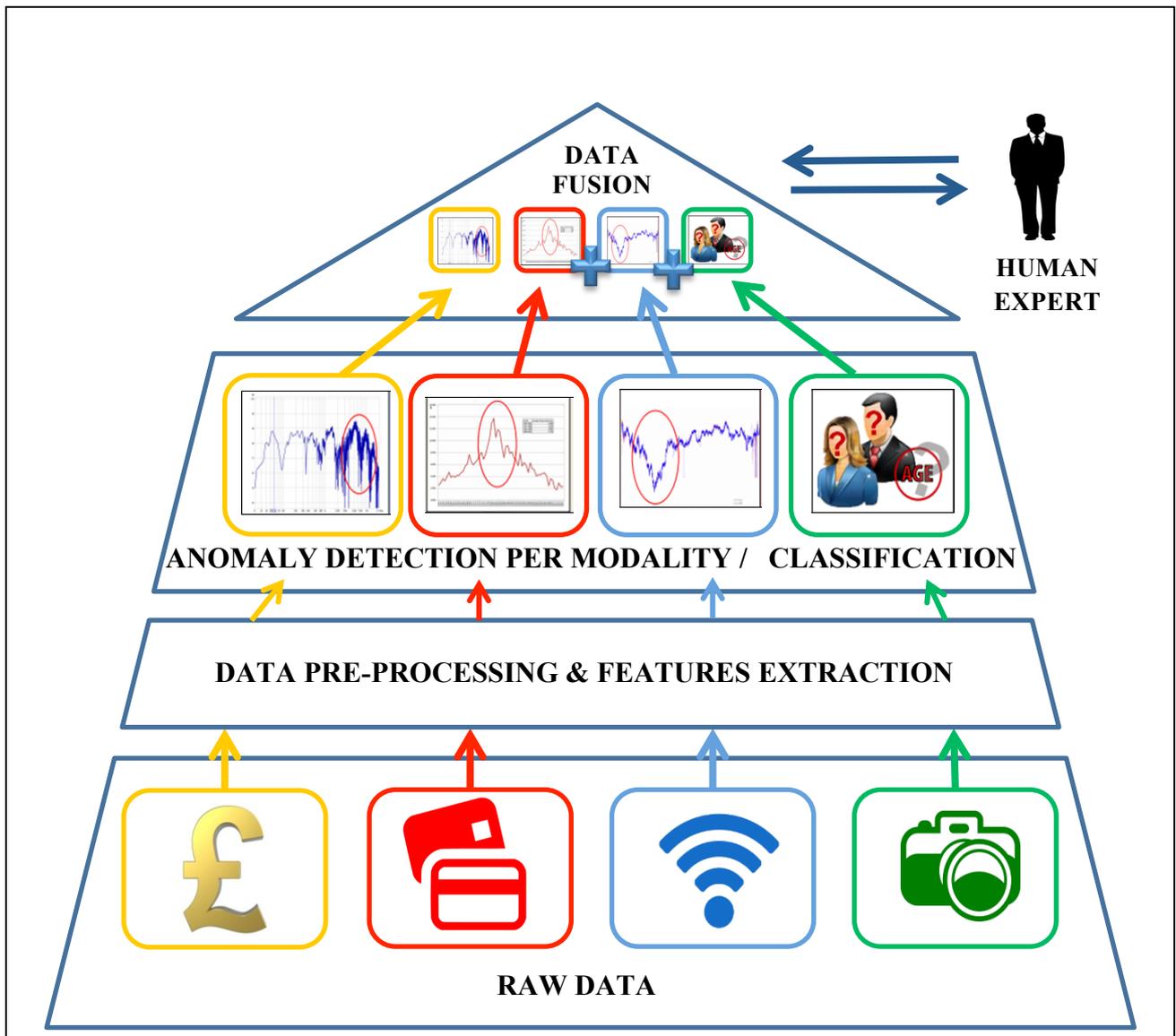


Figure 3.1: Research Framework

3.2 Proposed Approach

This study proposes using heterogeneous data for detecting anomalies. It combines the information/indication about the degree of anomaly of the data with a new data fusion technique. Figure 3.1 shows the research framework starting with raw data, data pre-

processing, anomaly detection and, finally, data fusion. It demonstrates the use of heterogeneous data which include three types of data:

- a) signals (GPS data)
- b) financial (credit card and loyalty card data)
- c) image data

Heterogeneous data may consist of various data types depending on the application. Data can be numerical, text, signal, or image. For instance, in medical applications, there are signal data such as from ECG, image data from X-ray or ultrasound, report from the blood test and urine test. All the data has different values, types and dimensions. If there is a patient with all that data, the medical doctor may need to combine it to make a decision or prediction of what disease the patient has and the appropriate treatment. In another case, such as a terrorist attacks the investigators may obtain evidence from CCTV (video and image data), a telephone call made by the suspect (signal), and emails or text messages (text data). This shows that most applications or events produce heterogeneous data. The heterogeneous data have to be combined to solve problems. Heterogeneous data can give more confidence in making decisions when combined. The top level of this approach is combining the data using new data fusion technique. In the proposed approach, raw heterogeneous data is at the low level. Each form of data needs to be identified and understood before proceeding to the next level.

In the next level, data pre-processing and feature extraction are applied. Every data type has different features and different ways to pre-process. If the data type is numerical, then the data may have an integer, double or float value. For example, financial data always includes numerical data. The numerical data can be normalised to make it comparable. Numerical data can be normalised between 0 and 1 or between -1 and 1. Numerical data is easier to process because it is already in a form which a computer can understand and process. Other data

types such as images, text or signal have to be transformed into numerical data first. However, the pre-processing technique is different and depends on the data type. If the data is image or signal, then image or signal processing technique is required to pre-process the data. There are many types of image pre-processing techniques, including image segmentation, image transformation, image rotation or scaling and image degradation. Techniques such as segmentation have also many ways to be applied. Examples of image segmentation include thresholding, clustering, compression-based method, histogram-based method and edge detection. Signals include different pre-processing techniques such as filtering, discrete transform, data compression and Fourier analysis. Image processing has two dimensions of data while signal processing is only one dimension. Nowadays, image and signal data sources are in a digital form. Therefore, it is easier to pre-process and make the computer process the data. However, text data is different from numerical, image or signal data. Text data requires natural language processing (NLP) techniques to be processed. The usual ways of pre-processing text are tokenisation or text normalisation. The tokenization process consists of removing the stop words (e.g., “the”, “a”, “of”, “and”); discarding unwanted stuff (e.g., HTML tags); stemming (e.g., the words “look”, “looked” and “looking” are from the same stem); word boundaries (e.g.: white space and punctuation); and capitalization (e.g.: “Us” and “us”). There are more steps to perform before all the text can be converted into numerical data. Natural language processing can be simple, such as using a bag of word model where a text (e.g., sentence or a document) is represented as a bag of its words.

The next level is *anomaly detection*. Anomaly detection is one of the most important methods to find and process abnormal data, as this method can distinguish between normal and abnormal behaviour. Abnormal behaviour is significant in some applications such as fraud detection, intrusion detection system, medical application, surveillance system and forensic application. In this proposed method, anomaly detection is implemented using a new

data analysis technique, EDA, which considers the standardised eccentricity and is based on unsupervised techniques. This technique is a foundation of the data analytics for statistic and streaming data analysis. Instead of analysing data using classical probability theory and analytics, EDA can be applied, respectively. EDA is more sensitive and flexible when detecting anomalies as compared to the traditional probability approach because it does take into account explicitly all the data (not only the mean). An anomaly can be detected without making *a priori* assumption about the data distribution. In real-world applications such as engineering, natural science, biomedical and human behaviour modelling, distributions of data vary greatly and are hard to define. These are collected real-world data and not *prior* theoretical assumptions. Therefore, EDA is a method that can help detecting an anomaly in any distribution of data and require no *prior* assumption. In the study presented in this thesis, there are four types of data: financial data (credit and loyalty card data), signal data (GPS data) and image data. However, anomaly detection is only applied to financial data (credit and loyalty card data) and signal data (GPS), while image data is processed using deep learning techniques. The proposed technique applies face image data and classifies images regarding gender and age of the person. Financial and signal data differ from image data. As discussed before, the pre-processing method for image data, in this case, is only classified then the classification result can be utilised in the next stages.

The top level in this proposed approach is the combination of all the data types using the new approach of data fusion. At the lower level, there are four types of data presented in this proposed approach. All of the data have different sizes and dimensions. Therefore, the new data fusion approach is introduced at this level. There are few steps to be applied in the data fusion level. The first step is normalisation, in which all the data has to be transformed into a value between 0 and 1 to make the data comparable. Then, all the values from each data modality are summed together. After that, the total value is divided by the number of the data

modalities. Later, all the result is sorted in descending order. Therefore the highest value will be the most suspicious data item, and the lowest value may be the least suspicious data item. The method cannot simply pre-define suspicious and non-suspicious, but a human expert can distinguish the suspicious and non-suspicious results. Next, the results of using the data can be analysed more efficiently, since the amount of data is significantly smaller, more organised, and represent human-intelligible information in the form of graphs, tables, and visualisations. The new data fusion techniques can also represent expert knowledge, which can later be used to formulate the sequence of events. The proposed approach is not intended (and hardly allows) a fully automatic system that resolves investigations. The aim is rather to simplify the role of the humans and to facilitate their focus on a small number of cases to be looked at in more detail. A human expert will need to verify the analysis in the final step and produce a final decision based on highly intelligible information, which is much smaller amount than the raw data. The greatest significance of the new data fusion approach is in assisting human experts to reach the correct decisions with a minimum time while using a much larger amount of more complex heterogeneous data as a starting point.

3.3 Summary

This chapter presented the research methodology of this study. Human behaviour and heterogeneous data have been discussed in this chapter. Then, the proposed research framework is presented and explained. The framework has four levels beginning with the raw data. Then, the second level is anomaly detection which is applied to three different modalities of data. Image data is used to classify the gender and age of the suspect. After that, at the top level, all data after finding the possible anomalies and classification of images

are combined using a new data fusion technique. Then, the top level will produce results from data fusion which later can help or assist in decision making. The next chapter will discuss anomaly detection, image classification and data fusion.

Detecting Anomalous Behaviour

4.1 Introduction

Daily routines encompass most human behaviours. For example, consider starting in the morning when we start going to work by car, then clocking in at the office, having lunch at a restaurant, paying using debit or credit card, clocking out after finish work, and maybe going to buy food before going home. This is a normal behaviour for a person who is working office hours. This behaviour may be different if the individual is not working office hours, such as police, firefighters or doctors who have different working hours every week or may have to respond to a call. These patterns of working can produce data about the normal human behaviour which depend on their job. Data can also be generated from mobile applications such as WhatsApp and Telegram. All the information about communication applications can be processed to produce a pattern of this communication. Another example is the travel pattern. Travel data can be derived from GPS information to show travel patterns. All of this pattern data provides information about the normal human behaviour. Detecting human

behaviour is important, especially for the investigators of crime. Forensic investigators have to investigate all the possible evidence to find the suspect. Data can be collected from surveillance cameras, social media accounts, telephone calls, and credit card purchases. Surveillance cameras produce video and images; social media data may produce text, images and video; a phone call produces signal data, and credit card purchases exemplify financial data. In reality, most of the gathered data produce a pattern of normal human behaviour. However, if there is something suspicious in the data, then anomalous human behaviour can be identified automatically. Different anomaly detection methods can be applied to identify the anomalous human behaviour.

Anomaly detection is one of the most important methods to analyse the data streams [4]. In statistics, anomalies or outliers are removed to get a better result of the analysis [120]. On the other hand, in some applications, anomalies and outliers are very important. They can be detected locally (i.e., in part of the data) or globally (in all data). Anomaly data can help solving problems and giving better solutions for specific problems such as in the medical sector [5], fraud detection in finance [6], fault detection in machines [7], intrusion detection in networks [8], surveillance systems for security [9] as well as forensic investigations [6], [10].

4.2 Anomaly Detection

Traditionally, anomaly detection is addressed using statistical methods. The frequentistic technique represents the probabilities of random signals, and the *prior* assumption has to be made [4]. The main decision is traditionally made using threshold values. These thresholds are based on the normal distribution of random variables (usually assuming Gaussians) (see figure 4.1) while for arbitrary distributions (see figure 4.2), they are based on the well-known Chebyshev inequality [4]. However, in general, we do not know the type of the distribution

beforehand. Moreover, for a complex problem such as human behaviour modelling, data distribution may be difficult to define. For such general cases, Chebyshev theorem can be used. According to it [4], no more than $1/n^2$ data are abnormal. For example, people often use 3σ , which guarantees no more than $1/9$ (or $\sim 11\%$), or 6σ which guarantees no more than $1/36$ (or $\sim 3\%$), respectively, to be anomalous data. An example of dataset which contains the temperature in Manchester over three years (from 2010 to 2012) is visualised using a histogram in figure 4.3.

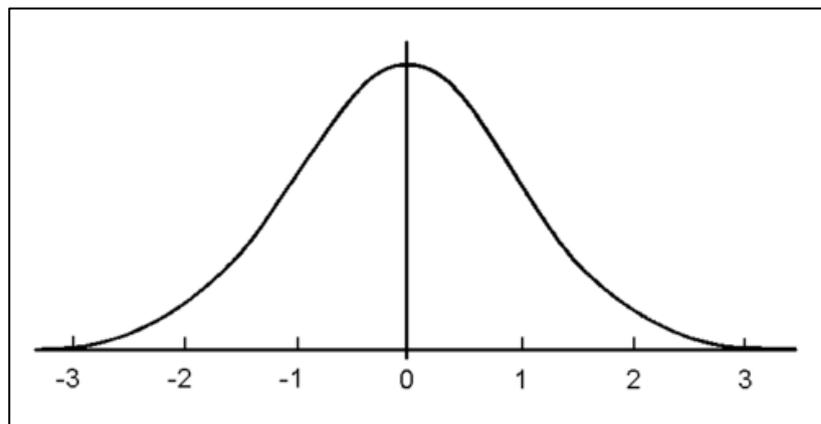


Figure 4.1: Normal Distribution

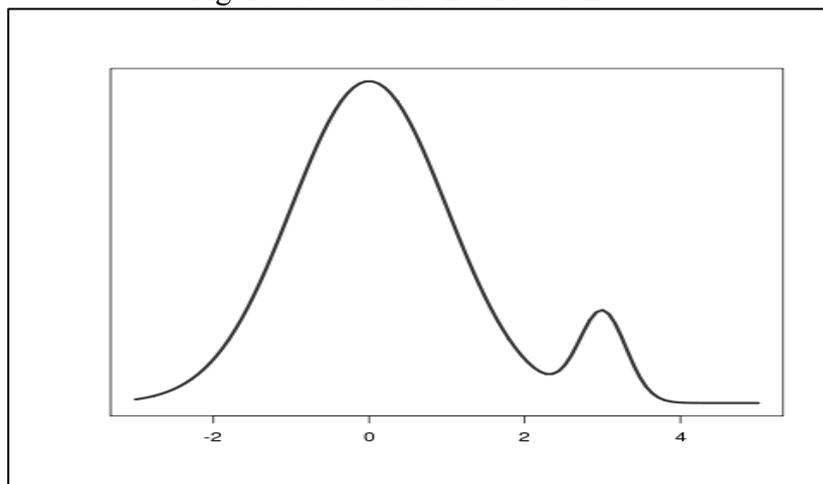


Figure 4.2: Arbitrary Distribution

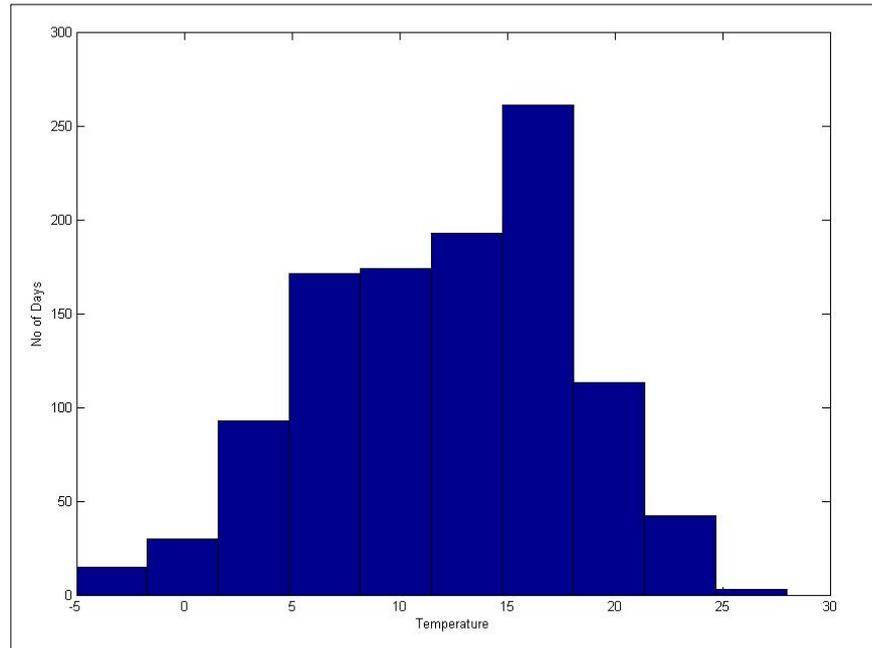


Figure 4.3: Histogram plot of temperature in Manchester from 2010 to 2012

In this example (see figure 4.3), an anomaly can be detected in the summer with a temperature between 25 and 30 degrees and temperature below -2 degree celsius in the winter. Then, the data is plotted for 2010, 2011 and 2012 using different histogram plots (see figure 4.4, 4.5 and 4.6 respectively). The three figures show that in 2010, the temperature is low at below -2 degrees and the highest temperature is above 25 degrees. This data can be considered anomalous. However, in 2011 and 2012, the lowest temperature is between -1 to 0 degree and the highest is around 20 to 24 degrees.

Then, the same data from 2010 to 2012 is plotted again using two-dimensional data (see figure 4.7). In this figure, it is easier to spot an anomaly in temperatures and days. It also gives the same results as histogram plot but is easier to see the exact time instance. When the range of normal data is set between -3 to 3 after the standardisation, no outliers or abnormal data were found in this data. However, if the range of standard data is set between -2σ and 2σ , then 36 outliers are found in which the lowest data is -5 degrees celsius and the highest is 28 degrees celsius. Figures 4.8, 4.9, and 4.10 show different temperatures and days on

different years, which are 2010, 2011, and 2012 respectively. There are nine outliers found in 2010, when the standard deviation is set between -2σ and 2σ , four outliers found in 2011, and 15 outliers in 2012.

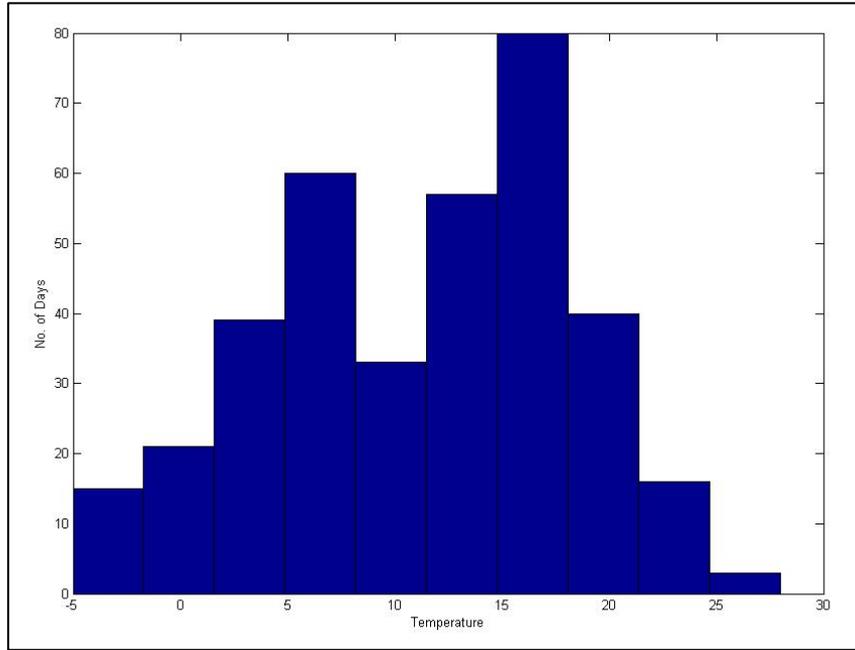


Figure 4.4: Histogram plot of temperature in Manchester in 2010

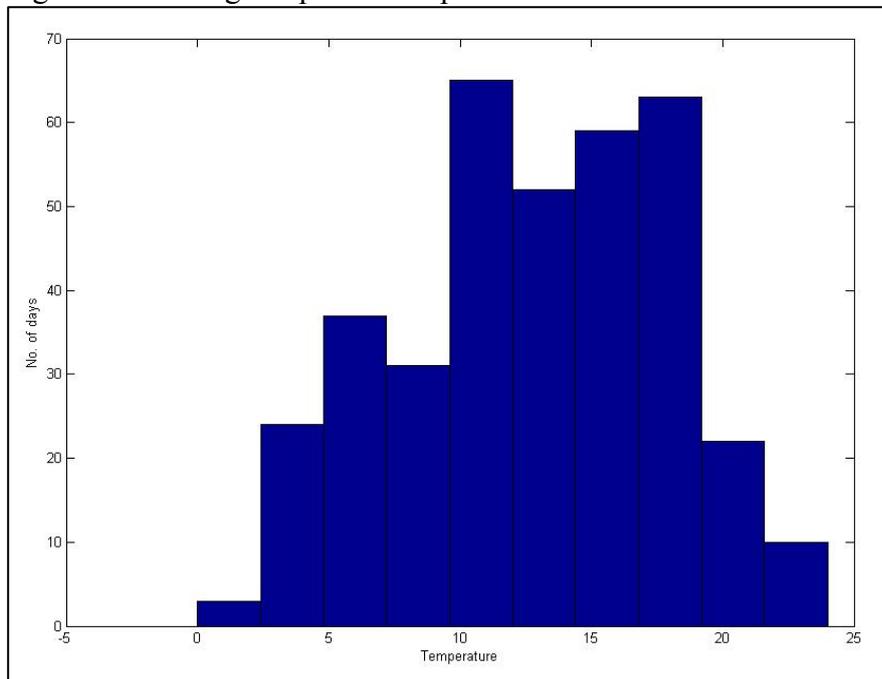


Figure 4.5: Histogram plot of temperature in Manchester in 2011

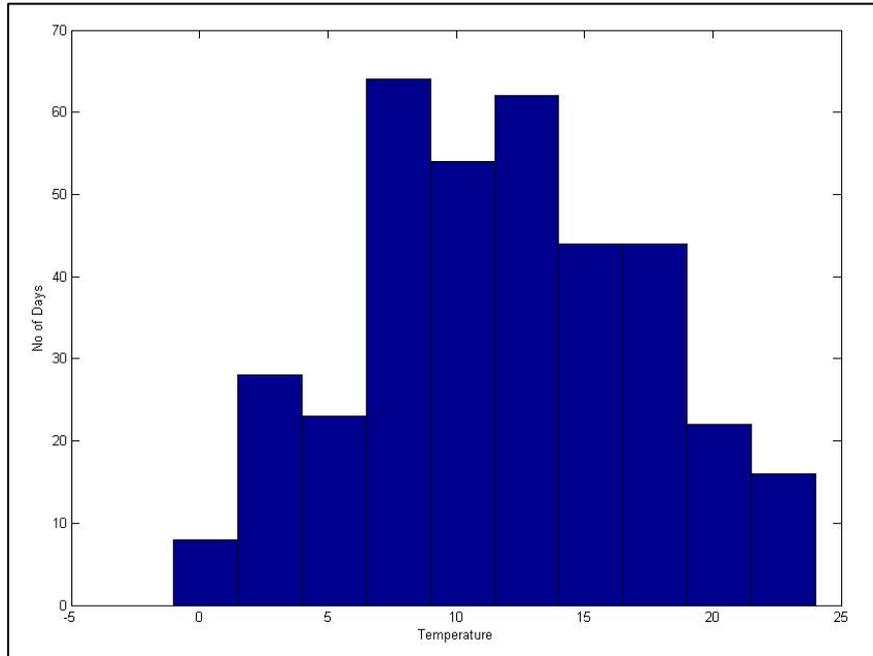


Figure 4.6 Histogram plot of temperature in Manchester in 2012

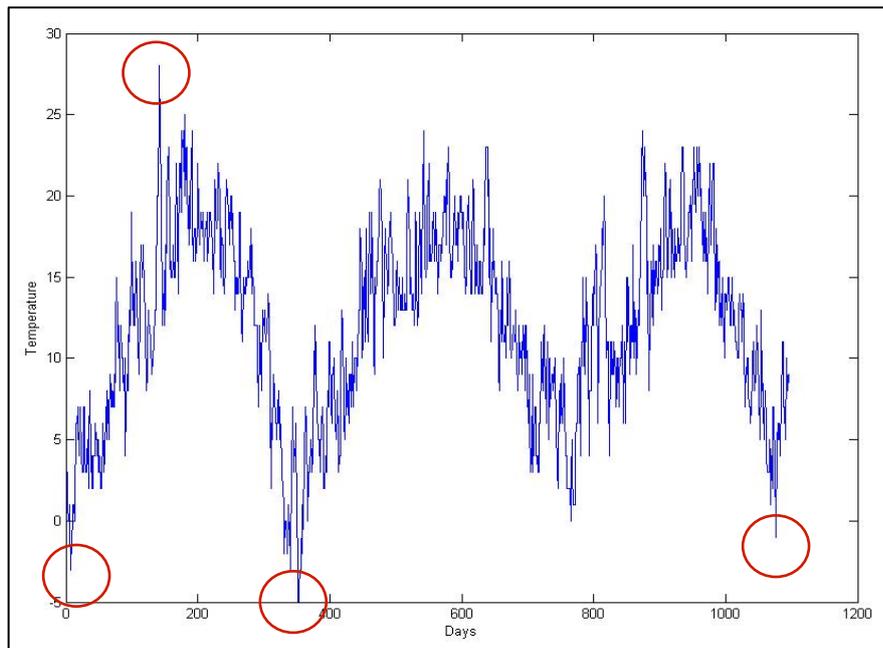


Figure 4.7: Plotting the temperatures and days from 2010 to 2012

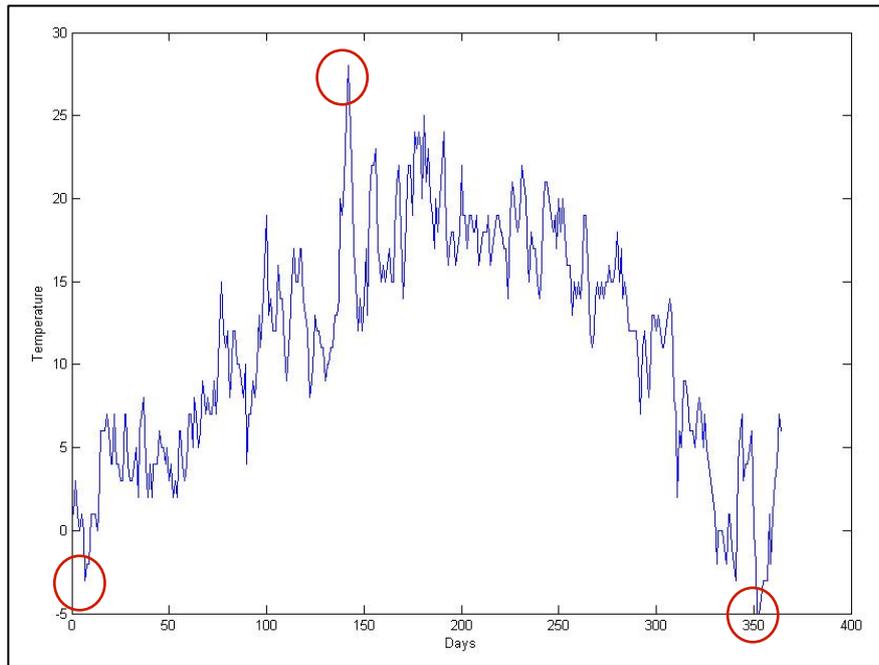


Figure 4.8: Plotting the temperatures and days in 2010

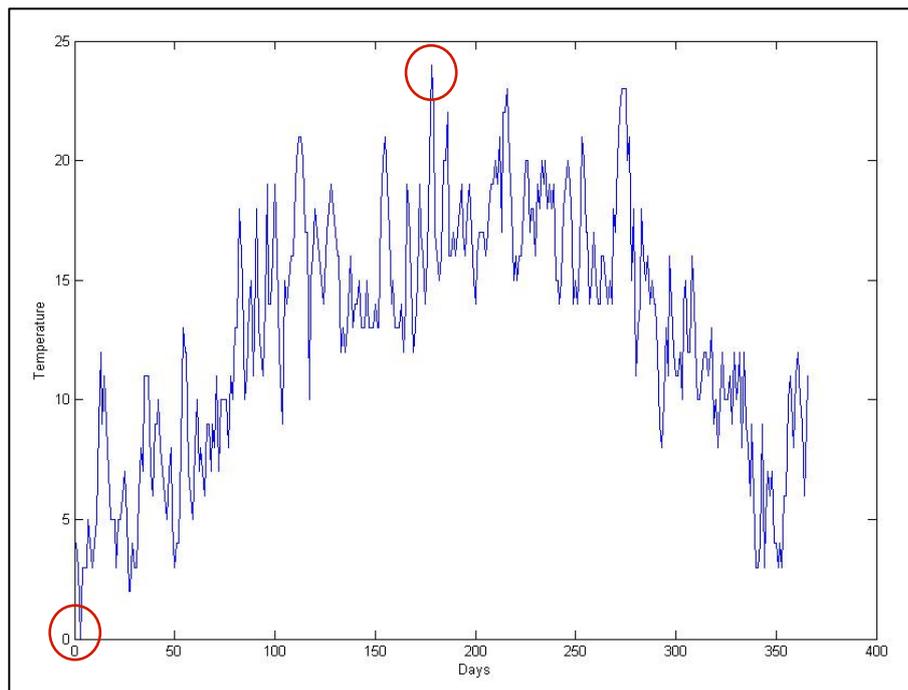


Figure 4.9: Plotting the temperatures and days in 2011

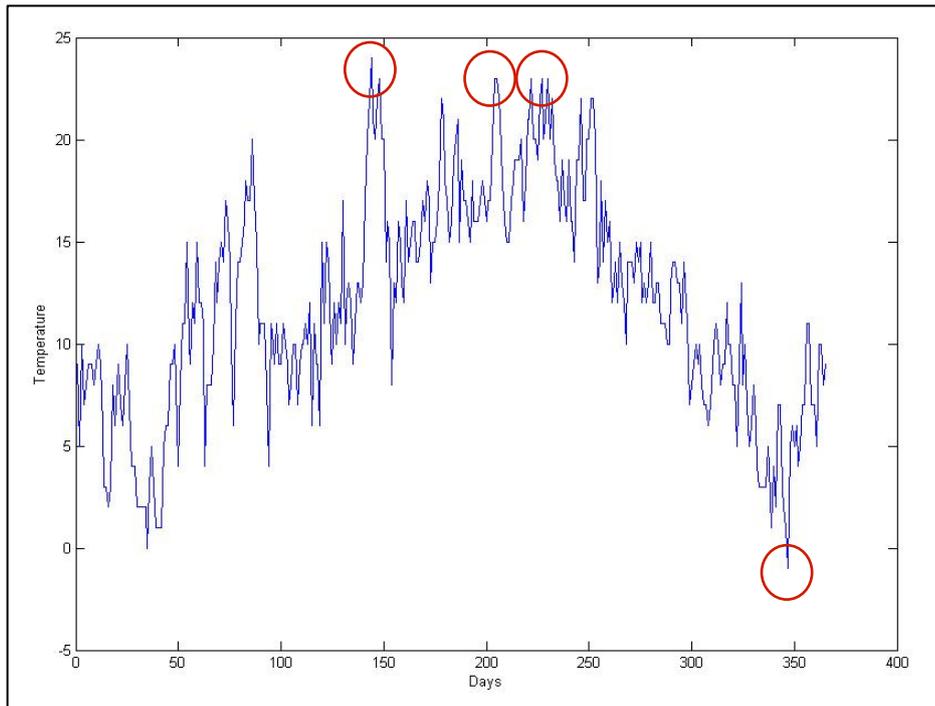


Figure 4.10: Plotting the temperatures and days in 2012

These approaches have the following disadvantages [4]:

- a) they require strict *prior* assumptions – in this example, a range between -2σ and 2σ is needed to find the anomaly.
- b) they relax the conditions too much to avoid false positives to the level where it misses many true positives (the 3σ rule sometimes fails to detect some obvious outliers) – if the example uses 3σ , then the outliers cannot be found, and everything is represented as normal.
- c) Many data samples are required – in this study we used 1095 data vectors.
- d) a single data sample is compared with the average, instead of comparing pairs of data samples; therefore, the information is blurred and is no longer point-wise and local.

Angelov [46], [4] and Angelov et al. [139] introduced a new method to avoid the disadvantages of the traditional statistical method. The new method called Recursive Density Estimation was introduced in 2012 and is based on data density calculated recursively. Later,

this method was improved with a new concept, namely the “eccentricity”. The newly introduced “eccentricity” is part of so called Empirical Data Analytics [141]. This approach does not require any prior assumption, and a σ_{gap} can be formulated between the “eccentricities” of the data samples with larger “eccentricity” [4].

4.2.1 Recursive Density Estimation (RDE)

RDE was designed based on Cauchy type Kernel function [142] and was initially introduced by Angelov and Buswell [143]. RDE was first described with the name RDE in 2008 [50]. Typically, many research uses Gaussians calculated offline once to detect an anomaly, but RDE has calculated density and updated it recursively. Only very small data sets and the means of all data samples are calculated in time, which must be stored in the memory and updated. The data sample x_k is also used, but there is no need to store or update it. D_k is the local density of the area k and N_k is the number of x_k linked to the area k [46]:

$$D_k(x_k) = \frac{1}{1 + \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{1}{|x_k - x_i|^2}} \quad (4.1)$$

$$D_k(x_k) = \frac{1}{1 + \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{1}{|x_k - \mu_k|^2}} \quad (4.2)$$

Where the mean μ_k and the scalar product \sum_k can be updated recursively as follows:

$$\mu_k = \frac{k-1}{k} \mu_{k-1} + \frac{1}{k} x_k \quad \mu_1 = x_1 \quad (4.3)$$

$$\sum_k = \frac{k-1}{k} \sum_{k-1} + \frac{1}{k} P x_k P^2 \quad \sum_1 = P x_1 P^2 \quad (4.4)$$

According to [46], an outlier or anomaly cannot be defined by any strict mathematical formula, but Chebyshev's theorem is widely used with 3σ as a threshold. 3σ is used because the probability of data sample to be abnormal is $<0.3\%$ based on Gaussian distribution ($>99.7\%$ of the data sample is normal if using 3σ [4]). Based on density $D_k(x_k)$, the mean density \bar{D}_k is calculated as follows [46]:

$$\bar{D}_k = \frac{k-1}{k} \bar{D}_{k-1} + \frac{1}{k} D_k \quad \bar{D}_1 = D_1 \quad (4.5)$$

The variance of the density can be updated recursively by [46]:

$$\left(\sigma_k^D\right)^2 = \frac{k-1}{k} \left(\sigma_{k-1}^D\right)^2 + \frac{1}{k} \left(D_k - \bar{D}_k\right)^2 \quad \left(\sigma_k^D\right)^2 = 0 \quad (4.6)$$

From equations (4.5) and (4.6), abnormal data can be easily detected and identified using the standard deviation, σ_k^D from the mean density, \bar{D}_k [46].

4.2.2 Empirical Data Analytics (EDA)

EDA is a new theoretical framework to analyse data and detect anomalies without prior assumption [140], [141]. EDA touches the very foundations of data analytics for statistic and streaming data. It serves as an alternative to the classical probability theory and analytics

[140], [141]. Applications of EDA include, but are not limited to, data analysis, clustering, classification, prediction and anomaly detection. The advantages of EDA are [19]:

- a) It is entirely driven by the empirically observed data and based on their mutual distributions.
- b) Free from *prior* assumption and predefined parameters
- c) No need to assume any model of data generation (Gaussian or other)
- d) No need to assume the observed data to be independent or identically distributed
- e) No requirement for an unlimited number of observations (it can work with only two data samples)
- f) Free from certain well-known paradoxes of the traditional probability theory
- g) Quantities can be calculated recursively for many types of distance metrics

In regards to anomalies detection, EDA has proven to be more sensitive and flexible than the traditional approach [140], [141]. Also, it is essential to analyse local anomalies rather than global ones [4]. The three EDA factors employed in this chapter are [140]:

- a) standardised eccentricity, ε
- b) cumulative proximity, π
- c) data density, D

Consider the data point denoted as x_k . “Standardised eccentricity, ε ” is a measure of the ensemble property related to the tail of the distribution. It is sensitive to abnormal data and thus plays an important role in anomaly detection. It can be defined as [140], [141]:

$$\varepsilon_k(x_k) = \frac{2k\pi_k(x_k)}{\sum_{j=1}^N \pi_k(x_j)} \quad (4.6)$$

where π denotes cumulative proximity, π from a particular, j^{th} , ($j \geq 1$) data point. Cumulative proximity is a square of the summed distance between all data points. This can be defined as [140], [141]:

$$\pi_k(x_k) = \sum_{j=1}^k d^2(x, x_j) \quad (4.7)$$

These quantities (π and ε) can be defined either locally (for a part of) or globally (for all data points) and calculated recursively for a particular type of distance [141].

$$\varepsilon_k(x_k) = 1 + \frac{P\mu_k - x_k}{X_k - P\mu_k} \frac{P^2}{P^2} \quad (4.8)$$

where μ is the recursively updated (local or global) mean; X is the recursively updated squared norm sum and the recursive update is made as in [141].

Data density is a measure of the ensemble property related to the centre of the distribution. Density can be defined as [140]:

$$D_k(x) = \frac{1}{\varepsilon_k(x)} \quad (4.9)$$

Further, in EDA a condition which provides the same result for the Chebyshev inequality without making any assumptions about the amount of data and their independence was introduced for Euclidean distance by [4], [141]:

$$P\left(\varepsilon_k(x) > n^2 + 1\right) \leq 1 - \frac{1}{n^2} \quad (4.10)$$

After finding the gap, the σ -gap principle is used to compare each of the data samples, with the aim to identify anomalies [4]:

$$\text{IF } (\varepsilon_k(x) > n^2 + 1) \text{ THEN } (x \text{ is an outlier}) \quad (4.11)$$

The significance of the proposed method is to assist human experts to reduce the time spent and reach the correct conclusion at the right moment in time. While at the same time allowing access to a massive amount of data. Such an approach can shorten the pre-processing phase and increase the efficiency of human experts.

4.2.3 Example of EDA

The same example which is the weather temperature in Manchester from 2010 to 2012 is applied using standardised eccentricity. In this example n is set to 2, therefore, if the eccentricity is more than 5, then the data is considered abnormal. Figure 4.11 shows that there is 19 anomalous data prints. Then, to see the local anomalies which occur every year, the data is separated into three years. Figure 4.12 shows the data in 2010, figure 4.13 shows the data in 2011, and figure 4.14 shows the data in 2012. In 2010, there are nine outliers found in this data while in 2011 there are only four outliers. In 2012, there are 15 outliers found in the data.

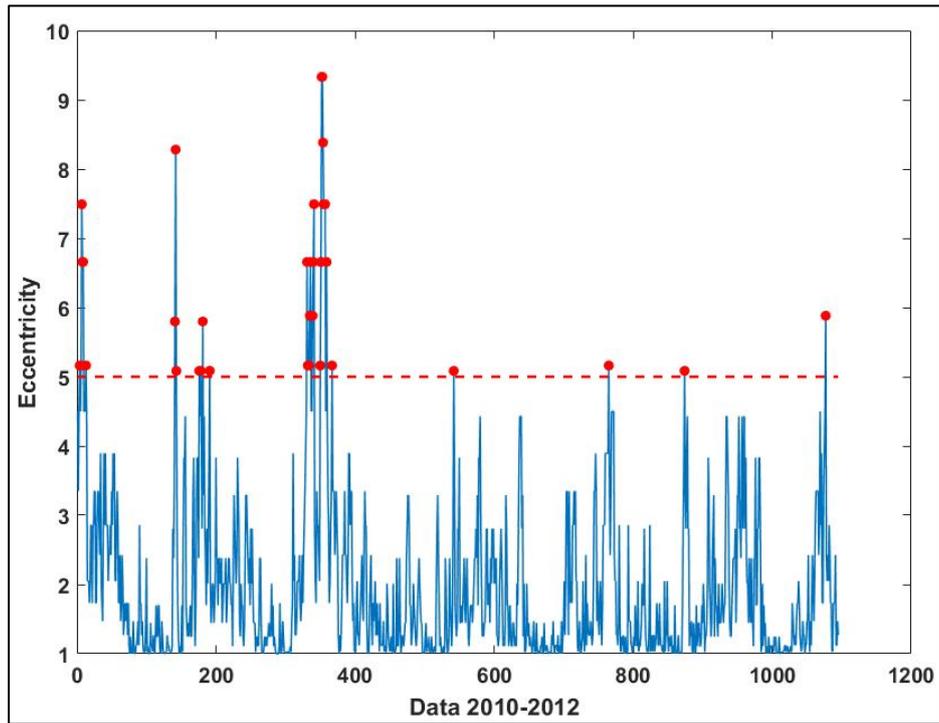


Figure 4.11: Eccentricity of temperature from 2010 to 2012

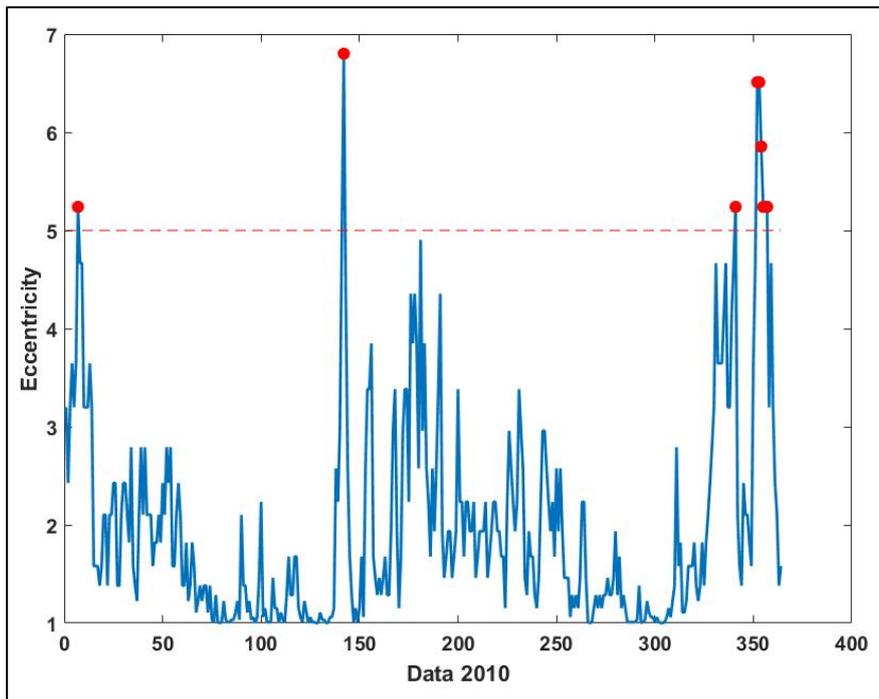


Figure 4.12: Eccentricity of temperature in 2010

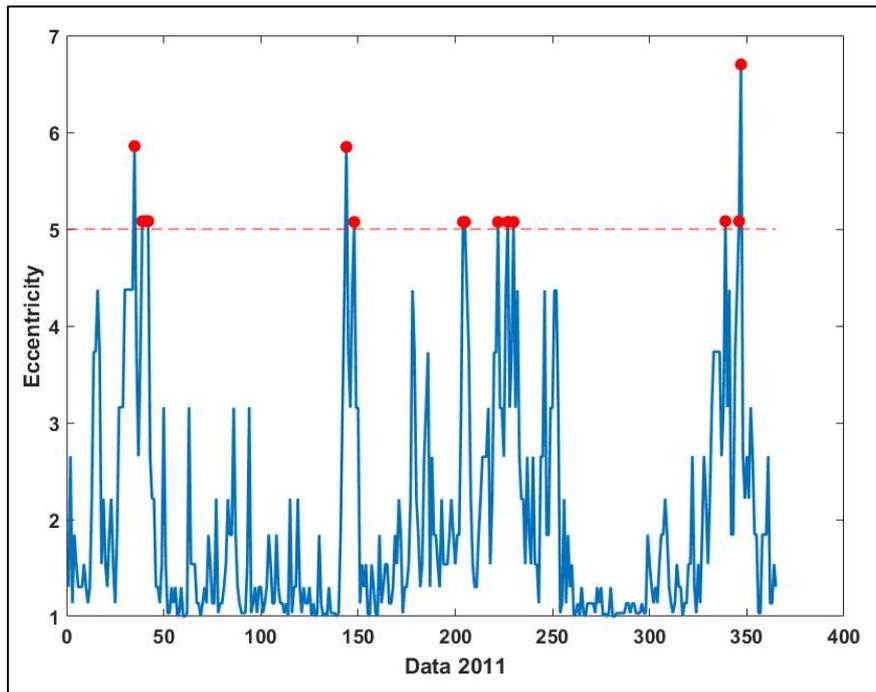


Figure 4.13: Eccentricity of temperature in 2011

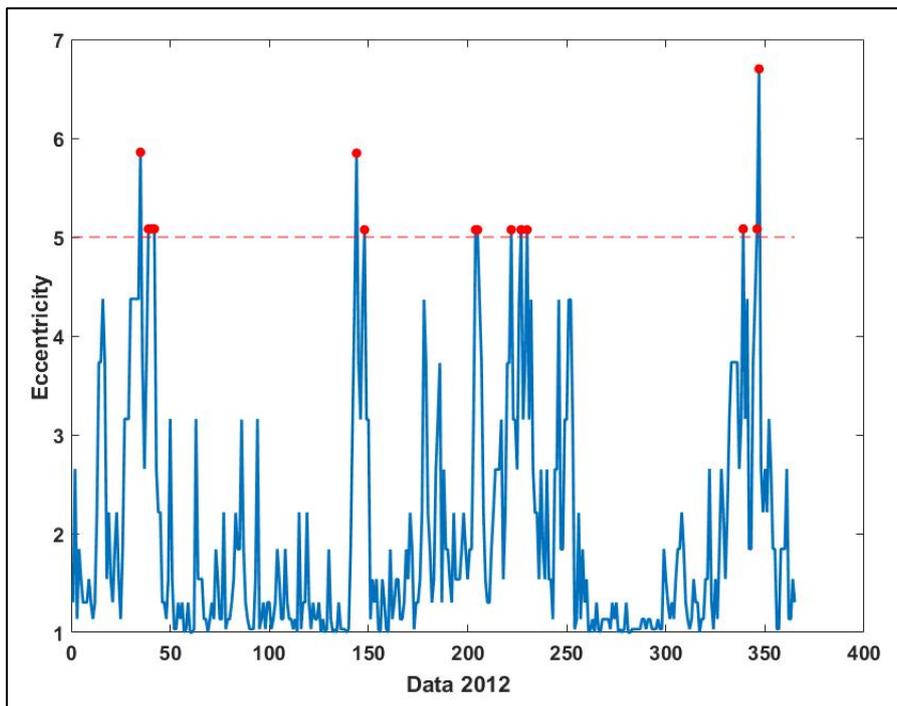


Figure 4.14: Eccentricity of temperature in 2012

4.3 Summary

This chapter discusses anomalous behaviour, which can occur in any situation or application. There are many applications in real-world situations which require applying anomaly detection to make a decision. Anomaly detection is a technique that can be applied to detect anomaly or outliers or abnormality in the dataset. The most popular technique is the statistical technique using a threshold value known as sigma. However, there are few disadvantages of this technique. Therefore, in this study, the new technique of EDA is applied. EDA has many advantages which can produce better results. It was demonstrated on an example of weather data with illustrate purposes.

Gender and Age Classification based on Images for Detecting Anomalous Behaviour

5.1 Introduction

This chapter introduces an approach to classify gender and age from images of human faces, which is an essential part of this research framework for autonomous detection of anomalous human behaviour. This chapter is a continuous study from the previous chapter on heterogeneous data in which face images are applied as supporting evidence. Feature extraction for the face images is done using a pre-trained deep learning model. Deep learning has been successfully applied to image classification area. Then, a Support Vector Machine is applied for the classification technique. One dataset named GAFace is built and used with the proposed method. It has to be stressed that in this dataset because few data sets with labels of gender and age exist for face images.

5.2 Gender and age classifications

Human behaviour is often uncertain and sometimes is affected by emotion or environment. Various human behaviours can be seen in public places because there are many people, therefore, there are many human behaviours. It is not easy to detect anomalous behaviour. However, if the normal behaviour is known or identified, then it is easy to compare and identify anomalous behaviours. Recently research on a surveillance systems in public places such as in train stations [115], football stadiums or concert events with crowded people [126] and other public places [144] has been increasing. When there are many people, it is not easy to detect the identity of the person especially getting access to the identity of every person in the world. Automatic detection can facilitate surveillance process in the airports, shopping malls or subway stations, helping to recognise human behaviour which later can assist in investigating suspicious events.

Gender and age classification is an important aspect of recognising a human. This research study is about “soft biometrics”. Soft biometrics include physical traits such as skin colour, hair colour, eye colour, height, weight, age and gender. Some face recognition systems can help in classifying gender and age, but not all are perfect, especially in recognising exact age. Even human eyes cannot perfectly guess human age because it is fuzzy. Some people have a baby face but are already adults. It is difficult to detect the exact age of this kind of person. Gender is also not so easy to classify by using face images compared to full body images.

5.3 Deep learning

Deep learning is an advanced machine learning method and a computational model that consists of multiple processing layers to produce learning representations of data with

multiple levels of abstraction. This model allows the computer to build complex concepts and improve and advance the image processing, speech recognition, and object recognition [145]. An example of deep learning model is the deep feedforward network, which is composed of many simpler mathematical functions and used to map a set of input to output values. Convolutional neural networks (CNN) is a type of deep learning based on feed forward network which is much easier to train and have better generalisation than networks with full connectivity between adjacent layers [145].

Recently, most data and information such as images, audios, and videos produced from the Internet applications such as Facebook, Instagram, WhatsApp and so on. The request for computational efficiency and high accuracy systems are increased. Due to some limitation of traditional machine learning techniques, transfer learning has become popular in the area of machine learning and computer vision. Transfer learning has changed the approach to how the machine is used to learn and perform classification tasks. Figure 5.1 shows the difference between traditional machine learning approaches and transfer learning approaches. The traditional machine learning approaches can only handle or process the same sample of training and testing. For example, the sample must have all face images, not full body images, etc. In contrast, transfer learning can handle and process mismatch data distribution through specific knowledge transfer methods. For instance, the training set can be images that share some knowledge of face images, e.g. eyes, nose, ears or irrelevant objects, e.g.: logo which has a shape similar to a face image. It has been successfully applied in a visual representation; image classification and object recognition [146]. Two approaches are employed in transfer learning [147]:

- a) preserving the original pre-trained network and updating the weights based on the new training dataset.
- b) pre-trained network in feature extraction, and representation followed by a

classifier such as SVM.

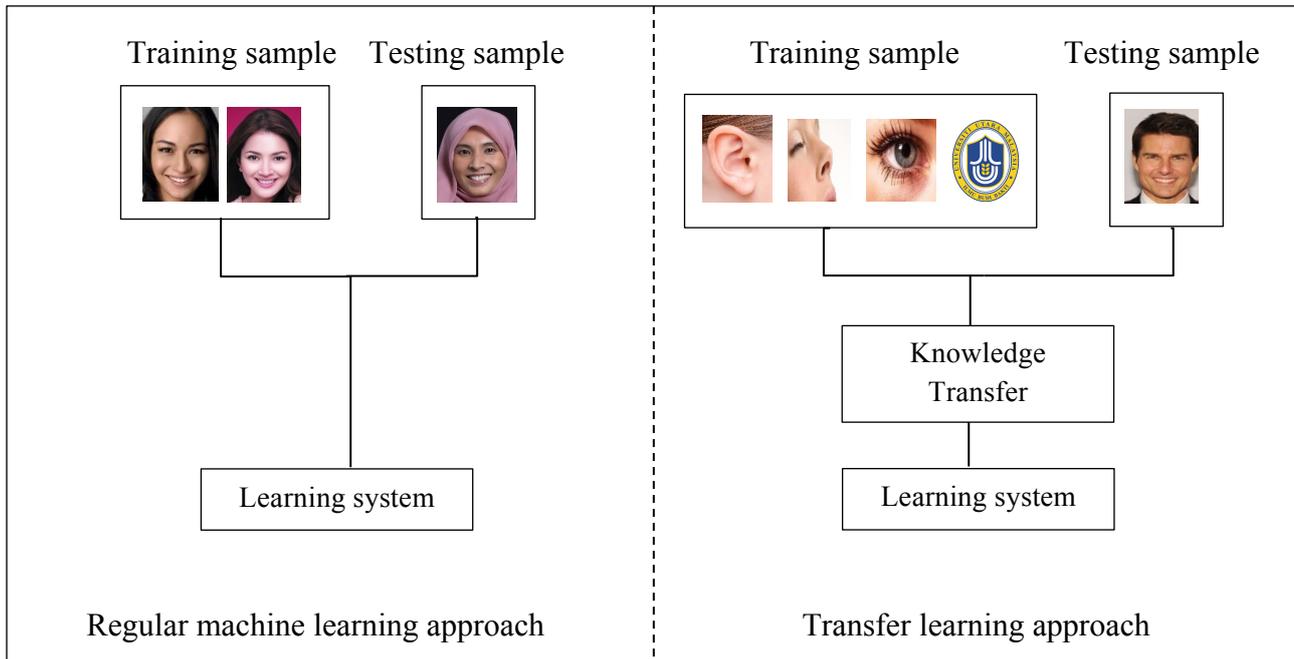


Figure 5.1: Difference between machine learning approaches and knowledge transfer approaches [Ling Shoa 2015]

In this study, transfer learning and deep convolutional neural network (CNN) are applied to extract features from images of possible suspects. A vast amount of data and high computational resources are required to train a new deep learning model “from scratch”. In some cases, the task is more challenging and requires training of a few days. A large dataset is used to train the CNN which can then be fine-tuned for a specific task even in a different domain. Transfer learning includes solving a new task by applying previously-learned knowledge. Transfer learning involves two major techniques:

- a) updating the weights on the new training dataset and keeping the original pre-trained network, and
- b) feature extraction and representation which will use pre-trained network then classify by a general classifier such as SVM [147].

CNN is likely to overfit with a small dataset. Therefore, transfer learning is suitable for

model training with a limited size of the dataset. Figure 5.2 shows an AlexNet structure. The input of AlexNet is an RGB image with $227 * 227$ pixels and has five convolutional layers (from C1 to C5) and three fully connected layers (from Fc6 to Fc8). The activation function is Rectified Linear Unit (ReLU) [148]. ReLU is defined as:

$$\text{ReLU}(x) = \max(x, 0) \quad (5.1)$$

where x denotes the data points. After running the pre-trained net, all the features have been obtained. The features of each image have 4096 dimensions. There are 60 million parameters in this architecture. Due to a large number of parameters and the few thousand training images, it is time-consuming. Therefore, transfer learning is very convenient. Figure 5.3 shows the combination of AlexNet and SVM classifiers to classify new images. There are three steps in the algorithm of this proposed solution, namely [15]:

- a) All the face images are normalised to a fixed size ($227 * 227$).
- b) AlexNet will extract the 4096 image features using pre-trained deep learning network.
- c) SVM will classify the images using the 4096 features extracted at step 2.

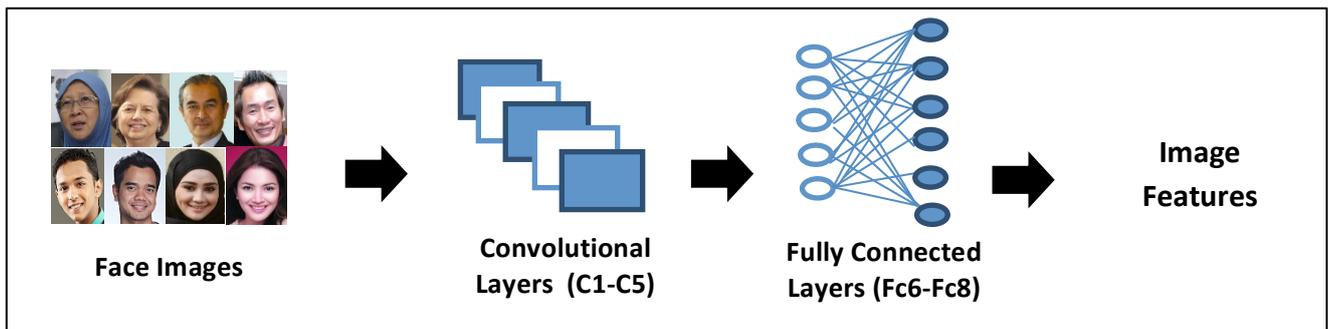


Figure 5.2: AlexNet Structure

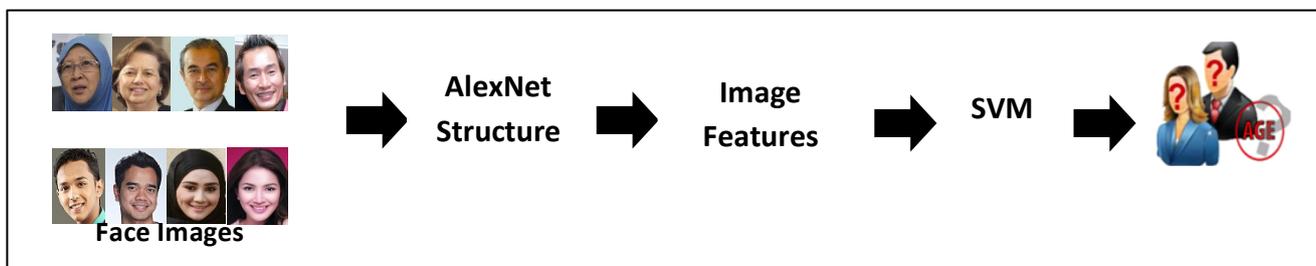


Figure 5.3: Pre-trained net application structure

5.4 Summary

This chapter presents the image data which is an essential part of this research framework for the autonomous detection of anomalous human behaviour. Face image data are applied as supporting evidence to the partial conclusion made based on the methods described in the previous chapter and to show the heterogeneous nature of the data. A face image dataset is built and named GAFace. Features from the face image data are extracted using pre-trained deep learning model. Then, these features are used to classify the age and gender of the human suspect. Age is divided into two groups: “young” and “old”. Gender also has two groups: “male” and “female”. The face images are classified using a support vector machine classifier.

Data Fusion of Heterogeneous Data for Decision Making

6.1 Introduction

This chapter discusses the data fusion technique proposed and used in this study. It is the last stage in the proposed method before sending all results and analysis to a human expert. All the five types of data will be combined using this data fusion technique. The significance of this chapter is that it presents the new data fusion method which can later assist a human expert in making the overall decisions.

6.2 Data Fusion

Data fusion is a technique which integrates or combines data from multiple sensors to receive more accurate information as compared to single, independent sensor. Humans have developed the ability to use multiple senses to survive in life. For example, sense of vision can see something, but cannot feel or touch. However, if combined with multiple senses such as touch, smell and taste, then much more information can be obtained. This is an example of

data fusion, in which combining more data modalities will produce more efficient or confident overall result. Data fusion itself is not a new technique. However, new fusion techniques have emerged along with new sensors, processing techniques and hardware. Therefore, data fusion has become more advanced with a new design and terminology and novel techniques.

This chapter demonstrates a data fusion technique using heterogeneous data based on the proposed overall framework for anomaly detection from heterogeneous data streams/sets. There are five types of data with three different modalities examined in this thesis. All the data has different dimensionalities and cannot be simply combined and integrated. Therefore, a new data fusion technique is introduced which first analyses the abnormalities in each data type separately, determines the degree of suspicion, and, finally, sums up all the degrees of suspicion.

6.3 Data transformation

Data transformation is the process of transforming all data into a value which can be compared. In this situation, all the data is being transformed into the value between 0 and 1. Four methods are introduced to transform all the data. In this case, there are five types of data, but only four methods are needed in this transformation stage. Images will produce two types of data, which are gender and age; then, it will apply the same approach. This is why only four methods are implemented in this transformation stage. The four methods estimate credit card data “eccentricity”, find the difference between credit card and loyalty card data, locate the distance of the suspicious person’s car to the store where the credit or loyalty cards were used, and finally, using images, the accuracy of the classification based on their gender and age is applied with a degree of confidence. After all the values are derived, the data is

summed and divided by the number of types of data resulting in the normalised value of the level/degree of suspicion.

6.3.1 Degree of suspicion – Credit card

The degree of suspicion based on the credit card data λ_k^{cc} is calculated based on the standardised “eccentricity” result. Standardised “eccentricity” has previously been applied to find anomalies in credit card data. From the “eccentricity” results, the degree of suspicion of the credit card transactions data is calculated as follows:

$$\lambda_k^{cc} = 1 - \frac{1}{\varepsilon_k} \quad (6.1)$$

k denotes the data points. Figure 6.1 show the data (refer chapter 7.2 and 8.4) after applying λ_k^{cc} . If the value of λ_k^{cc} is more than 0.9615, then the specific use of the credit card to be suspicious, and if it is less than 0.9615, then it is considered normal. The abnormal value can be determined by the value of n we set in the data set when calculating the standardised “eccentricity” ε . For example, in this data, the value of $\varepsilon = 26$ corresponds to 5σ according to Chebyshev inequality [16]. Thus,

$$\lambda_k^{cc} = 1 - \frac{1}{26} = 0.9615$$

is considered as a threshold. The value of λ_k^{cc} satisfies the following inequality:

$$0 \leq \lambda_k^{cc} < 1 \quad (6.2)$$

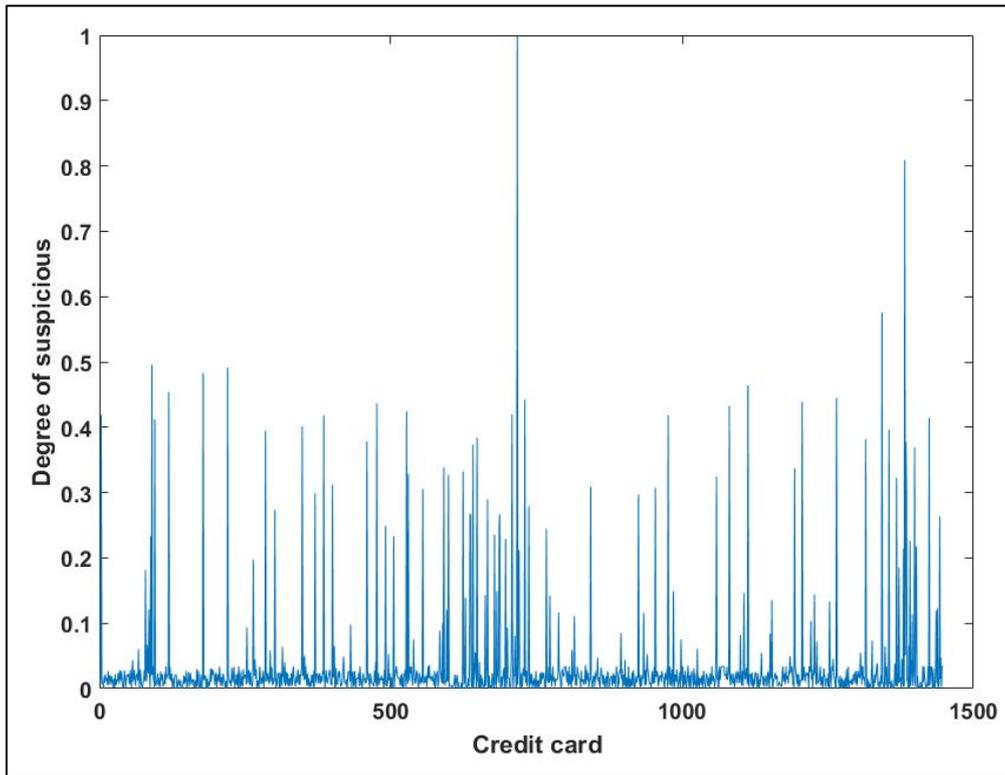


Figure 6.1: Degree of suspicion – credit card

6.3.2 Degree of suspicion – Loyalty card

The degree of suspicion in loyalty card use is calculated as the disagreement between the credit card and loyalty card data, λ_k^{dis} (refer chapter 7.2 and 8.4). Firstly, the difference between the credit and loyalty card data is calculated as follows:

$$\delta_k = Pcc_k - lc_k \quad (6.3)$$

All the values are matched based on the timeline. After that, the difference is calculated to get the absolute disagreement. If the credit card and the loyalty card have the same value, then the value of δ_k is 0, which is considered as not suspicious. Then, the degree of suspicion in loyalty card is calculated based on the value of δ_k transformed into the value between 0 and 1:

$$\lambda_k^{dis} = 1 - \frac{1}{1 + \delta_k^2} \quad (6.4)$$

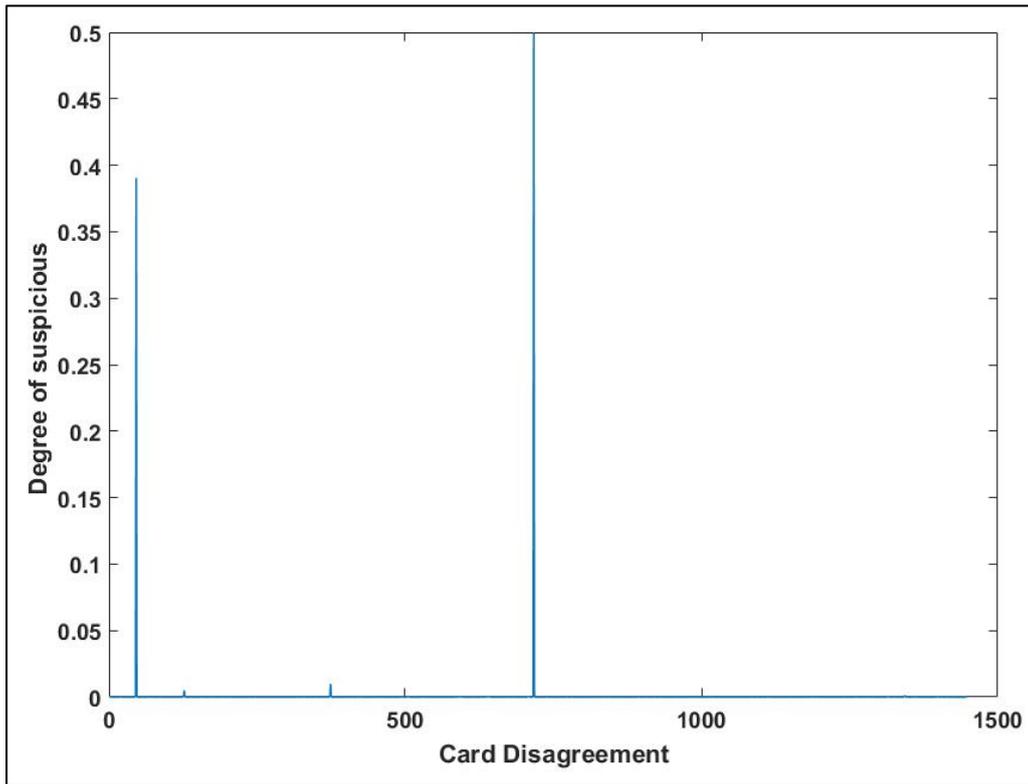


Figure 6.2 : Degree of suspicion – Loyalty card

Figure 6.2 show the degree of suspicion after applying the disagreement between the credit card and loyalty card.

6.3.3 Degree of suspicion - Distance

Distance d is calculated from the location of a store to every person's car location. After that, the degree of suspicion based on the distance between persons, car, and the location of the store λ_k^{loc} is calculated (refer chapter 7.2 and 8.4). The calculation of λ_k^{loc} is conducted as follows:

$$\lambda_k^{loc} = e^{-\frac{d_k^2}{2\sigma_k^2}} \quad (6.5)$$

The value of sigma (σ) can be set based on the distance between a car park and the store location. The standard deviation for the distance from every store location to the car park for all trips every day is calculated, and the average was found. Based on this, the sigma value was calculated to be $\sigma = 555 \text{ metres}$ for the data from VAST Challenge 2014. According to Waerden and Timmermans [149], the normal distance between a car park and a store is 50 metres to 700 metres. The value that has been determined (555 metres) is well within these limits. If the distance between the person's car and the store's location is more than 555 metres, then we consider the degree of suspicion to be high. This study also demonstrates the difference between distances of 50 metres, 555 metres and 700 metres (see figure 6.1, 6.2 and 6.3 respectively). A distance of 50 metres shows that many car parks are far away from the store location. Distances of 555 metres and 700 metres have almost the same value (see table 6.1).

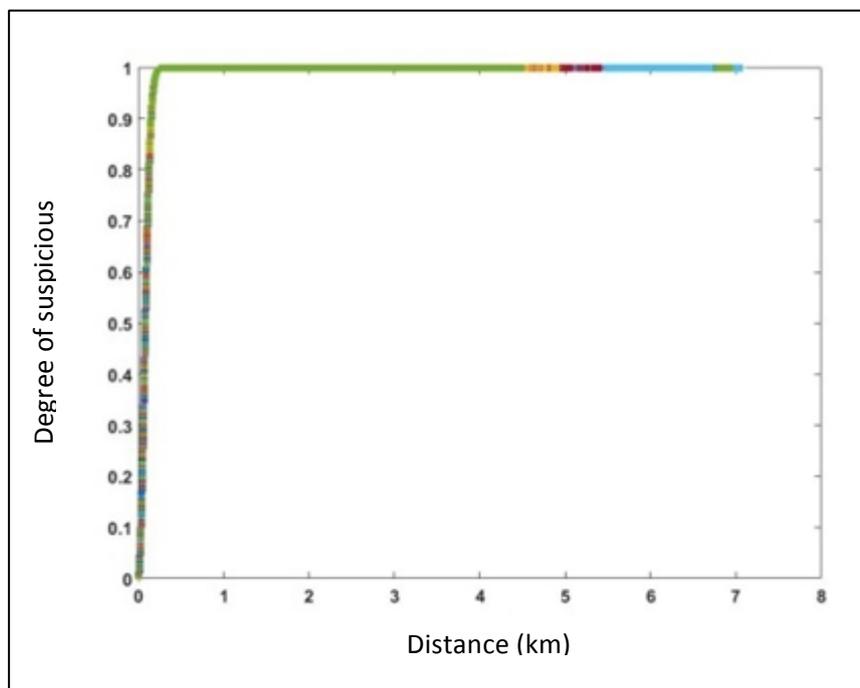


Figure 6.3: 50 metres distance from the car park to the shop location

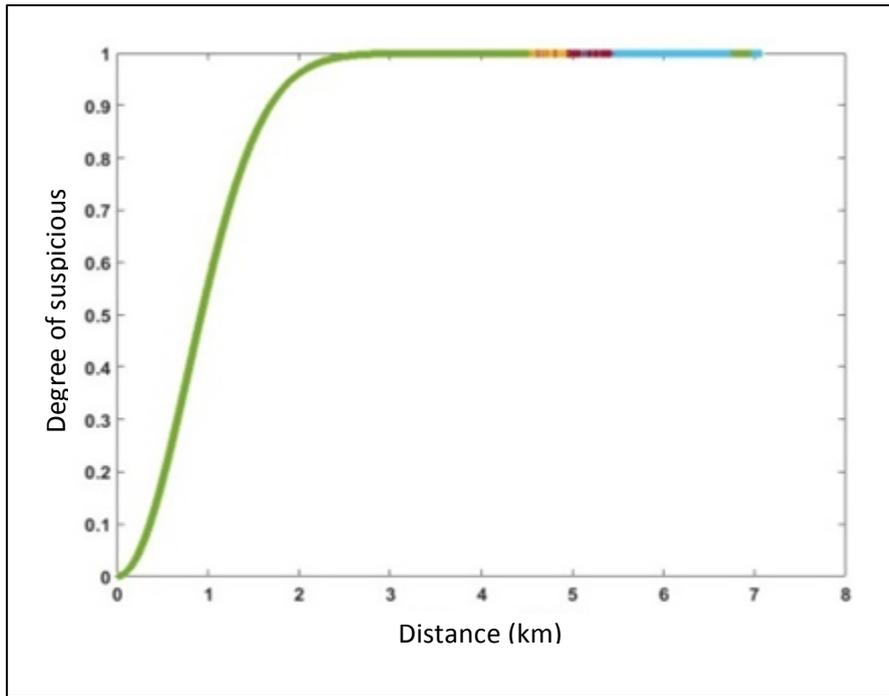


Figure 6.4: 555 metres distance from the car park to the shop location

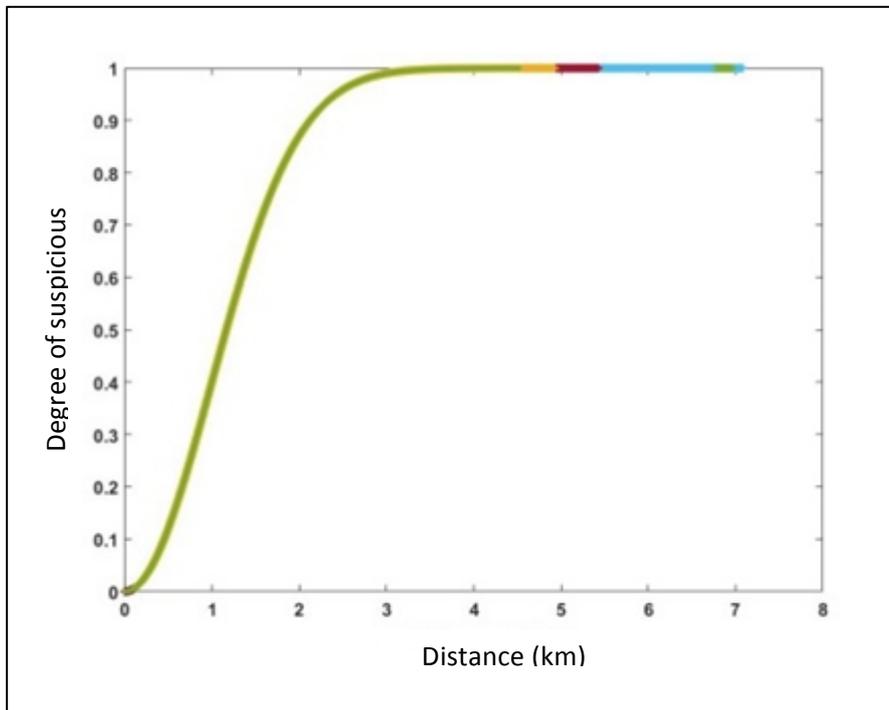


Figure 6.5: 700 metres distance from the car park to the shop location

Table 6.1: Sample data after calculating the degree of suspicion in terms of the distance

Distance (metres)		
50	555	700
1	0.998970985	0.986758381
0.999987913	0.087806342	0.056135129
1	0.873553663	0.727455069
0.999991536	0.09044038	0.057849354
1	0.999992804	0.999415114
0.00361117	0.0000294	0.0000185
0.006133849	0.0000499	0.0000314
1	0.999998696	0.999800158
1	0.999873835	0.996460329
0.999987732	0.087696192	0.056063484

6.3.4 Degree of suspicion – Face Images

Data containing face images are used to identify the gender and the age of the person who used the card. Gender and age are recognised by using pre-trained deep CNN learning and an SVM classifier. As a result, the classification accuracy is used to get the value of the degree of suspicion (if the gender or age do not match the true ones) based on the images. If both the age and the gender based on the images are the same as the true ones, then calculate as follows:

$$\lambda_k^{gen} \text{ or } \lambda_k^{age} = 1 - \text{classification accuracy} \quad (6.6)$$

To accommodate uncertainty due to the classifier error, if the images have a difference in the gender or age, take only the classification accuracy value for the same reason.

$$\lambda_k^{gen} \text{ or } \lambda_k^{age} = \text{classification accuracy} \quad (6.7)$$

6.4 Fusion technique

After all the data have been transformed into comparable values which are between 0 and 1, then proceed to the next step, fusion technique. The next step is a fusion of all partial degrees of suspicion based on the partial data, λ_k^{total} . All the values λ_k^i (for $i=1,2,3,4,5$) are summed

and multiplied by the weights, w . The weights can be set by experts based on the importance of the specific type of data. By default, weights can be set to 1:

$$w_i = 1, \quad i = \forall$$

Then, the sum of all values is divided into the number of types of data. λ_k^{total} is defined as follows:

$$\lambda_k^{total} = \frac{\sum_{i=1}^N w_i \lambda_i}{\sum_{i=1}^N w_i} \quad (6.8)$$

$$\lambda_k^{total} = \frac{w^{cc} \lambda_k^{cc} + w^{dis} \lambda_k^{dis} + w^{loc} \lambda_k^{loc} + w^{gen} \lambda_k^{gen} + w^{age} \lambda_k^{age}}{\sum_{i=1}^N w_i} \quad (6.9)$$

The data fusion technique represents a weighted average. The main novelty is the heterogeneous nature of the data, and each variable represents a different type of data. All the data has already been transformed into values between 0 and 1 to make them comparable. Then, the result of the λ_k^{total} is ranked in a descending order. The highest value will be the most suspicious data. The lowest will be the least suspicious data. The method cannot simply pre-define the suspicious and non-suspicious label. The real meaning is being assigned by a human expert who can determine is a result suspicious or non-suspicious and can set a threshold value. If there is a need to have a line between suspicious and non-suspicious, then 3 sigma or Chebyshev inequality may apply to the result. Finally, the proposed approach does not allow for a fully automatic system that resolves investigations (which would likely be unacceptable to many parties). The aim is rather to simplify the role of the humans, to facilitate, so that they can focus on a small number of cases to be looked in more detail. The proposed approach does simplify the way of processing such huge amount of data, and later this method can assist the human expert in their investigation and making the final decisions.

6.5 Summary

This chapter discusses the data fusion techniques. It is the last stage in the proposed method before sending all the results and analysis to a human expert. All the five data sets will be combined using this weighted average data fusion technique. Before being combined, the data is transformed into the value 0 to 1 to make all the data comparable. This is one of the novelties of this study. Then, all results are sorted into a descending order. The highest value is the most suspicious case. The significance of this chapter is simplifying the processing method for such huge amount of data, which will assist a human expert in the investigation and making the final decision.

Experimental Data

7.1 Heterogeneous Data

This study applies heterogeneous data, which consists of several different modalities of data including images and various signals. In real situation or application, there are many types of data involved in one application or situation. For instance, when a crime takes place, the forensic investigator needs the personal data of the suspect, such as bank account information, travel information from his/her GPS, video surveillance at the crime scene, and, maybe, an email account. There are many types of data involved in such scenarios such as financial data (bank account information), signal data (GPS), image data (video surveillance) or text data (email). It is hard to find many types of data in one application or situation publicly. Most available public data offer one dataset for one application. For example, the UCI Machine Learning Repository [150] provides more than 300 datasets, but every dataset applies to a single application. This web page offers empirical analysis for machine learning algorithms. Therefore, in this study VAST Challenge 2014 data are used. A popular example is considered to be applied by creating a GAFace dataset as an illustration only, as a proof of concept, without limiting the overall methodology to it.

7.2 VAST Challenge 2014

IEEE Visual Analytics Science and Technology (VAST) Challenge is a well-known competition organised by the Visual Analytics community. To demonstrate the framework, a case study based on the VAST Challenge 2014 [17] is applied. The VAST Challenge is based on fictitious (virtual) data, but is quite realistic and includes many problems that need to be solved.

This challenge concerns a fictional company named as GASTech. They experienced an incident in January 2014 when, celebrating the achievements of the company, some staff went missing. There were three mini-challenges and one grand challenge offered in this competition to investigate the missing staff. However, only mini challenge two is chosen with four datasets. The datasets are about financial data (credit card and loyalty card), GPS data and car assignment data. Some of the employees have high-quality company cars which they cannot afford to buy themselves. However, the company did not trust their employees and installed geospatial tracking software to track their employees' trajectories. The tracking data is used for the law enforcement for investigation if something happens. At the same time, law enforcement can also access the financial data of employees (credit and loyalty card). However, data is not available on the day the employees went missing. The data is available for two weeks from 6 January to 19 January 2014. The description of the data in this challenge is defined in Table 7.1. There are four datasets, including GPS data with 685170 data points (sample of data in table 7.2). The smallest data concern assigning a car to each employee at 45 data points, equivalent to the number of employees in GASTech (sample of data is shown in table 7.3). The truck drivers' data has been removed because there are missing values for the IDs of truck drivers. Financial data such as credit card data and loyalty card data includes transactions of every staff member (sample of data in table 7.4 and 7.5,

respectively). The values of credit and loyalty cards are almost the same for every transaction. However, some of the transactions have different values because the point of every transaction in the credit card depends on the item they buy.

Table 7.1: Description of Datasets

Datasets	No. of data points	Attributes
1. Credit Card	1492	1. Timestamp 2. Location 3. Price 4. First Name 5. Last Name
2. Loyalty Card	1393	1. Timestamp 2. Location 3. Price 4. First Name 5. Last Name
3. GPS Data	685170	1. Timestamp 2. Car ID 3. Latitude 4. Longitude
4. Car Assignment	45	1. First Name 2. Last Name 3. Car ID 4. Current Employment Type 5. Current Employment Title

Table 7.2: Sample of GPS Data

Timestamp	ID	Latitude	Longitude
01/06/2014 06:28	35	36.07623	24.87469
01/06/2014 06:28	35	36.07622	24.8746
01/06/2014 06:28	35	36.07621	24.87444
01/06/2014 06:28	35	36.07622	24.87425

Table 7.3: Sample of Car Assignment Data

Last Name	First Name	Car ID	Current Employment Type	Current Employment Title
Alcazar	Lucas	1	Information Technology	IT Helpdesk
Azada	Lars	2	Engineering	Engineer
Balas	Felix	3	Engineering	Engineer
Barranco	Ingrid	4	Executive	SVP/CFO

Table 7.4: Sample of Credit Card Data

Timestamp	Location	Price	First Name	Last Name
01/06/2014 07:28	Brew've Been Served	11.34	Edvard	Vann
01/06/2014 07:34	Hallowed Grounds	52.22	Hideki	Cocinaro
01/06/2014 07:35	Brew've Been Served	8.33	Stenig	Fusil
01/06/2014 07:36	Hallowed Grounds	16.72	Birgitta	Frente

Table 7.5: Sample of Loyalty Card Data

Timestamp	Location	Price	First Name	Last Name
01/06/2014	Brew've Been Served	4.17	Cornelia	Lais
01/06/2014	Brew've Been Served	9.6	Mat	Bramar
01/06/2014	Hallowed Grounds	16.53	Emile	Arpa
01/06/2014	Coffee Shack	11.51	Varro	Awelon

7.3 Face Image Dataset - GAFace Dataset

A face image dataset is not available from the VAST Challenge 2014 case study. Image data is important to show the different types modalities of data. It is also important to aid the overall decision. For example, if the gender and age of the person who is using a credit card are the same as that of the owner of the card there is no suspicion in this respect. If, however, the gender or age of the person using credit card are different this is suspicious. It is hard to find a complete dataset which has examples of heterogeneous data. Therefore, images have been collected separately. The face image data for this study is collected from Google Image (see samples of images in table 7.7). All the face images are of celebrities and politicians in Malaysia. We obtained 310 face images in total. The distribution of the data is provided in Table 7.6. Male images are 166, and female images are 144. There are 160 images of those below 40 years old and 150 images of those above 40 years old. Data were collected from December 2016 to February 2017. They are shown in Table 7.8.

Table 7.6: Descriptions of datasets

Face Image Class	Number of face images
Male	166
Female	144
Age below 40 years old	160
Age above 40 years old	150

Table 7.7: Samples of images from GAFace dataset

Gender	Age Group	Images
female	<40 years old	
	>=40 years old	
male	<40 years old	
	>=40 years old	

Table 7.8: Sample images of each class

Class	Younger (below 40 years old)	Older (above 40 years old)
Male	82	84
Female	78	66

7.4 Data Pre-processing

Data pre-processing is a data preparation process which involves transforming raw data into an understandable format. Data can have a variety of problems, such as incompleteness or missing values, noise which leads to impossible data combination (e.g., Age: 40, Date of birth: 20/07/1985), and inconsistent data like out-of-range values (e.g., Income: - 100). Incomplete data may come from data collection. Data may not be applicable at the time it is collected and noisy or incorrect because of hardware failure, due to human or computer error at data entry and inconsistent data from different sources. Data preprocessing can help to

overcome these problems and produce high-quality data. There are two important phases involved in data preprocessing which are:

- a) feature selection
- b) feature extraction

Feature selection is the process of selecting the important features to be used for further processing. Feature extraction is the process of transforming the original features (e.g., GPS coordinate) into a set of new features (e.g., speed, direction). In this study, we apply feature selection in financial data and feature extraction in GPS data and image data.

7.4.1 Financial data (credit and loyalty card data)

Financial data is data related to personal income, transactions or withdrawal and balance sheet. Normally, financial data can be derived from a bank statement. A bank statement can be an account statement or credit card statement. Transactions of debit and credit in the bank statement are the features in financial data. Daily transactions can create a behavioural pattern of a person especially in the withdrawal or spending activity. Spending activity can only be traced if a person is using the credit or debit card. If a person is using cash to buy something, transaction activity cannot be traced, but the withdrawal can be seen in the account statement. Financial data can provide significant information, especially for business activity and banking. The banking sector has information of their customer making it easy to promote their finance products based on the patterns in the customer financial data. In this scenario, financial data concerns the money spent using the credit card. Transactions of credit cards will also appear in the loyalty card if they are used together. Normally, all transactions are the same except when they are not using the loyalty card, or maybe the shop did not accept the loyalty card. Therefore, for some of the transactions, the two values are not the same. In the

VAST 2014 Challenge data that we consider, there are five attributes (timestamp, location, money spend, first name and last name) in credit card and loyalty card data, but only money spent is extracted from these datasets. Money spent according to the credit card, C_i and money spent according to the loyalty card are as follows:

$$x_i = [C_i, L_i] \quad (7.1)$$

where i denotes the data points. From these datasets, different features can be extracted including:

- a) Total spending per person.
- b) Total spending per person and per day.
- c) Total spending per location.
- d) Total spending per location per day.

These features can be extracted from the credit card and loyalty card data. Features including “total spending per person” can be extracted by summing up every spending made with the credit card and loyalty card for every staff member. There are 45 data points which represent 45 staff members. From this data, it will be easy to determine which staff members spent more and which staff members spent less. Then, the total spending by every staff member per day can be extracted. This can show which day the staff member spent more or less. After that, the total amount of spending per location can be extracted. From the location, places in which people spend more can be identified. The last feature is creating data for every location per day. Hence, it will be easier to find which day and which location is significantly different from others. These features can give results concerning which person has a suspicious spending behaviour, when the suspicious behaviour of spending happened

and where the suspicious spending took place. Money spent is normalised between 0 and 1 to make the data comparable. Normalization requires a range (min, max) per feature:

$$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (7.2)$$

7.4.2 Global Positioning System (GPS) data

GPS provides positioning, navigation and timing. From GPS satellites, information about coordinates can be derived. GPS coordinates consist of latitude and longitude. Latitude is a horizontal line from east to west across the globe. The longest and the main line of latitude is equator. The Equator represents 0° latitude. Each line represents 1°, 2° and up to 90° (North Pole). All the lines of latitude above the Equator are indicated with the letter N which denotes north of the equator. If moving to the south of the equator, coordinate will indicate the letter S (South Pole). Longitude is a vertical line from the North Pole to the South Pole. The main line of longitude at 0° is the prime meridian. All lines of longitude east of prime meridian are indicated with the letter E, and those to the west of the prime meridian are indicated with the letter W. Values of latitude and longitude are in three different formats:

- a) degrees in (minutes and seconds), for example 36° 3' 48" N 24° 51' 2.93" E
- b) degrees in (decimal and minutes), for example 40° 26.767' N 79° 58.933' W
- c) decimal degrees, for example 36.06647° N 24.85327° E

GPS coordinates use a decimal degree format. There are three attributes of this data:

- a) Timestamps
- b) ID
- c) GPS coordinates (longitude and latitude)

The most important features are the GPS coordinates. Trajectory information can be determined from the GPS coordinates such as the projection of the trajectory, average speed, ratio of the trajectory angle and distance. Trajectory can represent mobility of people (e.g. people moving by bicycle or jogging carrying mobile phone), mobility of transportation (e.g. vehicle supported by GPS – taxis, bus, aircraft), mobility of animals (e.g. biologist collecting the moving trajectories animal – migrating or behaviour) or mobility of natural phenomena (e.g. meteorologists, environmentalists, climatologists – collecting trajectories of some natural phenomena – hurricanes, tornados) [151]. After the preprocessing, the GPS data can be compressed as a vector:

$$y_i = [N_i, d_i, \bar{R}_i] \quad (7.3)$$

In [152] several steps were proposed to get different features from the trajectory type data.

a) Trajectory, T_i

$$T_i = \{(a_{i,j}, b_{i,j}); j = 1, \dots, N_i\} \quad (7.4)$$

Where $a_{i,j}$ denotes the latitude and $b_{i,j}$ denotes the longitude, N_i is a duration (in seconds) of the trajectories contained in that specific sample.

b) Projections of the trajectory, r_i

Projections include the horizontal and vertical projections of the trajectory. This shows the trajectory of a person from start point to end point or destination. The advantage of using this feature is to separate the trajectory per axis. Projections r_i are defined as follows:

$$r_i = (a_{i,N_i} - a_{i,1}, b_{i,N_i} - b_{i,1}) \quad (7.5)$$

c) Average Speed, v_i

The next feature is average speed. It is derived from the distance of the route. This feature can also differentiate vehicles with varying speed. Average speed v_i is calculated as follows:

$$v_i = \frac{1}{N_i - 1} \sum_{j=1}^{N_i-1} (a_{i,j+1} - a_{i,j}, b_{i,j+1} - b_{i,j}) \quad (7.6)$$

d) Distance, d_i

The distance shows the span of the destination from the start point till the end point is. The distance d_i is calculated as follows:

$$d_i = N_i v_i \quad (7.7)$$

e) Ratio of the trajectory change angle, \bar{R}_i

The trajectory angle $\theta_{i,j}$ is calculated to find the sharpness of turns in the trajectory, T_i . The angle $\theta_{i,j}$ is then compared to check whether it is less than 90° or not. 90° is chosen to compare the angle because normally when people move, they will go straight to the destination and sometimes turn back but not always. The trajectory may reach 90° , but it is quite abnormal if we turn more than 90° . R_i denotes the time the angle exceeds 90° during single trajectory. If the angle is less than 90° , then it will give the number of normal values, $R_i \leftarrow R_i + 0$, else the number of abnormal values, $R_i \leftarrow R_i + 1$. Then, the ratio \bar{R}_i is calculated. The trajectory change angle $\theta_{i,j}$ and the ratio \bar{R}_i are defined as follows:

$$\theta_{i,j} = \arctan(b_{i,j+1} - b_{i,1}, a_{i,j+1} - a_{i,1}) \quad (7.8)$$

$$\text{IF } (\theta_{i,j} < 90^\circ) \text{ THEN } (R_i \leftarrow R_i + 0) \text{ ELSE } (R_i \leftarrow R_i + 1) \quad (7.9)$$

$$\bar{R}_i = \frac{R_i}{N_i} \quad (7.10)$$

All features are normalised to make the data comparable. The process is similar to equation (7.2) used for the credit and loyalty card data.

7.5 Summary

This chapter presents the experimental data that were used in this study. The data was acquired from the VAST challenge 2014. Three different datasets have been applied which include credit card, loyalty card and GPS data. This study also applied face images data. These image data is collected from the Google Images and named as a GAFace dataset. The combination of VAST challenge 2014 signal type/modality data of various types(financial, GPS, etc) was combined with the image data from the GAFace to form a heterogeneous mixture of data used in this study uniquely. Then, the chapter explained how the data was processed. The next chapter will discuss the results and analysis after applying the proposed techniques to these datasets.

Results and Analysis

8.1 Introduction

This chapter represents the results and analysis after applying the research framework. There are three sections. The first result is after applying anomaly detection, followed by image results and data fusion results. This chapter also explains the analysis of the overall results.

8.2 Anomaly Detection Results

This study applied to credit card and loyalty card data using RDE to find anomalies automatically and in real-time. Figure 8.1 shows the anomaly of the money spent using RDE. There are three anomalies detected using a very conservative threshold of 6σ . These three anomalies are spending no. 774 which is the highest spending, \$10000 and spending no. 1264 (\$600) and no. 1291 (\$361.54). However, for the loyalty card data, if 6σ is used as a threshold, there is no anomaly detected (see figure 8.2). An anomaly in the loyalty card data is only detected when 3σ is used, but the number of anomalies is too high (61 anomalies), see figure 8.3. For the credit card data, if 3σ is set, the number of anomalies detected is also high (45 anomalies), see figure 8.4.

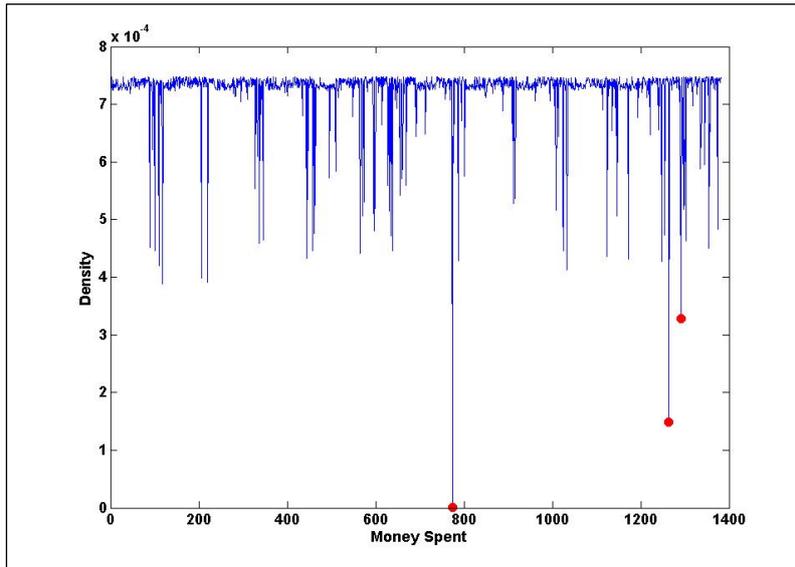


Figure 8.1: RDE on credit card data – 6σ

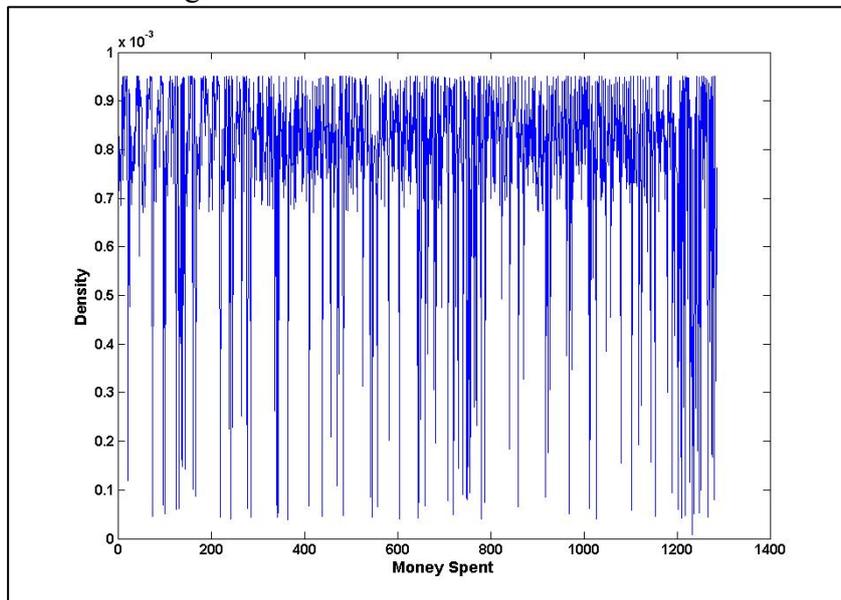


Figure 8.2: RDE on loyalty card data – 6σ

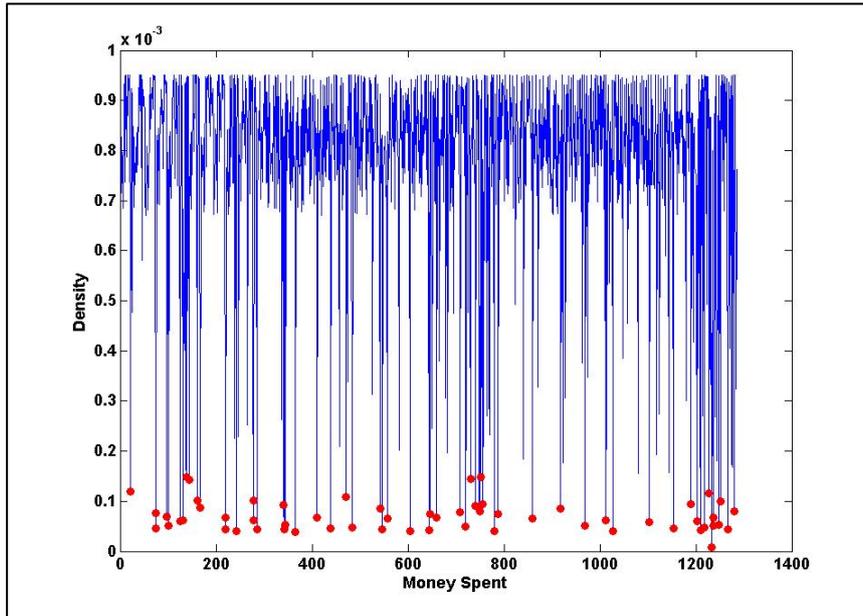


Figure 8.3: RDE on loyalty card data – 3σ

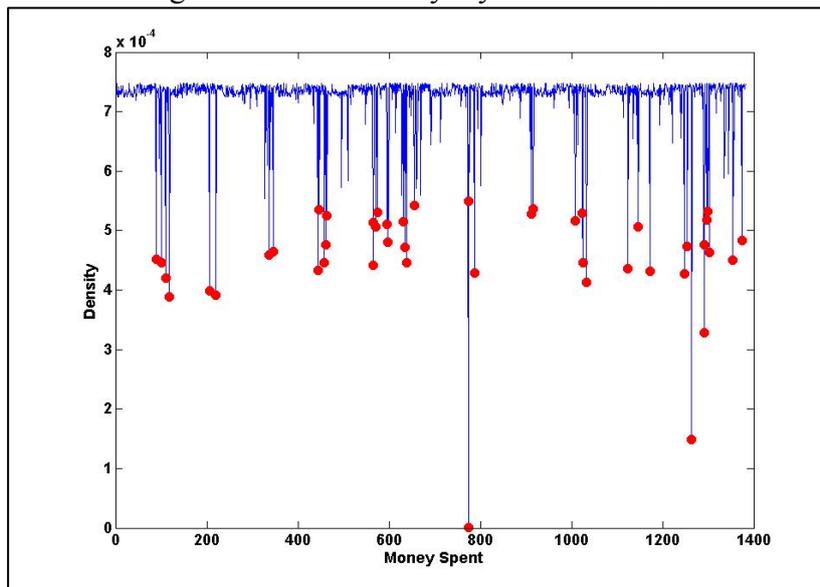


Figure 8.4: RDE on credit card data – 3σ

Then, the approach (empirical data analytics) EDA is applied to detect anomalies. Figure 8.5 shows the standardised eccentricity, $\varepsilon_k(x)$ for the money spent based on the credit and loyalty card data. There is one noticeable anomaly in this data where the staff member no. 31 spends \$10,000. Other attributes have also been analysed (money spent, day and staff member (see figure 8.6)). This figure clearly shows that only one staff member spent too

much, which is abnormal. Staff member no.31 spent an obviously high amount in one day compared to the other 13 days and other staff members. Then, we remove the abnormal data of \$10,000 (see Figure 8.7). The results show two anomalies in this data. There is one case which shows an obvious anomaly (spending of \$600 by staff member no. 43 (see figure 8.8)) which is $> 5\sigma$ away from the mean. According to the Chebyshev inequality, this translates to $<4\%$ of the data. Figure 8.9 shows that there is one anomaly in the loyalty card data. The same staff member no. 43 spends 600 and is the noticeable one. Again, this is $> 3\sigma$ while the other anomalies are above 5σ . Analysing daily spending on credit card and loyalty card data shows two obvious spending patterns by staff member no. 43 and staff member no. 40 (see figure 8.10). The two staff members are the top management of this company. Staff member no. 43 is the CEO (Chief Executive Officer) of the company and staff member no. 40 is the COO (Chief Operating Officer). Therefore, they have the power of spending much money. However, the suspicious thing is that staff member no. 31 spends \$10,000 on the credit card but is not listed in the loyalty card data for this transaction. Normally, when people spend that much, they will swipe their loyalty card together with the credit card. Assuming that maybe (s)he is not bringing the loyalty card when spending \$10,000. When analysing the location, where the money was spent, it shows clearly in figure 8.11 and 8.12, that the highest credit card spending was at Frydos Autosupply, but the loyalty card is just half of the spending on the credit card. Compared to other locations, the spending using the credit card is almost the same as the loyalty card. This again shows suspicious spending using a credit card for staff member no.31. Figure 8.13 shows the total spending of every staff member using a credit card, and again staff member no.31 spends obviously highest compared to the other staff members. For the loyalty card, the spending pattern is almost the same, and there is no anomaly detected in this dataset (see Figure 8.14). For all staff members, the spending patterns were analysed for each day. Figures 8.15 and 8.16 show the total spending per day

for staff member no. 31 using credit and loyalty card, respectively. It demonstrates that the highest spending for this staff member is on day 8, and there is no anomaly in the loyalty card. After that, every location has to be analysed. Location no. 11, Frydos Autosupply has the highest number of money spent in this location using a credit card, but there is no difference with the loyalty card. Based on this analysis and discrepancies, conclusions about the credit card and loyalty card spending behaviour can be made. Staff member no. 31 has suspicious behaviour based on the spending using the credit card on day eight at a location no. 11, Frydos Autosupply. Nevertheless, the spending behaviour using a loyalty card is normal, and there is no other suspicious behaviour detected in this dataset.

A possible explanation can be that someone else, but not staff member no.31, used his/her credit card. This can reduce the huge amount of raw data into a much smaller amount of suspicious data (in this case, regarding staff member’s no.31, 40 and 41) and location, Frydos Autosupply, which maybe further clarified with video from the CCTV (if available). As it will be demonstrated later, this can also be identified by analysing the travel data.

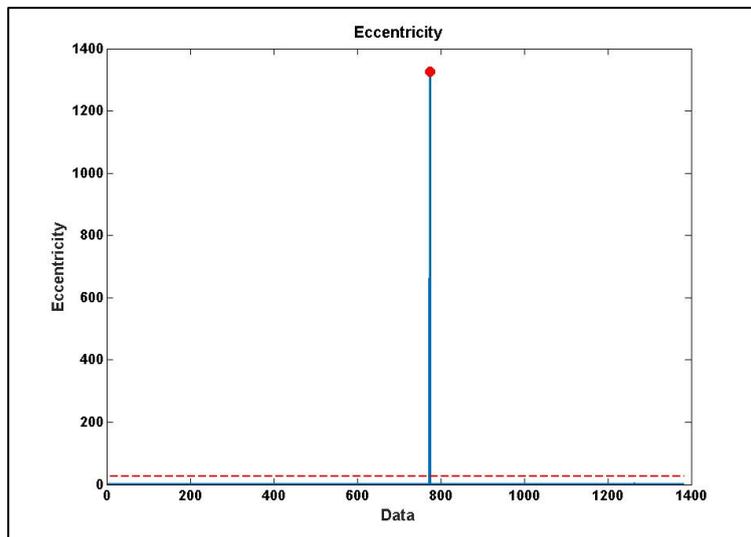


Figure 8.5: Anomaly on credit card usage

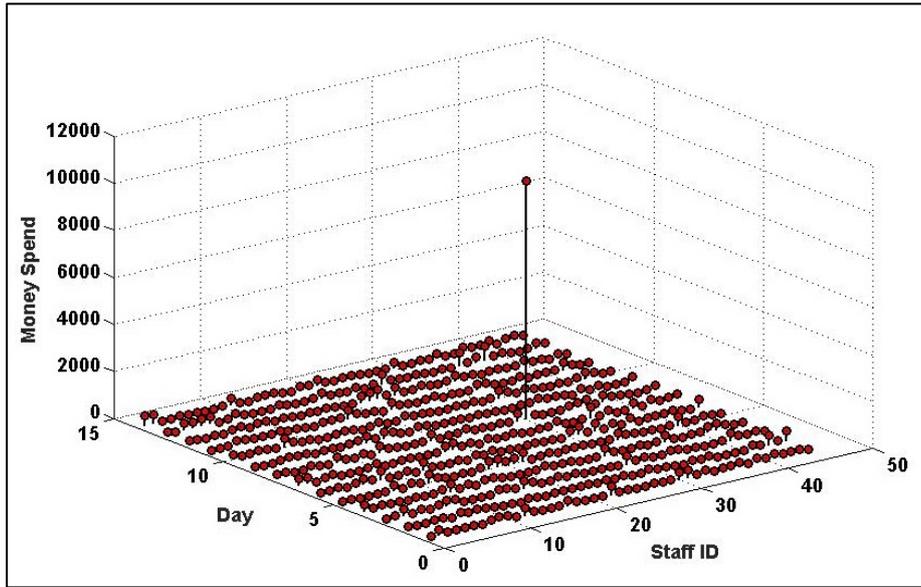


Figure 8.6: Anomaly on money spend, day and staff ID

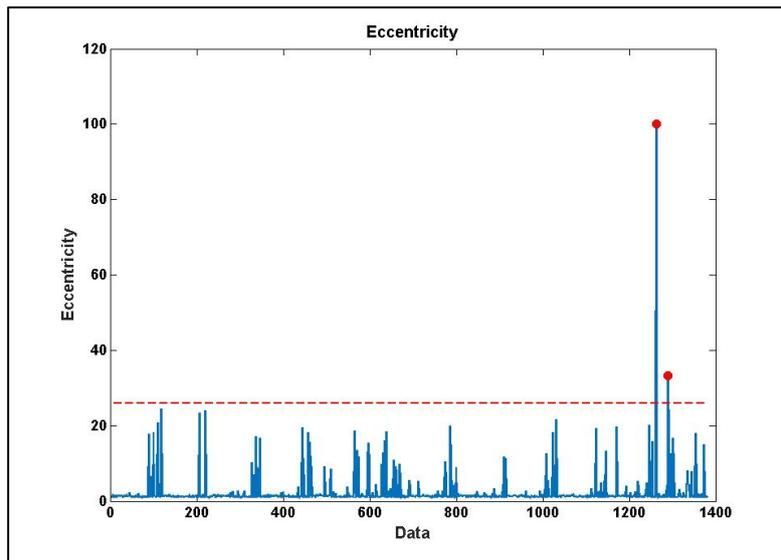


Figure 8.7: Anomalies based on the credit card transaction data after removing the first anomaly

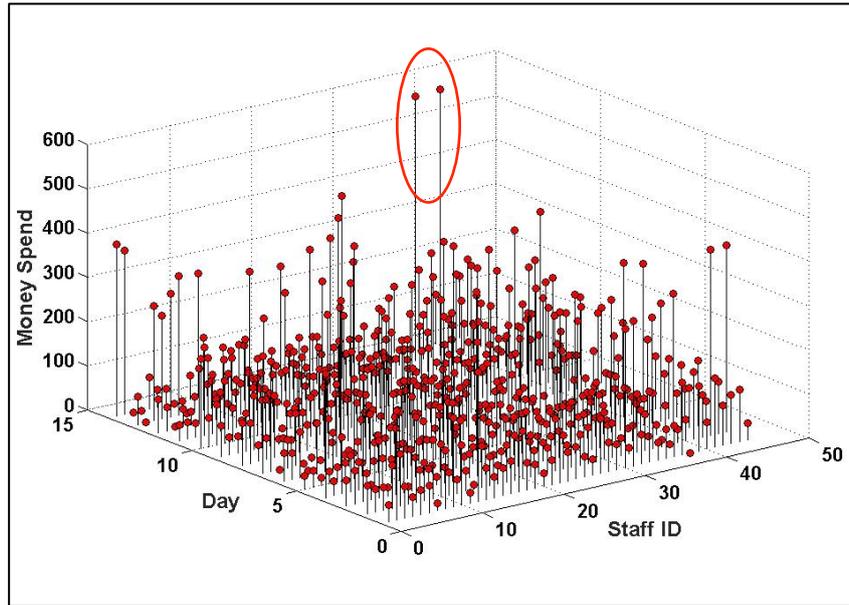


Figure 8.8: Anomalies based on the money spend, day and staff ID after removing the first anomaly

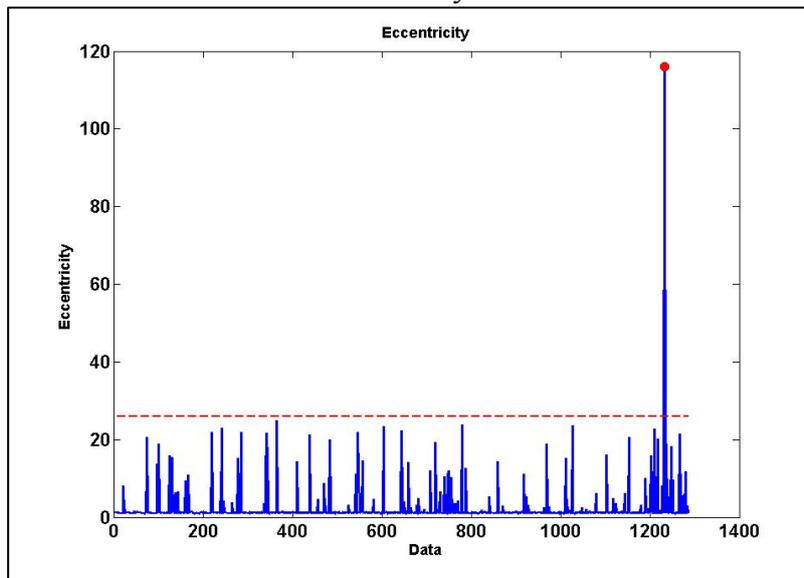


Figure 8.9: Anomalies by the loyalty card data

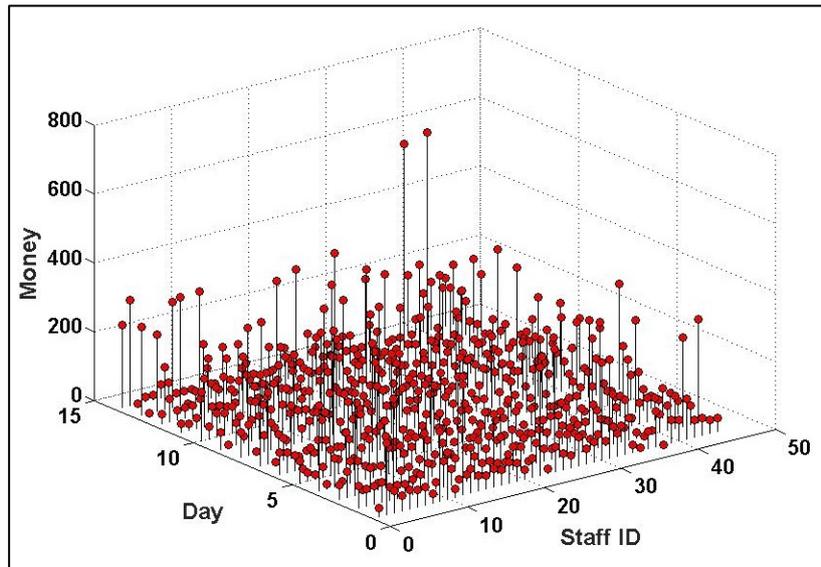


Figure 8.10: Anomalies by the loyalty card data, day and staff ID

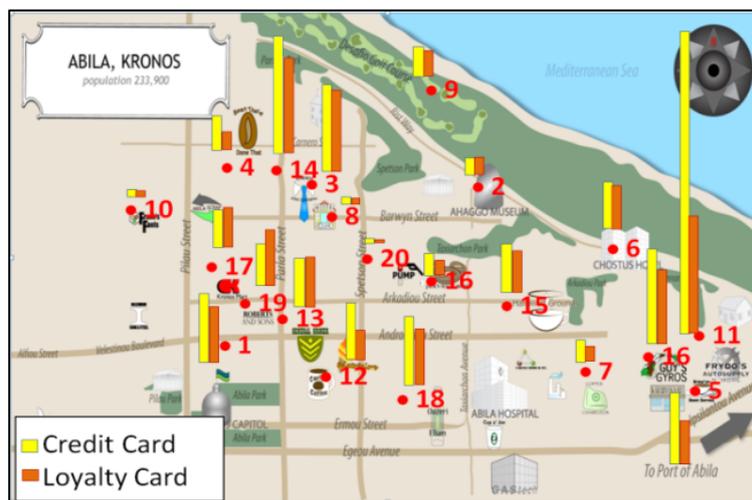


Figure 8.11: Comparison of the total spending using credit card and loyalty card in different locations

Table 8.1: Location legend

Locations			
1	Abila Zacharo	11	Frydos Autosupply n' More
2	Ahaggo Museum	12	Gelatogalore
3	Albert's Fine Clothing	13	General Grocer
4	Bean There Done That	14	Hallowed Grounds
5	Brew've Been Served	15	Jack's Magical Beans
6	Chostus Hotel	16	Katerina's Café
7	Coffee Cameleon	17	Ouzeri Elian
8	Coffee Shack	18	Shoppers' Delight
9	Desafio Golf Course	19	Roberts and Sons
10	Frank's Fuel	20	U-Pump

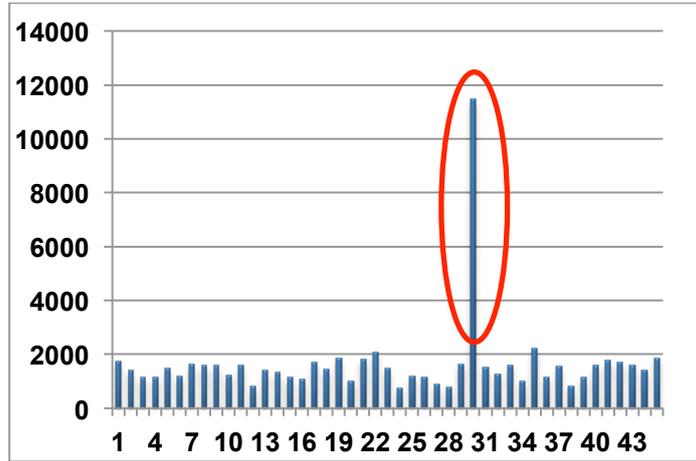


Figure 8.12: Total spending per person using credit card data

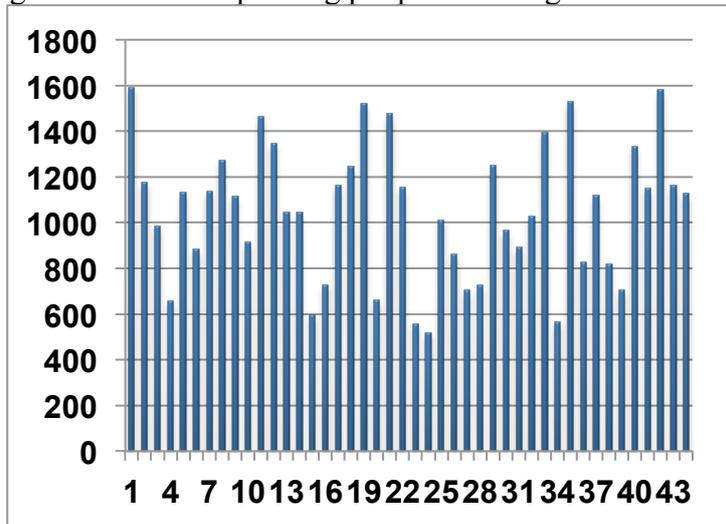


Figure 8.13: Total spending per person using loyalty card data

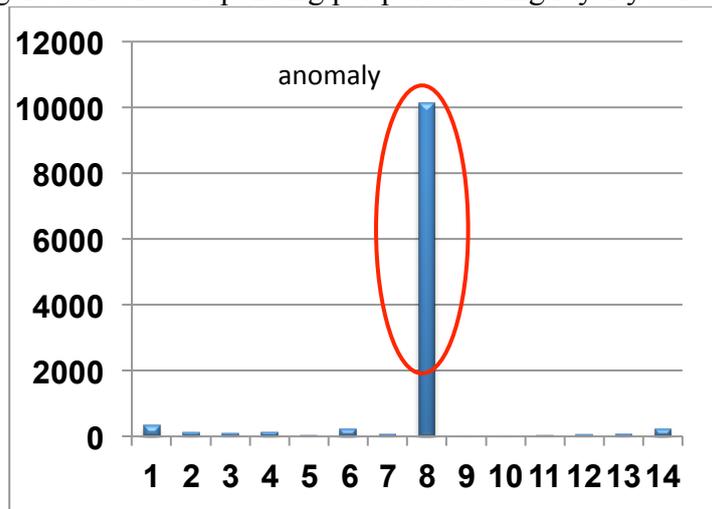


Figure 8.14: Total spending per day using credit card data- Staff Member no.31

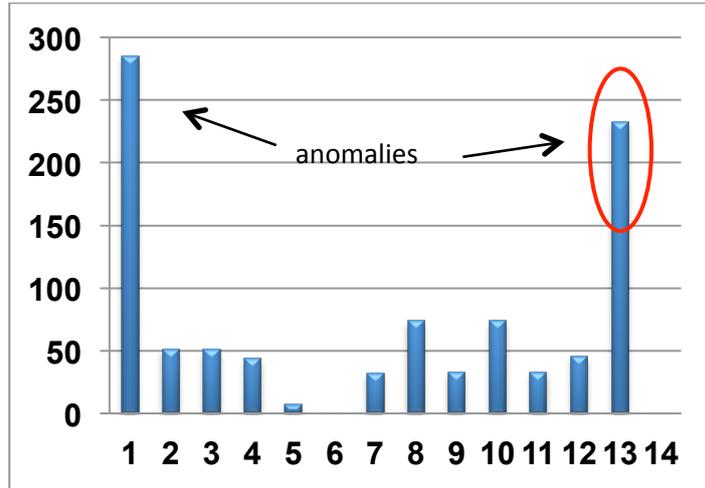


Figure 8.15: Total spending per day using loyalty card data – Staff Member no.31

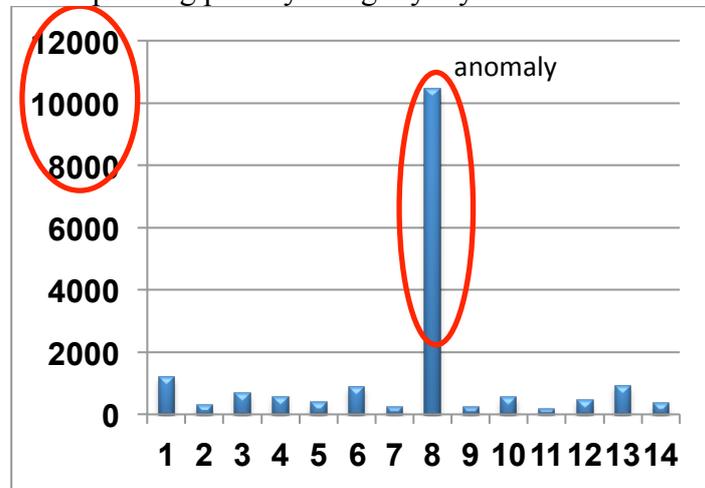


Figure 8.16: Total spending using credit card data per day at Frydos Autosupply

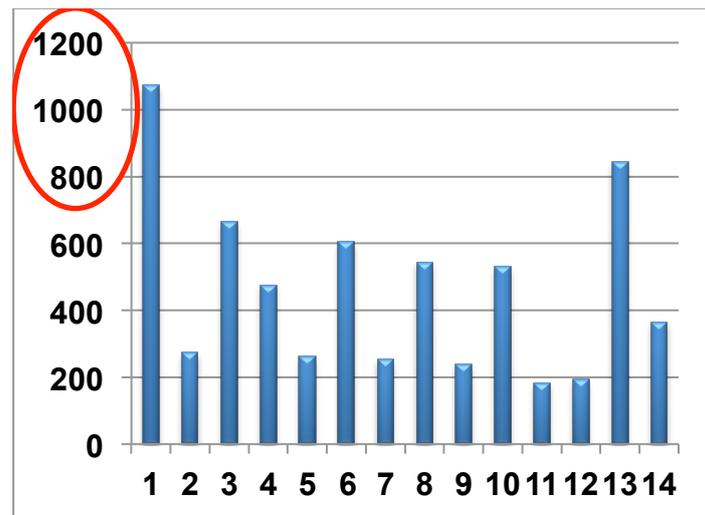


Figure 8.17: Total spending using loyalty card data per day at Frydos Autosupply

The data about the GPS position of the vehicles used by staff members is the biggest dataset and gives much useful information about the trajectory of their movement. All information about the travel time, travel distance and trajectory angle ratio has been transformed and used to calculate eccentricity to discover the suspicious behaviour among the staff members. Figure 8.19 shows the eccentricity based on the travel time, distance and ratio of the trajectory angle. The anomalous data shows that there is one staff member that travelled 14 times between 6 January and 19 January 2014. The anomalies are concerning staff member no. 18. Figure 8.20 also shows examples of patterns of both normal and abnormal travel behaviour. Figures 8.21, 8.22, and 8.23 shows the eccentricity and anomaly of travel time, distance and ratio of the trajectory angle based on 5σ . Table 8.1 shows the ID, date and travel behaviour (abnormal). In this table, all of the travel behaviours are abnormal. Related to the suspicious spending from the credit card, analysis of the trajectory is made on the day of the transaction. The credit card of staff member no.31 was charged 10,000 on 13 January 2014 at 19.20 pm. Figure 8.24 shows a comparison of the trajectories of staff member no.31 and staff member no.41. The trajectory is shown from 17.57 pm to 20.10 pm on 13 January 2016. The trajectory shows that staff member no.31 did not go to the location where the credit card was charged, while staff member no.41 has a trajectory, to the location at the same time the credit card has been charged. It shows how automatically using the newly proposed data analysis method detected something suspicious for staff member no.41. After the highest amount of spending on 13 January 2014, staff member no.31 is not using the credit card until 16 January 2014. Assuming the credit card is not with him/her from 13 January 2014 to 15 January 2014. (S)he started using the credit card again only after 16 January 2014. A possible explanation can be that staff member no.31 went somewhere else and left his/her credit card which was misused by staff member no.41. This was detected

fully automatically, and also it was determined that this amount spent has extremely high value of eccentricity ($> 35\sigma \rightarrow < 0.1\%$ of data samples).

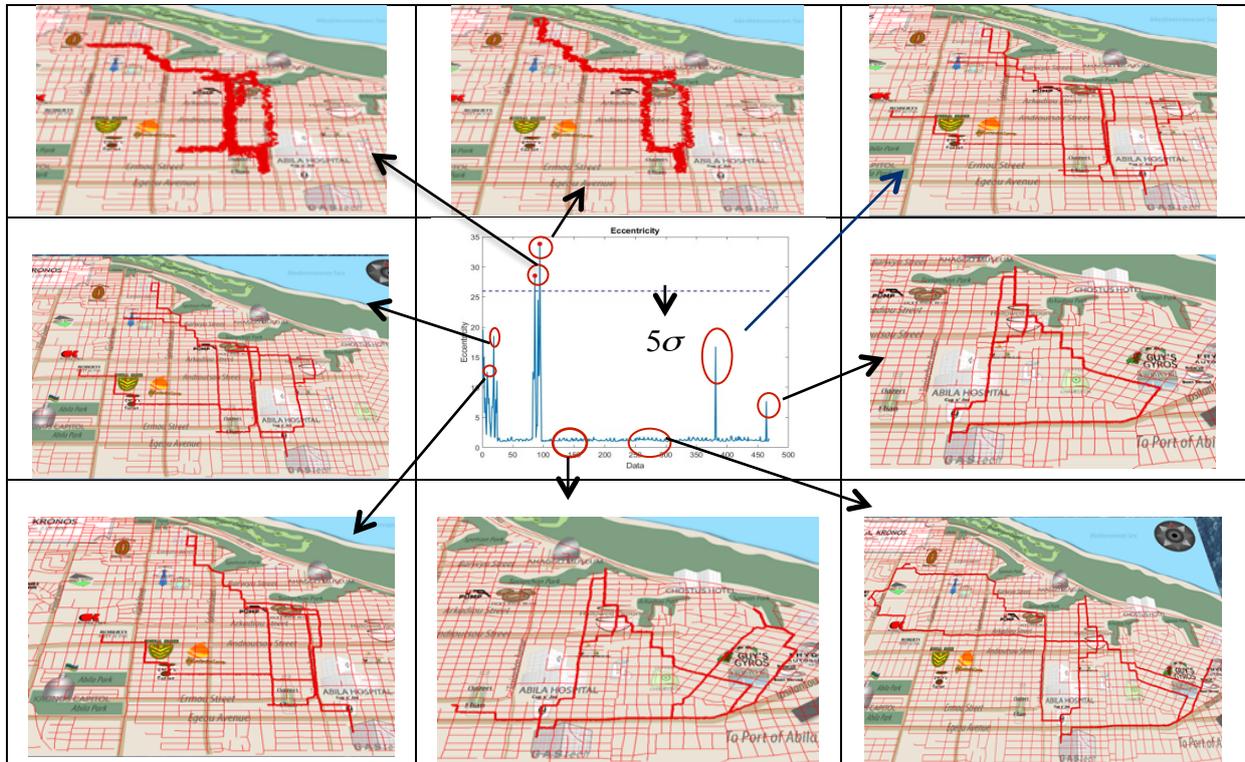


Figure 8.18: Eccentricity of travel time, distance and trajectory angle ratio and normal and abnormal behaviour.

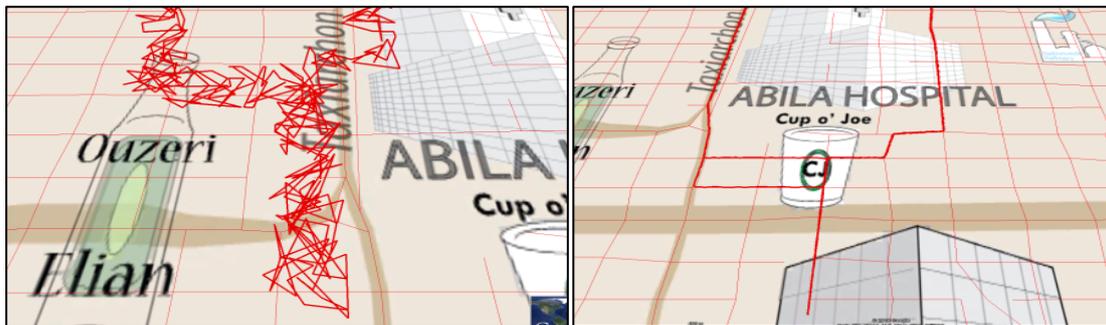


Figure 8.19: Example of abnormal and normal trajectory

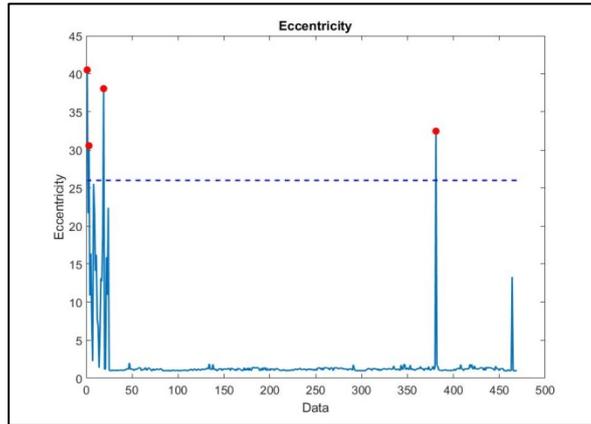


Figure 8.20: Anomalies detected based on the travel time using 5σ

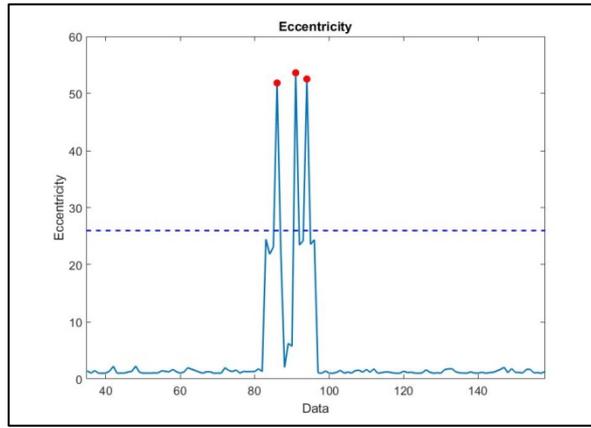


Figure 8.21: Anomalies detected based on distance using 5σ

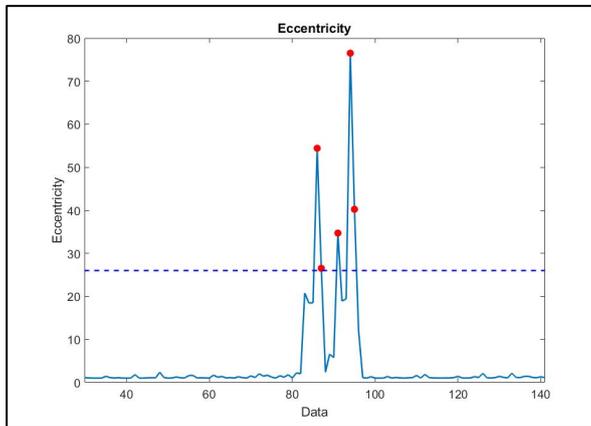


Figure 8.22: Anomalies detected based on the ratio of the trajectory angle using 5σ

Table 8.2: Description on abnormal trajectory

ID	Date	Abnormal
18	9 Jan 2014	Seven trips – abnormal
18	17 Jan 2014	Seven trips – abnormal

in 1998. These features motivated Viola and Jones in 2004 to propose a face detection algorithm [154] which uses three kinds of features. The value of a two-rectangle feature is the difference between sums of the pixels in two rectangular regions. Size and shape of the area are the same and horizontally or vertically adjacent (see figure 8.24A and 8.24B). A three-rectangle feature calculates the sum of two outside rectangles subtracted from the sum in a centre rectangle (see figure 8.24C). Lastly, a four-rectangle feature calculates the difference between diagonal pairs of rectangles (see figure 8.24D). Oualla et al. [155] state that there are five advantages of Haar-like features and only one disadvantage. Advantages of Haar-like features are:

- a) More robust to illumination changes compared to the colour histogram.
- b) Compared to the pixel-based system, the feature-based system operates much faster.
- c) The sum of pixel responses within a given sub-rectangle of an image to be computed quickly.
- d) To extract a Haar-like feature respond, only several accesses to the integral image are required
- e) Can be applied in real-time detection.

The disadvantage of Haar-like features is that an object which rotates is sensitive to angle change and will make Haar-like features challenging to solve. From these face images, three different feature sets are processes, namely 48 features (4x4 images), 108 features (6x6 images) and 192 features (8x8 images).

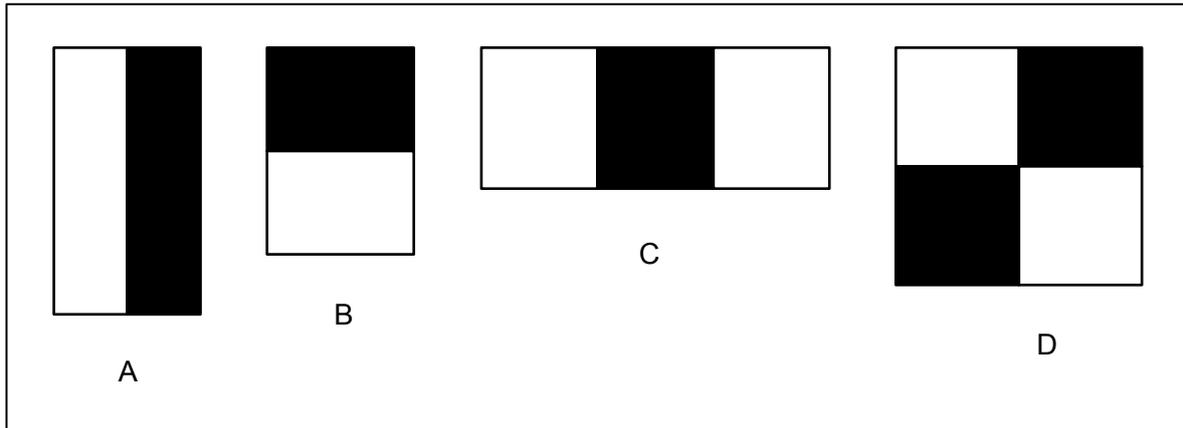


Figure 8.24: Example rectangle features shown relative to the enclosing detection window. Figure (A) and (B) show two-rectangle features. The three-rectangle feature shown in figure (C), and (D) shown a four-rectangle feature.

8.3.2 Classification using Haar-like features

All the images have to be pre-processed to extract the features. The facial features are extracted using Haar-like features, introduced by Viola and Jones in 2001 [29] who applied integral images which represent two-dimensional lookup table. An integral image is the sum of all pixels from the top left to the right and down. From these face images, 3 different feature sets are processed, namely 48 features (4x4 images), 108 features (6x6 images) and 192 features (8x8 images). After extracting the features, SVM is used to classify the images. Results from using Haar-like features and SVM are represented in Table 8.3. When using 192 features, the highest percentages of accuracy for age classification and gender classification are 57.43% and 72.73%, respectively. Meanwhile, the classifier which used 48 features produced the lowest result, namely 51.79% and 69.8% for each respective classification.

Table 8.3: Haar-like features and classification SVM rates (accuracy)

<i>Feature Extraction Techniques</i>	<i>Age</i>	<i>Gender</i>
Haar-Like Features		
4 x 4	51.79%	69.8%
6 x 6	52.82%	72.49%
8 x 8	57.43%	72.73%

8.3.3 Classification using pre-trained network features

In this study, two sets of images (for training and testing) are used to be run by the pre-trained deep learning network to extract features. Each dataset has several subsets, and each subset describes a particular class. Next, all images are resized to 227*227 to input images to the Alex net [148]. After running the pre-trained net, all the features have been obtained. The features of each image have 4096 dimensions. Subsequently, SVM is used to classify the images and obtained 80.17% average accuracy for age classification using 10-cross-validation (Table 8.4). The confusion matrix is used to tabulate the classifications per class (Table 8.5).

Table 8.4: Pre-trained net results

<i>Age</i>	<i>Gender</i>
80.17%	90.33%

Table 8.5: Confusion matrix of age classification

	<i>Below 40 years old</i>	<i>Above 40 years old</i>
<i>Below 40 years old</i>	131	29
<i>Above 40 years old</i>	33	117

Then, the whole dataset is utilised for gender classification using the same process as used for the age classification. The average accuracy achieved for gender classification is 90.33% (Table 8.4) including 136 females and 143 males. The confusion matrix is shown in Table 8.6.

Table 8.6: Confusion matrix of gender classification

	<i>Male</i>	<i>Female</i>
<i>Male</i>	143	23
<i>Female</i>	8	136

The results show that the new method proposed in this study allows promising accuracy is approaching human capabilities (humans are also not perfect in determining age). Therefore, this method can be applied successfully for age and gender classification based on face images and the results can be used for automatic detection of anomalous human behaviour.

In this study, a method for anomalous human behaviour detecting gender and age classification using a pre-trained deep CNN architecture and SVM classifier is proposed for use in an automatic system for detecting anomalous human behaviour. This work is a continuous study from previous research on the use of heterogeneous data for autonomous detection of anomalous human behaviour. Face images are a type of data which can improve the evidence in investigating a suspect or an event. This method will reduce the scope in identifying suspicious persons in the forensic investigation area. The first step in the methodology involves extracting the features from face images using a pre-trained deep learning convolutional network. Then, SVM is applied to classify the images. The transfer learning was then used to successfully utilise the already learnt knowledge for a new task with limited dataset. The proposed method was tested on the GAFace dataset, and the results

are very encouraging. The percentages of accuracy for gender and age classifications are 90.33% and 80.17%, respectively, which demonstrates that the solution is feasible.

8.4 Data Fusion Results

Data fusion is applied to combine all types and modalities of datasets. Before all data is combined, there are several steps to be applied to ensure that all the data is comparable and has the same range, $[0,1]$. Firstly, the degree of suspicion is calculated in credit card data using equation (6.1). Secondly, the disagreement between credit card and loyalty card data is calculated (see equation 6.3 & 6.4). Thirdly, the degree of suspicion is calculated based on the distance between person's car and store location (see equation 6.5). This study set the distance = 555 metres from the car park to the store location. Fourthly, the classification accuracy for face image data is applied. Finally, all the data is sum up and divide by the number of datasets (see equation 6.9). After getting the fusion for every data, then, the result is sorting in ascending order to see which data are most suspicious. Table 8.7 shows the results of data fusion. The first row of data is the suspicious data, where the degree of suspicion using a credit card is 0.999055. Degree of suspicion based on the credit card disagreement is 0.999298367. Finally, degree of suspicion based on the gender of the suspected person taking into account the uncertainty due to gender accuracy is 0.0967, same for the age accuracy is 0.8017, and degree of suspicion in term of distance between the store location and the person's car is 0.998971. The overall data fusion result is 0.779145, which is the highest for any transaction made. From this result, the original data shows that on 13th January 2014 at 19.20pm, the suspect spent \$10 000, with no credit card point recorded. The gender is male, age is 24, and the distance between the person's car and the store location is 1.6 km (see figure 8.24). The distance reveals that this person is not in the location when the

card is being used. Then, the day and the time at the location is examined, and it shows that there is another person at the location. This person is suspected to be the one who misused the credit card. The overall degree of suspicion is just over 78%, which is somewhat less than 100% mostly because the gender of the person who used it and was on the surveillance cameras is the same as the real owner.

Table 8.7: Results of data fusion

No	Degree of Suspicion					
	Credit Card	Loyalty Card	Age	Gender	Distance	Total
1	0.999055	0.999298367	0.8017	0.0967	0.998971	0.779145
2	0.350594	0.489662559	0.1983	0.0967	0.999993	0.42705
3	0.307303	0.447548007	0.1983	0.0967	0.965699	0.40311
4	0.275661	0.415541258	0.1983	0.0967	0.928997	0.38304
5	0.005113	0.004701328	0.8017	0.0967	0.998307	0.381304
6	0.005961	0.004701328	0.8017	0.0967	0.996663	0.381145
7	0.02854	0.014683478	0.8017	0.0967	0.901165	0.368558
8	0.026451	0.003722793	0.8017	0.0967	0.879321	0.361579
9	8.41E-05	0.019266556	0.8017	0.0967	0.873554	0.358261
10	0.393363	0.006760843	0.1983	0.0967	0.999993	0.339023
.
.
1400	1.14E-05	4.74405E-05	0.1983	0.0967	4.36E-05	0.05902

The result from the second row in table 8.7 shows an anomaly in the credit card, loyalty card and distance. This person spent \$277.26, and his loyalty card is 0 points. This shows that he is not using his loyalty card. This person is male and 51 years old. His location at the shop can be considered too far at 2.03 km (see figure 8.26). Then, result no.3 shows that the person is male and 45 years old. The anomaly was detected based on the distance from his location to the store which is 1.8 km (see figure 8.27). The amount that this person used with the credit card is \$256.24, which is not suspicious as compared with the first result, but suspicious if the first result is removed.

Another example of analysis is that if the gender of the person using the card is different, then the results will significantly change. Per this case study, for most of the cases, the gender matches the gender of the owner. Therefore, this study intends to demonstrate that there will be a different result when the gender is different. Table 8.7 shows four cases in

which the gender of the user is not the same as the one of the owner. Data line 1 shows an increase of the total degree of suspicion after the data fusion from 0.779145 to 0.950455. Another example is in data no 2, 3 and 4, the total degree of suspicion after the data fusion rises from 0.42705 to 0.59837, 0.40311 to 0.574087 and 0.38304 to 0.55365, respectively. This shows that the difference in the gender can help in finding suspicious cases.

Figure 8.9 shows another example of data when weight is added. Weight can be added and adjusted by human experts. They can set the weights based on the importance of the data. In this example, the study assumes that gender is most important because if the gender is different that apparently, it is suspicious. Therefore, in this example gender is set to the higher weight which is 0.6. Then, for credit card and age data, weight is set to 0.2 and 0.1 respectively. For the other two variables, loyalty card and distance location, the weight is set to 0.05. After calculating the weighted degree of suspicion, the most suspicious case is still the same data as 1 for which the degree is 0.921874468. Then, the degree of suspicion for lines 2, 3 and 4 rises from 0.59837 to 0.706411578, 0.574087 to 0.69393295 and 0.55365 to 0.684169113, respectively.

Figure 8.28 presents the comparison of the number of data to be process and number of 10 most suspicious fused data. Before we get the most suspicious fused data to be investigated, there are many data to process, and it is time-consuming and tedious. Using our proposed method can help investigator or expert to simplify all the data and sort based on the level of suspicion. From these three analyses, it is obvious that every data item has been simplified and helpful for the investigator to shorten the time in investigating fraud cases or other crime or anomalous behaviour.

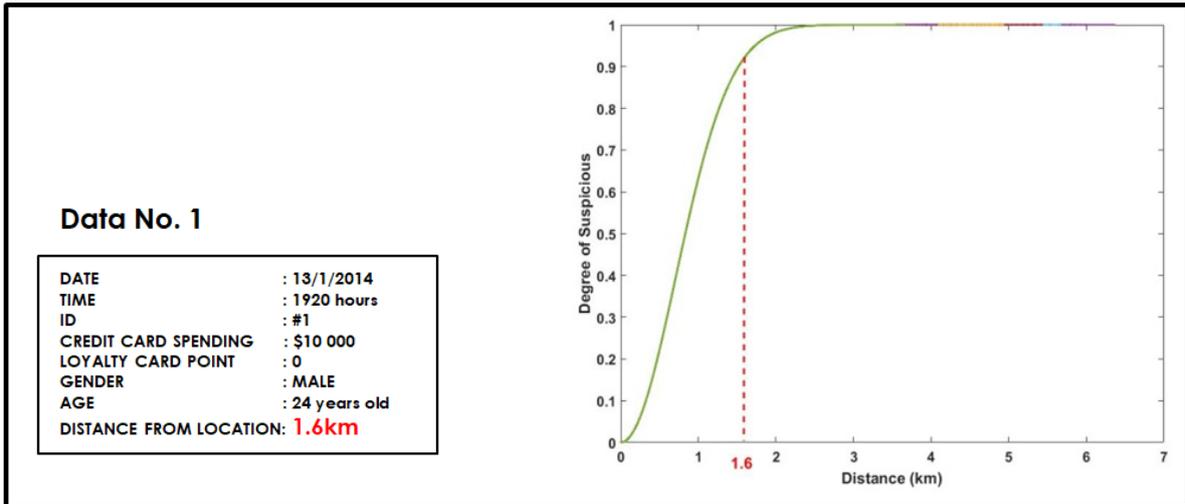


Figure 8.25: Details on data no.1

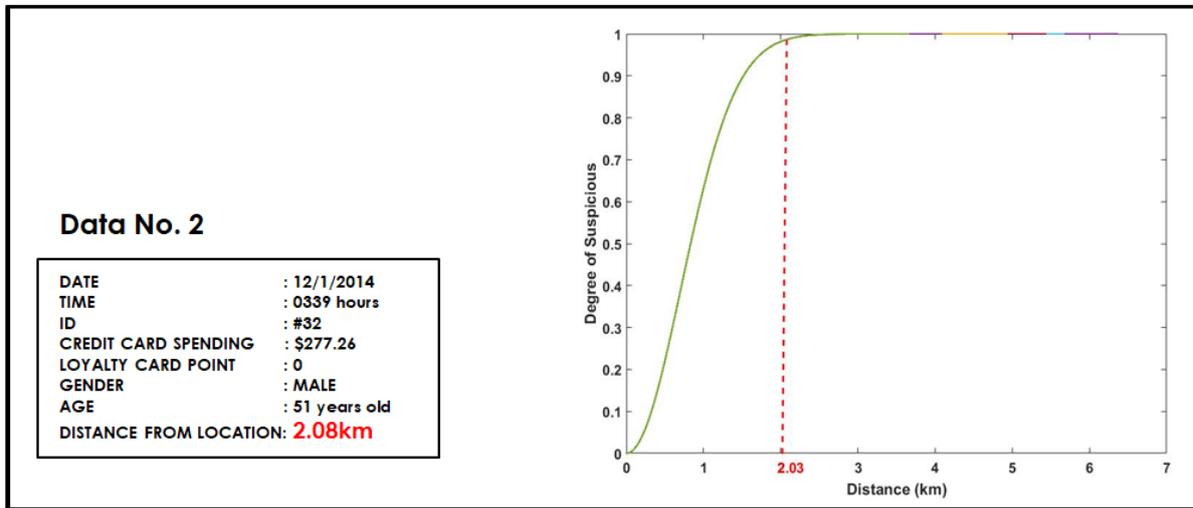


Figure 8.26: Details on data no.2

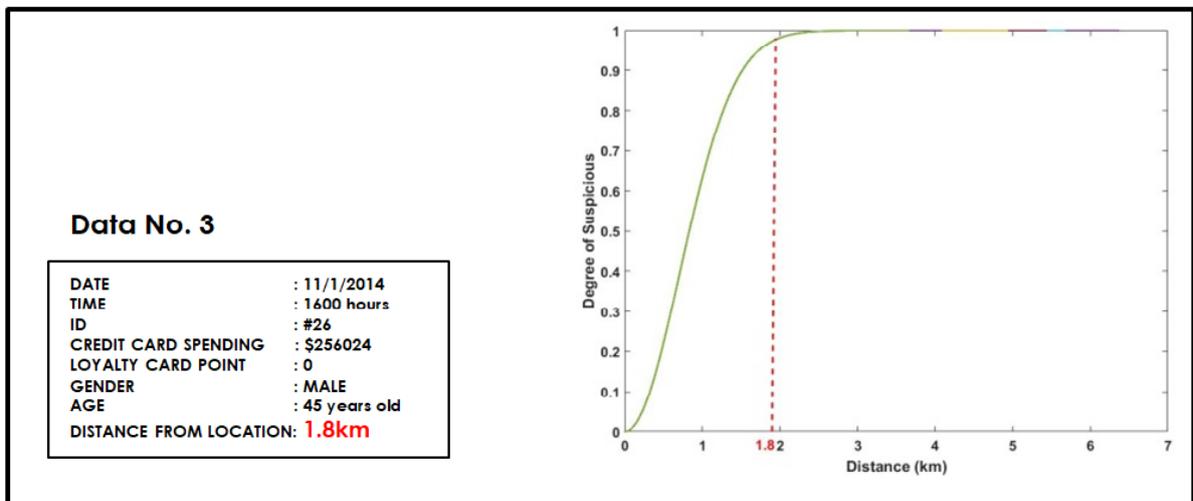


Figure 8.27: Details on data no.3

Table 8.8: Example of several cases of different gender

No	Degree of Suspicion					Total
	Credit Card	Loyalty Card	Age	Gender	Distance	
1	0.999298	0.999055	0.8017	0.9033	0.998971	0.950455
2	0.489663	0.350594	0.1983	0.9033	0.999993	0.59837
3	0.447548	0.307303	0.1983	0.9033	0.965699	0.574087
4	0.415541	0.275661	0.1983	0.9033	0.928997	0.55365
5	0.004701	0.005113	0.8017	0.0967	0.998307	0.391287
6	0.004701	0.005961	0.8017	0.0967	0.996663	0.391112
7	0.014683	0.02854	0.8017	0.0967	0.901165	0.377569
8	0.003723	0.026451	0.8017	0.0967	0.879321	0.370372
9	0.019267	8.41E-05	0.8017	0.0967	0.873554	0.366996
10	0.006761	0.393363	0.1983	0.0967	0.999993	0.349023
.
.
1400	0.000190569	4.74405E-05	0.1983	0.0967	0.00000151	0.059048

Table 8.9: Weighted Total

No	Degree of Suspicion					Weighted Total
	Credit Card	Loyalty Card	Age	Gender	Distance	
1	0.049965	0.199811	0.08017	0.54198	0.0499486	0.921874468
2	0.024483	0.0701188	0.01983	0.54198	0.0499997	0.706411578
3	0.022377	0.0614606	0.01983	0.54198	0.048285	0.69393295
4	0.020777	0.0551322	0.01983	0.54198	0.0464499	0.684169113
5	0.000338	0.0786726	0.01983	0.05802	0.0499997	0.206860292
6	0.000235	0.0010226	0.08017	0.05802	0.0499154	0.189363016
7	0.000235	0.0011922	0.08017	0.05802	0.0498332	0.189450416
8	0.000734	0.005708	0.08017	0.05802	0.0450583	0.189690424
9	0.000186	0.0052902	0.08017	0.05802	0.0439661	0.18763239
10	0.000963	0.00001682	0.08017	0.05802	0.0436777	0.182847848
.
.
1400	0.000190569	4.74405E-05	0.1983	0.0967	0.00000151	0.077891

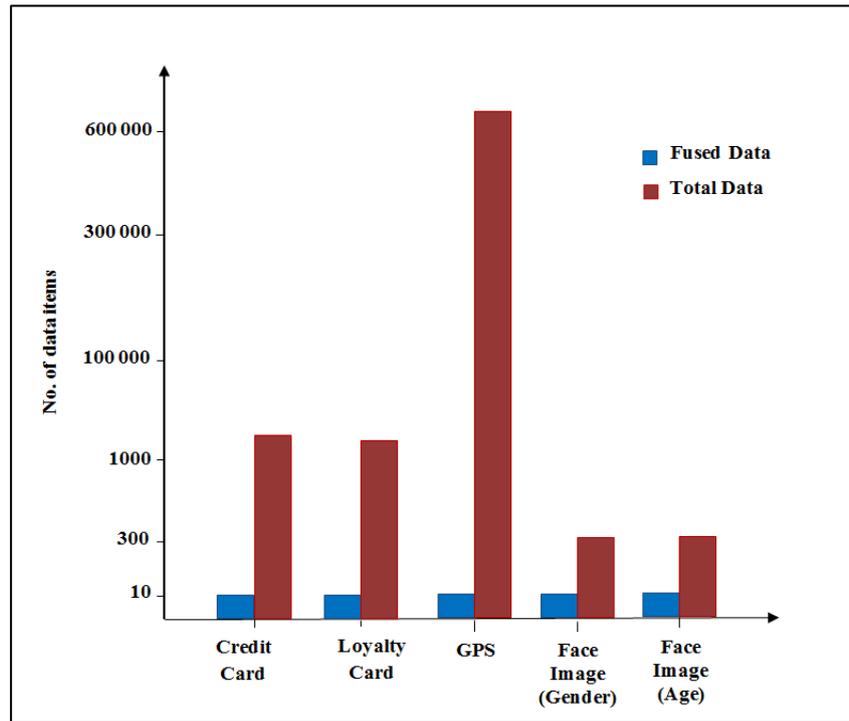


Figure 8.28: Comparison between original data and fused data

8.5 Summary

Result and analysis of this study are presented in this chapter. There are three sections of results presented, which are anomaly detection, image classification and data fusion. Anomaly detection shows satisfactory results which help finding the suspicious person in this case study. This study also achieves high accuracy on image classifications for age and gender. Data fusion helps by simplifying the huge amount of data and assists in making the final overall decision.

Evaluation

9.1 Introduction

This chapter discusses evaluation part of anomaly detection in this study. Anomaly detection is applied to three different datasets which are a credit card, loyalty card and GPS data. This chapter will discuss the comparison of anomaly detection based on the distribution of data and statistical technique.

9.2 Comparison of anomaly detection

In this study, anomaly detection is applied to the VAST challenge data. This data has no labels, and therefore only unsupervised anomaly detection can be applied to the dataset. Anomaly detection can be categorised into supervised, semi-supervised and unsupervised. An evaluation of unsupervised anomaly detection is not as straightforward as supervised anomaly detection. Supervised anomaly detection can be evaluated using receiving operating characteristics (ROC) or precision-recall (PR) [156]. For example, in classification, an instance classified wrongly is a mistake. Unsupervised anomaly detection is different because the data do not have any labels. In a real situation, most data do not have a label, and it is costly for experts to label the data manually. Therefore, more unlabelled data is available. In

this situation, this study compares Gaussian distribution, Chebyshev inequality and EDA techniques. Gaussian distribution or normal distribution can be achieved when the number of data is huge for some data. Examples include student heights in school, income in one country such as the United Kingdom or student examination marks. Gaussian distribution can be calculated as follows:

$$f(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (9.1)$$

where μ = mean, σ = standard deviation, σ^2 = variance. Figure 9.1 shows an example of the normal distribution of UK income from 2012 to 2013. Very often, however the nature of the data does not follow Gaussian law, e.g. rainfall data.

Chebyshev inequality is applied to the data with any distribution. In most situations in the real world, the distribution of data is not normal. Therefore, Chebyshev inequality can be applied to find anomalies. In Chebyshev inequality, no more than $\frac{1}{n^2}$ of the data samples/points are more than $n\sigma$ away from the mean (where σ denotes the standard deviation).

Based on the Gaussian distribution and Chebyshev inequality, this study has compared numbers of outliers detected using these two techniques and proposed technique in this study. There are a different number of standard deviation is demonstrated to the different datasets. The table has 8 columns starting with n where n denotes the sigma or standard deviation, column number 2 is labelled as A is percentage of Gaussian distribution on different n ; B in column number 3 is number of anomalies detected based on the Gaussian distribution; column C is percentage of Chebyshev inequality on different n ; D is the number of anomalies detected based on the Chebyshev Inequality; E is number of anomalies detected; F is percentage of anomalies detected in regards to the theoretical max, according to the

Chebyshev inequality theory and last column; and G is percentage of anomalies detected in regards to the total number of data points (see table 9.1).

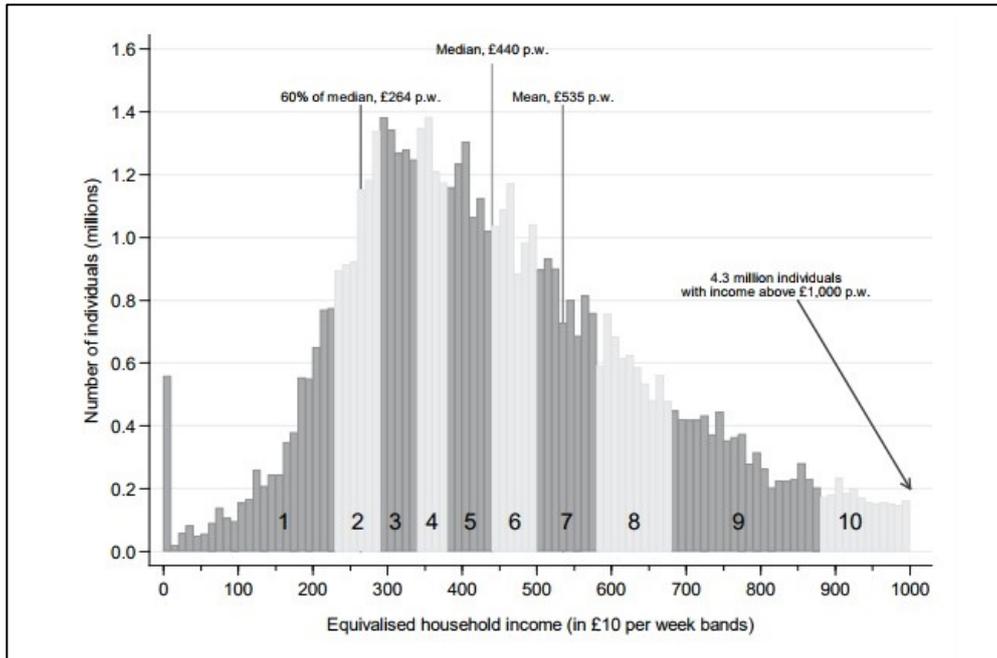


Figure 9.1: Distribution of UK income from 2012 to 2013 [157]

Table 9.2 shows a comparison of anomalies detected using credit card data. There are 1382 data points in credit card data. There are five different sigma have been applied which are 2, 3, 4, 5 and 6. In Gaussian distribution, the highest number of anomalies detected is 62.88 which is 4.55% using two sigma. After applying three sigma, only 4.15 data is anomalous at 0.3%. Four, five and six sigma detected the same number of anomalies which is 1.38, or 0.1% of the total data. Based on Chebyshev inequality, the highest percentage of anomalies will be detected using 2 sigma which is 25% of the total data (345.5 anomalies). The lowest number of anomalies detected is using 6 sigma which is 3% of the data (38.30 anomalies). Three sigma will detect 11% of anomalies or 153.55 anomalies. There are 55.28 and 38.39 anomalies detected after applying 5 and 6 sigma which is 4% and 3% of the data respectively. A number of outliers detected when $n = 2$ is only two and only one anomalies detected when $n = 3, 4, 5$ and 6. The highest percentage of anomalies detected regarding the

theoretical max, according to the Chebyshev inequality, is 2.6% anomalies based on 6 sigma. The lowest percentage of anomalies detected in regards to the theoretical max, according to the Chebyshev inequality is 0.58% based on 2 sigma. Starting from 3 sigma, the percentage of anomalies detected regarding the theoretical max, according to the Chebyshev inequality is 0.65% and double when applying 4 sigma (1.21%) and triple from 3 sigma when applying 5 sigma (1.8%). The last column explains the percentage of anomalies detected in regards to the total number of data points. The highest percentage is 0.14% which is 2 sigma and 0.7% when applying 3, 4, 5 and 6 sigma.

Table 9.1: Labels in the comparison table

Label	Description
<i>n</i>	Sigma/standard deviation
A	Gaussian distribution
B	Number of outliers based on the Gaussian distribution
C	Chebyshev Inequality ($1/n^2$) %
D	Number of outliers based on the Chebyshev Inequality
E	Number of outliers detected
F	% of outliers detected in regards to the theoretical max, according to the Chebyshev Inequality Theory
G	% of outliers detected in regards to the total number of data points.

Table 9.2: Comparison of anomalies detected using credit card data

<i>n</i>	A	B	C	D	E	F	G
2	4.55%	62.88	25%	345.5	2	0.58%	0.14%
3	0.3%	4.15	11%	153.55	1	0.65%	0.07%
4	0.1%	1.38	6%	82.92	1	1.21%	0.07%
5	0.1%	1.38	4%	55.28	1	1.8%	0.07%
6	0.1%	1.38	3%	38.39	1	2.6%	0.07%

Table 9.3 provides a comparison of anomalies detected using credit card data after removing one of the obvious data which is spending is \$10 000. Therefore, the number of data points is 1381. Gaussian distribution in column A and number of outliers are the same as table 9.2. The highest number of anomalies is 62.83 with a sigma of 2, and the lowest is 1.38 which is sigma = 4, 5 and 6 (0.1%). If the sigma = 3, then the anomalies detected are 4.15 (0.3%). Column C is using Chebyshev inequality, and the result based on sigma value is the

same with table 9.2. There is slightly different with table 9.2 because a number of data points is minus 1. Therefore, the highest anomalies are 345.25 when sigma is 2, followed by 151.91 anomalies (sigma = 3). The lowest anomalies detected are 41.43 when sigma is 6. Sigma = 4 and 5 detected 82.86 and 55.24 anomalies based on Chebyshev inequality respectively. The real number anomalies of detected are in column E which show that sigma = 2 and number of anomalies is 70 (the highest anomalies detected). The lowest anomalies detected are only 1 using 6 sigma and 2 anomalies using 5 sigma. There are 45 and 20 anomalies detected when sigma is set to 3 and 2, respectively. Then, the real anomalies detected are compared to the number of anomalies detected using Chebyshev inequality. The highest percentage of anomalies detected regarding the theoretical maximum, according to Chebyshev inequality, is 29.6% (sigma = 3) and the lowest percentage is 2.4% (sigma = 6). Percentage of anomalies detected in regards to the theoretical maximum, according to Chebyshev inequality when sigma = 2, 4 and 5 is 20.2%, 24.14% and 3.6%. The last column shows the percentage of anomalies detected regarding the total number of data points. The highest percentage of anomalies detected is 5.07% (sigma = 2) and the lowest percentage of anomalies detected is 0.07% (sigma = 6). Percentage of anomalies detected based on sigma = 3, 4 and 5 are 3.26%, 1.75% and 0.14% respectively.

Table 9.3: Comparison of anomalies detected using credit card data (without 10000 spending data)

<i>n</i>	A	B	C	D	E	F	G
2	4.55%	62.83	25%	345.25	70	20.2%	5.07%
3	0.3%	4.14	11%	151.91	45	29.6%	3.26%
4	0.1%	1.38	6%	82.86	20	24.1%	1.75%
5	0.1%	1.38	4%	55.24	2	3.6%	0.14%
6	0.1%	1.38	3%	41.43	1	2.4%	0.07%

Table 9.4 provides a comparison of anomalies detected based on loyalty card data which is 1285 data points. There are 5 different number of sigma are compared in this table same

with table 9.2 and 9.3 which are 2, 3, 4, 5 and 6 (column n). Column A and C is the same value with column A and C in table 9.2 and 9.3. The highest and the lowest number of anomalies detected based on Gaussian distribution is 58.47 (sigma = 2 (4.55%)) and 1.29 (sigma = 4, 5 and 6 (0.1%)). When the sigma is set to 3, the percentage of Gaussian distribution is 0.3%, and a number of anomalies detected are 3.85. Based on Chebyshev inequality, the highest and the lowest number of anomalies detected are 321.25 (sigma = 2 (25%)) and 38.46 (sigma = 6 (3%)). After setting the sigma value to 3, the percentage of data based on Chebyshev inequality is 11%, and a number of anomalies detected is 141.35. There are 77.1 (6%) and 51.4 (4%) anomalies detected based on Chebyshev inequality when sigma is set to 4 and 5 respectively. The highest number and the lowest number of detected anomalies in loyalty card data are 70 (sigma = 2) and 1 (sigma = 5, 6) respectively. After setting the sigma value to 3 and 4, the numbers of anomalies detected are 51 and 24 anomalies. Then, the number of detected anomalies is compared with Chebyshev inequality theory. The highest percentage is 36% (sigma = 3) and the lowest percentage is 1.9% (sigma = 5). When a number of sigma is set to 2, 4 and 6, the percentages of anomalies detected regarding theoretical maximum, according to Chebyshev inequality, reach 21.8%, 31.13% and 2.6%. The percentages of anomalies detected in regard to the total number of data points are 5.44% (sigma = 2), 3.97% (sigma = 3), 1.87% (sigma = 4), and 0.07% (sigma = 5, 6).

Table 9.4: Comparison of anomalies detected using loyalty card data

n	A	B	C	D	E	F	G
2	4.55%	58.47	25%	321.25	70	21.8%	5.44%
3	0.3%	3.85	11%	141.35	51	36%	3.97%
4	0.1%	1.29	6%	77.1	24	31.1%	1.87%
5	0.1%	1.29	4%	51.4	1	1.9%	0.07%
6	0.1%	1.29	3%	38.46	1	2.6%	0.07%

The final GPS dataset has been processed regarding time, distance and ratio of trajectory angle. Table 9.5 present comparison of anomalies detected based on 469 data points in this

dataset. When sigma is set to 2, the percentage of Gaussian distribution is 4.55%, and Chebyshev inequality is 25%, and numbers of detected anomalies are 21.34 and 117.25 respectively. The real number of detected anomalies based on 2 sigma is 22 (4.69%). 0.47 anomalies are detected in Gaussian distribution based on 4, 5 and 6 sigma (0.1%). According to Chebyshev inequality, when sigma is set to 4, 5 and 6, the number of outliers detected are 28.14 (6%), 18.76 (4%) and 14.07 (3%) anomalies. In a real situation, the number of anomalies detected after setting sigma to 4, 5 and 6 are 9 (1.92%), 4 (0.85%) and 2 (0.14%) anomalies respectively. The percentages of outliers regarding theoretical maximum, according to the Chebyshev inequality after sigma, are set to 4, 5 and 6 are 31.98%, 21.3% and 14.2%. When sigma is set to 3, Gaussian distribution is 4.55% (0.42 anomaly detected), and Chebyshev inequality is 11% (51.59 anomalies detected). The real numbers of anomalies detected based on 3 sigma are 18 (3.83%) and 34.8% of anomalies detected regarding the Chebyshev inequality.

Table 9.5: Comparison of anomalies detected using time, distance and ratio of trajectory angle

<i>n</i>	A	B	C	D	E	F	G
2	4.55%	21.34	25%	117.25	22	18.7%	4.69%
3	0.3%	0.42	11%	51.59	18	34.8%	3.83%
4	0.1%	0.47	6%	28.14	9	31.98	1.92%
5	0.1%	0.47	4%	18.76	4	21.3%	0.85%
6	0.1%	0.47	3%	14.07	2	14.2%	0.14%

9.3 Summary

This chapter discusses the comparison of the applied method in this study with other methods. It begins with anomaly detection, followed by feature extraction in face images classification. Data distribution and statistical techniques are compared with anomaly

detection part. The result shows that the proposed method which is EDA produces better results than statistical methods.

Conclusion

10.1 Conclusion

This study demonstrates the application of heterogeneous data by identifying anomaly detection in the first stage, then applying data fusion technique to combine the heterogeneous data. The illustrative example of using VAST Challenge data [17] and GAFace dataset shows the application of data fusion technique on heterogeneous data. This demonstrates five types of data, namely, a credit card, loyalty card, GPS, age, and gender based on face images. The anomaly is detected from the credit card data, loyalty card and GPS data. Anomaly detection applied a new method for the autonomously detecting of anomalous behaviour based on heterogeneous data. The method applied EDA to detect anomalies in various heterogeneous data sets.

The next step is detecting gender and age classification using a pre-trained deep CNN architecture and SVM classifier. This classification is applied in the automatic system for detecting anomalous human behaviour. This work is a continuous study from a previous stage on the use of heterogeneous data for autonomous detection of anomalous human behaviour. Face images are a type of data which can improve the evidence in investigating a suspect or an event. This method will reduce the scope in identifying suspicious persons in the forensic investigation area. The first step in the methodology involves extracting the features from

face images using a pre-trained deep learning convolutional network. Then, SVM is applied to classify the images. Transfer learning was then used to successfully utilise the already learnt knowledge for a new task with the limited dataset. Even though, when the dataset is small, the results are much better than using traditional handcrafted features like Haar. The proposed method was tested on the GAFace dataset, and the results are very encouraging: the percentages of accuracy for gender and age classifications are 90.33% and 80.17% respectively, which demonstrates that the solution is feasible.

After analysing results from anomaly detection, there is one person who is consistent in all the anomalies. After this data is removed, then a new anomaly is detected, but after the details of this person are checked, it shows that he is the CEO of the company. Therefore, it is normal for him to spend more than the other staff members. Then, data on trajectory and location was analysed, and the suspicious person was automatically discovered. To confirm the suspect, face recognition can be applied to support information from the EDA. Anomalous results and image classification results are combined through the data fusion technique. The result from fusion technique is then ranked ascendingly to examine which is the suspicious data. After analysing the result, a new suspicious person is discovered. Location distance between the staff members and the store is far (1.6 km). Then, the study concludes that he is not at the store when the credit card is charged. Analysis of this result can assist human expert in simplifying their job and helping them in making a decision.

10.2 Limitations and Challenges

This study applies heterogeneous data in one application or situation. Heterogeneous data needs a variety of datasets with different data modalities. In this study, the limitation is only one case study is applied to show the heterogeneous data in a complete situation. Most of the research always applied one dataset for one application. For example, in image

processing, there is only image data which is ready to be classified or detected depends on the research aim such as detecting identity. Financial data generally for processing something related to finance or economy only. Intrusion detection data only applied to networking problems. Therefore, it is hard to find datasets that share the same situation. In a real situation, there are a variety of datasets which finally can be combined or integrated to produce a result or decision.

The most challenging part of this study was to get a case study and data that demonstrates heterogeneous data. Many public databases offer only one dataset for one application or one research study. For example, if the researcher wants to research classification, there are many databases which offer data on classifying the data. However, in this study, full data and information are needed to show the application of heterogeneous data. This study found fictional competition dataset and adds another dataset to simulate heterogeneous data. Dealing with heterogeneous data is challenging work. Every data modality has a different way to process. For example, processing numerical data is more straightforward than processing signals or images. In this study, the most challenging part is the pre-processing especially the GPS data. The data is big because it takes every second of the car movement. A few steps need to be taken before processing to get other variables such as the distance, speed and direction. It takes almost six months to pre-process the data before the data is ready to be processed for the next step.

Image data also involves challenging work, such as finding the best datasets and pre-processing the images. Finding the suitable dataset that matches the problem is really hard. In this study, set of images that represent people at the counter is required to simulate the process of buying something. The size of the image is also small. This is a standard problem in face recognition when taking images from the surveillance camera. A similar image can be found in the Google Images, but it has a problem to label the age. Many benchmark data offer

the features of age and gender label but when the data have been downloaded then so many steps to pre-process. This study does not cover in detail image processing and face recognition. Therefore, it will take time to find the best datasets with labels, followed by pre-processing and finally classifying the images.

10.3 Futures Works

This research can be continued in the future adding more data modalities such as text or streaming data. Then, more data modalities can show more heterogeneous data. Text data can be short text messaging such as text from iMessage or WhatsApp applications. Much information can be gathered from text data, especially when dealing with criminal cases. Recently, many research applying anomaly detection to the text data to find the suspicious things. Example of streaming data is real-time data in which the data is created continuously. Normally streaming data is acquired from sensors or in financial is stock market data. This data is challenging and will be more fascinating using heterogenous data combined with data fusion. Another work can be done for this study is applications to more datasets. If the data can be gathered from the investigation department, then the finding will be more interesting. Real situation data can demonstrate the propose of the research study. More data or cases means that better findings can be acquired.

In the future, more data modalities can be added to the framework for example in this study if there is text data it will show more heterogeneous data. Text data may include text messaging or email. Much information can be found in the text data especially when dealing with the criminal cases. Research has also applied anomaly detection to text data to find something suspicious. Ultimately, the proposed research is intended to help the authorities or investigation departments. This research can help such departments handle the big data,

integrate all data modalities, and produce results that may finally assist investigators in finding suspicious persons.

References

- [1] M. Tistarelli, E. Grosso, and D. Meuwly, “Biometric in Forensic Science: Challenges, Lessons and New Technologies,” in *Biometric Authentication*, vol. 8897, V. Cantoni, D. Dimov, and M. Tistarelli, Eds. Cham: Springer International Publishing, 2014, pp. 153–164.
- [2] S. F. Pratama, L. Pratiwi, A. Abraham, and A. K. Muda, “Computational Intelligence in Digital Forensics,” in *Computational Intelligence in Digital Forensics: Forensic Investigation and Applications*, vol. 1, Springer, 2014, pp. 1–16.
- [3] D. Dessimoz and C. Champod, “Linkages between Biometrics and Forensic Science,” *Handb. Biometrics*, pp. 425–459, 2008.
- [4] P. Angelov, “Anomaly Detection based on Eccentricity Analysis,” in *2014 IEEE Symposium on Evolving and Autonomous Learning Systems (EALS)*, 2014, pp. 1–8.
- [5] F. Mone, C. Walsh, C. Mulcahy, C.J. Macmahon, S. Farrell, Aoife.M, R. Segurado, R. Mahony, S. Higgins, S. Carroll, P.Mcparland and F.M. Mcauliffe, “Prenatal detection of structural cardiac defects and presence of associated anomalies : a retrospective observational study of 1262 fetal echocardiograms,” no. 1, 2015.
- [6] Y. Kim and A. Kogan, “Development of an Anomaly Detection Model for a Bank’s Transitory Account System,” *J. Inf. Syst.*, vol. 28, no. 1, pp. 145–165, 2014.
- [7] B. S. J. Costa, P. P. Angelov, and L. A. Guedes, “Real-time fault detection using recursive density estimation,” *J. Control. Autom. Electr. Syst.*, vol. 25, no. 4, pp. 428–437, 2014.
- [8] R. Zuech, T. M. Khoshgoftaar, and R. Wald, “Intrusion detection and Big Heterogeneous Data: a Survey,” *J. Big Data*, vol. 2, no. 1, pp. 1–41, 2015.
- [9] F. Nater, “Abnormal Behavior Detection in Surveillance Videos,” 2012.
- [10] A. Mohd Ali, P. Angelov, and X. Gu, “Detecting Anomalous Behaviour Using Heterogeneous Data,” in *Advances in Computational Intelligence Ssystems: Contributions Presented at the 16th UK Workshop on Computational Intelligence, September 7–9, 2016, Lancaster, UK*, 2016, pp. 253–273.
- [11] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, “Multisensor data fusion : A review of the state-of-the-art,” *Inf. Fusion*, vol. 14, no. 1, pp. 28–44, 2013.

- [12] D. Perez, I. Maza, F. Caballero, D. Scarlatti, E. Casado, and A. Ollero, “A ground control station for a Multi-UAV surveillance system: Design and validation in field experiments,” *J. Intell. Robot. Syst. Theory Appl.*, vol. 69, pp. 119–130, 2013.
- [13] D. Lahat, T. Adali, and C. Jutten, “Multimodal Data Fusion : An Overview of Methods , Challenges , and Prospects,” *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.
- [14] A. Mohd Ali and P. Angelov, “Applying Computational Intelligence to Community Policing and Forensic Investigations,” in *Community Policing - A European Perspective*, 2017, pp. 1–16.
- [15] X. Wang, A. Mohd Ali, and P. Angelov, “Gender and Age Classification of Human Faces for Automatic Detection of Anomalous Human Behaviour,” in *International Conference on Cybernetics (CYBCONF 2017)*, 2017, pp. 1–6.
- [16] A. M. Ali and P. Angelov, “Anomalous behaviour detection based on heterogeneous data and data fusion,” *Soft Comput.*, no. Dhar 2013, 2018.
- [17] “VAST Challenge 2014,” 2014. [Online]. Available: [vacommunity.org/VAST Challenge 2014](http://vacommunity.org/VAST_Challenge_2014).
- [18] A. Bera, D. Bhattacharjee, and M. Nasipuri, “Hand Biometrics in Digital Forensics,” *Comput. Intell. Digit. ForensicsForensic Investig. Appl.*, vol. 1, pp. 145–163, 2014.
- [19] A. Guarino, “Digital Forensics as A Big Data Challenge,” *ISSE 2013 Secur. Electron. Bus. Process.*, pp. 197–203, 2013.
- [20] D. Quick and K. R. Choo, “Impacts of increasing volume of digital forensic data : A survey and future research challenges,” *Digit. Investig.*, vol. 11, no. 4, pp. 273–294, 2014.
- [21] F. R. Mitchell, “An Overview of Artificial Intelligence Based Pattern Matching in a Security and Digital Forensic Context,” *Cyberpatterns*, pp. 215–222, 2014.
- [22] A. . Awad and E. . Hassanien, “Impact of some biometric modalities on Forensic Science,” *Comput. Intell. Digit. ForensicsForensic Investig. Appl.*, vol. 1, pp. 47–62, 2014.
- [23] M. Sawadkar, N. Keni, and R. Agarwal, “Machine Learning Based Gender Detection Using Craniometric Analysis,” vol. 6, no. 7, pp. 1773–1776, 2016.
- [24] D. Navega, R. Vicente, D. N. Vieira, A. H. Ross, and E. Cunha, “Sex estimation from the tarsal bones in a Portuguese sample : a machine learning approach,” *Int. J. Leg. Med.*, vol. 129, pp. 651–659, 2015.
- [25] Z. Ghasem, I. Frommholz, and C. Maple, “A Machine Learning Framework to Detect And Document Text-based Cyberstalking,” no. October, pp. 348–355, 2015.
- [26] K. Reynolds, A. Kontostathis, and L. Edwards, “Using Machine Learning to Detect Cyberbullying,” 2011.
- [27] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, “Computers in Human Behavior Cybercrime detection in online communications : The experimental case of

- cyberbullying detection in the Twitter network,” *Comput. Human Behav.*, vol. 63, pp. 433–443, 2016.
- [28] J. Bunk, J.H. Bappy, T.M. Mohammed, L. Nataraj, B.S. Manjunath, S. Chandrasekaran and A.M. Roy-chowdhury, “Detection and Localization of Image Forgeries using Resampling Features and Deep Learning,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1881–1889.
- [29] B. Bayar and M. C. Stamm, “A Deep Learning Approach To Universal Image Manipulation Detection Using A New Convolutional Layer,” in *IH & MMSec*, 2016, pp. 5–10.
- [30] N. S. Monson and M. Kumar, “Behaviour Knowledge Space-Based Fusion for Image Forgery Detection,” in *International Conference on Inventive Communication and Computational Technologies (ICICCT 2017)*, 2017, pp. 400–403.
- [31] M. Qi, Y. Liu, L. Lu, J. Liu, and M. Li, “Big Data Management in Digital Forensics,” *2014 IEEE 17th Int. Conf. Comput. Sci. Eng.*, pp. 238–243, 2014.
- [32] S. Kaisler, F. Armour, J. a Espinosa, and W. Money, “Big Data: Issues and Challenges Moving Forward,” *46th Hawaii Int. Conf. Syst. Sci.*, pp. 995–1004, 2013.
- [33] R. Kruse, C. Borgelt, F. Klawonn, C. Moewes, M. Steinbrecher, and P. Held, *Computational intelligence : A Methodological Introduction*. Springer-Verlag London, 2013.
- [34] C. Bitter, J. North, D. Elizondo, and T. Watson, “An introduction to the use of neural networks for network intrusion detection,” pp. 5–24, 2012.
- [35] F. N. Sibai, H. I. Hosani, R. M. Naqbi, S. Dhanhani, and S. Shehhi, “Iris recognition using artificial neural networks,” *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5940–5946, 2011.
- [36] M. Negnevitsky, *Artificial intelligence.*, Second., vol. 180, no. 8. Pearson Education Limited, 2005.
- [37] H. Roubos and M. Setnes, “Compact Fuzzy Models and Classifiers Optimization,” in *The Practical Handbook of Genetic Algorithms Applications*, Second., L. Chambers, Ed. Chapman & Hall/CRC, 2001, pp. 69–97.
- [38] M. S. Hoque, M. A. Mukit, and M. Ab. N. Bikas, “An Implementation of Intrusion Detection System using Genetic Algorithm,” *Int. J. Netw. Secur. Its Appl.*, vol. 4, no. 2, pp. 109–120, 2012.
- [39] S. Al Amro, D. Elizondo, and A. Solanas, “Evolutionary Computation in computer security and cryptography,” *Comput. Intell. Priv. Secur.*, vol. 23, no. 3, pp. 25–34, 2012.
- [40] L. A. Zadeh, “Fuzzy sets,” *Inf. Control*, vol. 8, no. 3, pp. 338–353, 1965.
- [41] S. Al Amro, F. Chiclana, and D. A. Elizondo, “Application of fuzzy logic in computer security and forensics,” *Stud. Comput. Intell.*, vol. 394, pp. 35–49, 2012.

- [42] B. Stahl, M. Carroll-Mayer, D. Elizondo, K. Wakunuma, and Y. Zheng, “Intelligence techniques in computer security and forensics: At the boundaries of ethics and law,” *Stud. Comput. Intell.*, vol. 394, pp. 237–258, 2012.
- [43] Yee Ling Boo and Daminda Alahakoon, “Building Multi-modal Crime Profiles with Growing Self Organizing Maps,” *Comput. Intell. Digit. Forensics Forensic Investig. Appl.*, vol. 1, pp. 97–124, 2014.
- [44] J. Garcia Rodriguez, A. Angelopoulou, F. J. Mora-Gimeno, and A. Psarrou, “Building visual surveillance systems with neural networks.,” pp. 181–198, 2012.
- [45] N. L. Beebe and L. Liu, “Clustering digital forensic string search output,” *Digit. Investig.*, vol. 11, no. 4, pp. 314–322, 2014.
- [46] P. Angelov, *Autonomous Learning Systems*. John Wiley and Sons, Ltd, Publication, 2012.
- [47] J. Macías-Hernández and P. Angelov, “Applications of Evolving Intelligent Systems to Oil and Gas Industry,” *Evol. Intell. Syst. Methodol. Appl.*, pp. 401–421, 2010.
- [48] B. Sielly, C. G. Bezerrat, L. A. Guedes, P. P. Angelov, C. Natal, and Z. Norte, “Online Fault Detection Based on Typicality and Eccentricity Data Analytics,” in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–6.
- [49] X. Zhou and P. Angelov, “An Approach to Autonomous Self-localization of a Mobile Robot in Completely Unknown Environment using Evolving Fuzzy Rule-based Classifier,” no. Cisd, pp. 131–138, 2007.
- [50] R. Ramezani, P. Angelov, and X. Zhou, “A Fast Approach to Novelty Detection in Video Streams using Recursive Density Estimation,” pp. 2–7, 2008.
- [51] P. Angelov, P. Sadeghi-Tehran, and R. Ramezani, “An approach to automatic real-time novelty detection, object identification, and tracking in video streams based on recursive density estimation and evolving Takagi-Sugeno fuzzy systems,” *Int. J. Intell. Syst.*, vol. 26, no. 3, pp. 189–205, Mar. 2011.
- [52] O. Salem, A. Guerassimov, A. Marcus, and B. Furht, “Sensor Fault and Patient Anomaly Detection and Classification in Medical Wireless Sensor Networks,” in *IEEE ICC 2013 - Selected Areas in Communications Symposium*, 2013, pp. 4373–4378.
- [53] J. A. Iglesias, P. Angelov, A. Ledezma, and A. Sanchis, “Modelling Evolving User Behaviours,” in *Evolving and Self-Developing Intelligent Systems, 2009*, 2009, pp. 1–5.
- [54] J. A. Iglesias, P. Angelov, A. Ledezma, and A. Sanchis, “Creating evolving user behavior profiles automatically,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 854–867, 2012.
- [55] P. Angelov, “Evolving Fuzzy Systems,” *Comput. Complex. Theory, Tech. Appl.*, vol. 2, no. 2, pp. 1053–1065, 2012.
- [56] R. Dutta Baruah and P. Angelov, “Evolving local means method for clustering of streaming data,” *IEEE Int. Conf. Fuzzy Syst.*, pp. 10–15, 2012.

- [57] R. D. Baruah and P. Angelov, “DEC: Dynamically evolving clustering and its application to structure identification of evolving fuzzy models,” *IEEE Trans. Cybern.*, vol. 44, no. 9, pp. 1619–1931, 2014.
- [58] R. Hyde and P. Angelov, “A Fully Autonomous Data Density Based Clustering Technique,” in *Evolving and Autonomous Learning Systems (EALS), 2014 IEEE*, 2014, pp. 116–123.
- [59] N. Kasabov, *Evolving Connectionist Systems: Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines*. Springer, 2002.
- [60] P. Angelov and A. Kordon, “Adaptive inferential sensors based on evolving fuzzy models,” *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 40, no. 2, pp. 529–539, 2010.
- [61] P. Angelov, X. Zhou, and F. Klawonn, “Evolving fuzzy rule-based classifiers,” *Comput. Intell. Image Signal Process. 2007. CIISP 2007. IEEE*, pp. 220–225, 2007.
- [62] P. Angelov, “Outside the Box: An Alternative Data Analytics Framework,” *J. Autom. Mob. Robot. Intell. Syst.*, vol. 8, no. 2, pp. 53–59, 2014.
- [63] P. Angelov, D. P. Filev, and N. Kasabov, *Evolving Intelligent Systems: Methodology and Applications*. John Wiley and Sons, Ltd, Publication, 2010.
- [64] P. Angelov and R. Yager, “Simplified fuzzy rule-based systems using non-parametric antecedents and relative data density,” *Evol. Adapt. Intell. Syst. (EAIS), 2011 IEEE*, pp. 62–69, 2011.
- [65] I. Skrjanc, S. Blazic, and P. Angelov, “Robust Evolving Cloud-based PID Control Adjusted by Gradient Learning Method,” in *Evolving and Adaptive Intelligent Systems (EAIS), 2014 IEEE*, 2014, pp. 1–8.
- [66] P. Angelov, I. Škrjanc, and S. Blažič, “A Robust Evolving Cloud-Based Controller,” in *Springer Handbook of Computational Intelligence*, 2015, pp. 1435–1449.
- [67] S. . Sangeetha and M. Karnan, “Survey of Biometric Systems using Iris Recognition,” vol. 6, no. 2, pp. 40–44, 2014.
- [68] A. K. Jain, P. Flynn, and A. K. Ross, *Handbook of Biometrics*. Springer Science & Business Media, 2007, 2008.
- [69] A. K. Jain, A. Ross, and S. Prabhakar, “An Introduction to Biometric Recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 1–29, 2004.
- [70] A. K. Jain, A. A. Ross, and K. Nandakumar, *Introduction to Biometrics*, vol. 1. Springer, 2011.
- [71] T. Kiertscher, R. Fischer, and C. Vielhauer, “Latent Fingerprint Detection using a Spectral Texture Feature,” in *Thirteenth ACM Multimedia Workshop on Multimedia and Security (MM&Sec '11)*, 2011, pp. 27–32.
- [72] W. Horng, C. Lee, and C. Chen, “Classification of Age Groups Based on Facial Features,” vol. 4, no. 3, pp. 183–192, 2001.

- [73] A. M. Bukar and D. Connah, “Automatic age and gender classification using supervised appearance model,” *J. Electron. Imaging*, vol. 25, no. 6, pp. 1–11, 2016.
- [74] G. Levi and T. Hassner, “Age and Gender Classification using Convolutional Neural Networks,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 34–42.
- [75] G. S. Chandel and A. Bhargava, “Identification of People by Iris Recognition,” vol. 14, no. 3, pp. 77–80, 2014.
- [76] I. Bouchrika, M. Goffredo, J. Carter, and M. S. Nixon, “On Using Gait in Forensic Biometrics,” *J. Forensic Sci.*, vol. 56, no. 4, pp. 882–889, 2011.
- [77] S. Pal, U. Pal, and M. Blumenstein, “Signature-Based Biometric Authentication,” *Comput. Intell. Digit. Forensics Forensic Investig. Appl.*, vol. 1, pp. 285–314, 2014.
- [78] S. M. Abu-soud, “An Enhanced Non-Static Biometric Keystroke Dynamics Model for Strengthening the Information Content Security,” vol. 4, no. 5, pp. 219–232, 2014.
- [79] P. S. Teh, A. B. J. Teoh, C. Tee, and T. S. Ong, “Keystroke dynamics in password authentication enhancement,” *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8618–8627, Dec. 2010.
- [80] M. Karnan, M. Akila, and N. Krishnaraj, “Biometric personal authentication using keystroke dynamics: A review,” *Appl. Soft Comput.*, vol. 11, no. 2, pp. 1565–1573, Mar. 2011.
- [81] P. H. Pisani and A. C. Lorena, “A systematic review on keystroke dynamics,” *J. Brazilian Comput. Soc.*, vol. 19, pp. 573–587, 2013.
- [82] M. Nakada, H. Wang, and D. Terzopoulos, “AcFR: Active Face Recognition Using Convolutional Neural Networks,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops AcFR:*, 2017, pp. 35–40.
- [83] Y. Wen, K. Zhang, Z. L. B, and Y. Qiao, “A Discriminative Feature Learning Approach,” in *ECCV 2016, PART 7, LNCS 9911*, vol. 1, B. Leibe and et.al, Eds. Springer International Publishing, 2016, pp. 499–515.
- [84] A. Ucar, Y. Demir, and C. Guzelis, “A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering,” *Neural Comput. Appl.*, no. 27, pp. 131–142, 2016.
- [85] C. C. Charalambous and A. A. Bharath, “A data augmentation methodology for training machine / deep learning gait recognition algorithms,” in *proceedings of the British Machine Vision Conference (BMVC)*, 2016, pp. 1–12.
- [86] F. Reynard and P. Terrier, “Determinants of gait stability while walking on a treadmill: a machine learning approach,” *arXiv:1705.05191*, no. April, pp. 1–11, 2017.
- [87] T. Shirakawa, N. Sugiyama, H. Sato, and K. Sakurai, “Gait analysis and machine learning classification on healthy subjects in normal walking,” *Int. J. Parallel, Emergent Distrib. Syst.*, vol. 31, no. 5, pp. 543–552, 2017.

- [88] M. K. Bashar, I. Chiaki, and H. Yoshida, “Human Identification from Brain EEG signals Using Advanced Machine Learning Method,” in *IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES) Human*, 2016, pp. 475–479.
- [89] J. K. Johannesen, J. Bi, R. Jiang, J. G. Kenney, C. A. Chen, and P. Service, “Machine learning identification of EEG features predicting working memory performance in schizophrenia and healthy adults,” *Neuropsychiatr. Electrophysiol.*, pp. 1–21, 2016.
- [90] D. Sanchez, P. Melin, O. Castillo, and F. Valdez, “Modular granular neural networks optimization with multi-objective hierarchical genetic algorithm for human recognition based on iris biometric,” *2013 IEEE Congr. Evol. Comput. CEC 2013*, pp. 772–778, 2013.
- [91] D. Sánchez and P. Melin, “Optimization of modular granular neural networks using hierarchical genetic algorithms for human recognition using the ear biometric measure,” *Eng. Appl. Artif. Intell.*, vol. 27, pp. 41–56, 2014.
- [92] V. Mai, I. Khalil, and C. Meli, “ECG biometric using multilayer perceptron and radial basis function neural networks,” *2011 Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 2745–2748, 2011.
- [93] M. Beltrán, P. Melin, and L. Trujillo, “Development of modular neural networks with fuzzy logic response integration for signature recognition,” *Fuzzy Inf. Eng.*, vol. 1, no. 4, pp. 345–355, 2009.
- [94] G. Tiwari, B. Gour, R. Kori and A. U. Khan, “Fast Image Retrieval Method based on Controlled Self Organization Map Neural Network on Biometric Feature,” vol. 105, no. 15, p. 8887, 2014.
- [95] S. Chu, “Kernel-Based Machine Learning for Biometrics Computing : Frameworks and Methods,” vol. 1, no. 1, pp. 47–55, 2014.
- [96] A. Joshi, S. Bhushan, and M. J. Kaur, “Gait Recognition of Human using SVM and BPNN Classifiers,” vol. 3, no. 1, pp. 281–290, 2014.
- [97] E. Gianaria, M. Grangetto, M. Lucenteforte, and N. Balossino, “Human Classification Using Gait Features,” *Biometrics*, pp. 16–27, 2013.
- [98] D. Boyadzieva and G. Gluhchev, “Neural Network and kNN Classifiers for On-line Signature Verification,” *Biometrics*, pp. 198–206, 2014.
- [99] IDC, “Where in the World is Storage: A Look at Byte Density Across The Globe,” 2013.
- [100] V. Dhar, “Data Science and Prediction,” *Commun. ACM*, vol. 56, no. 12, pp. 64–73, 2013.
- [101] Ernst & Young, “Forensic Data Analytics,” 2013.
- [102] H. V. Jagadish *et al.*, “Big data and its Technical Challenges,” *Commun. ACM*, vol. 57, no. 7, pp. 86–94, 2014.
- [103] D. Turcsany, A. Bargiela, and T. Maul, “Local Receptive Field Constrained Deep

- Networks,” *Inf. Sci. (Ny)*, vol. 349–350, pp. 229–247, 2016.
- [104] B. J. C. Principe and R. Chalasani, “Cognitive Architectures for Sensory Processing,” *Proceeding IEEE*, vol. 102, no. 4, 2014.
- [105] S. Maldonado and G. L’Huillier, “SVM-Based Feature Selection and Classification for Email Filtering,” *Pattern Recognit. - Appl. Methods*, vol. 204, pp. 1–11, 2013.
- [106] P. Angelov and P. Sadeghi-Tehran, “A Nested Hierarchy of Dynamically Evolving Clouds for Big Data Structuring and Searching,” *Procedia - Procedia Comput. Sci.*, vol. 53, pp. 1–8, 2015.
- [107] C. L. Borgman, *Scholarship in the Digital Age: Information, Infrastructure and the Internet*. The MIT Press, 2007.
- [108] J.V. Atanasoff, “Advent of Electronic Digital Computing,” *IEEE Ann. Hist. Comput.*, vol. 6, no. 3, pp. 229–282, 1984.
- [109] L. Mearian, “Data Storage: Then and Now,” *Computerworld*, 2014.
- [110] R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE Publications Ltd, 2014.
- [111] H. Liu, S. Shah, and W. Jiang, “On-line outlier detection and data cleaning,” vol. 28, pp. 1635–1647, 2004.
- [112] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection,” *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.
- [113] E. Lughofer and P. Angelov, “Handling drifts and shifts in on-line data streams with evolving fuzzy systems,” *Appl. Soft Comput. J.*, vol. 11, no. 2, pp. 2057–2068, 2011.
- [114] H. Om and A. Kundu, “A hybrid system for reducing the false alarm rate of anomaly intrusion detection system,” *2012 1st Int. Conf. Recent Adv. Inf. Technol. RAIT-2012*, pp. 131–136, 2012.
- [115] B. Delgado, K. Tahboub, and E. J. Delp, “Automatic Detection of Abnormal Human Events on Train Platforms,” in *IEEE National Aerospace and Electronics Conference*, 2014, no. 2009, pp. 169–173.
- [116] Y. Wu, A. Patterson, R. D. C. Santos, and N. L. Vijaykumar, “Topology Preserving Mapping for Maritime Anomaly Detection,” pp. 313–326, 2014.
- [117] M. Goldstein, “Anomaly Detection in Large Datasets,” University of Kaiserslautern, 2014.
- [118] B. G. Amidan, T. A. Ferryman, and S. K. Cooley, “Data outlier detection using the chebyshev theorem,” *IEEE Aerosp. Conf. Proc.*, vol. 2005, pp. 3–8, 2005.
- [119] V. Chandola, A. Banerjee, V. Kumar, J. Hawkins, P.N. Tan, M. Steinbach and V. Kumar, “Introduction to Data Mining,” *Introd. to Data Min.*, vol. 41, no. September, pp. 1–18, 2014.
- [120] T. V Pollet and L. van der Meij, “To Remove or not to Remove: the Impact of Outlier

- Handling on Significance Testing in Testosterone Data,” *Adapt. Hum. Behav. Physiol.*, pp. 1–18, 2016.
- [121] F. Palmieri, U. Fiore, and A. Castiglione, “A distributed approach to network anomaly detection based on independent component analysis,” no. June 2013, pp. 1113–1129, 2014.
- [122] A. Abdallah, M. A. Maarof, and A. Zainal, “Fraud detection system: A survey,” *J. Netw. Comput. Appl.*, vol. 68, pp. 90–113, 2016.
- [123] V. Jyothisna, “A Review of Anomaly based Intrusion Detection Systems,” *Int. J. Comput. Appl.*, vol. 28, no. 7, pp. 975–8887, 2011.
- [124] D. Malekian and M. R. Hashemi, “An adaptive profile based fraud detection framework for handling concept drift,” in *2013 10th International ISC Conference on Information Security and Cryptology, ISCISC 2013*, 2013, pp. 1–6.
- [125] E. Keogh, J. Lin, A. W. Fu, and H. Van Herle, “Finding Unusual Medical Time-Series Subsequences : Algorithms and Applications,” vol. 10, no. 3, pp. 429–439, 2006.
- [126] W. Li, V. Mahadevan, and N. Vasconcelos, “Anomaly detection and localization in crowded scenes,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, 2014.
- [127] G. Pallotta, M. Vespe, and K. Bryan, “Framework for Anomaly Detection and Route Prediction,” *Entropy*, vol. 15, pp. 2218–2245, 2013.
- [128] M. Liggins II, D. Hall, and J. Llinas, *Handbook of Multisensor Data Fusion: Theory and Practice, Second Edition*, Second Edi. CRC Press, 2017.
- [129] H. Boström, S.F. Andler, M. Brohede, R. Johansson, A. Karlsson, J.V. Laere, L. Niklasson, M. Nilsson, A. Persson and T. Ziemke, “On the Definition of Information Fusion as a Field of Research,” *IKI Tech. Reports*, pp. 1–8, 2007.
- [130] D. L. Hall, M. McNeese, J. Llinas, and T. Mullen, “A framework for dynamic hard/soft fusion,” *Proc. 11th Int. Conf. Inf. Fusion, FUSION 2008*, 2008.
- [131] F. Castanedo, “A Review of Data Fusion Techniques,” *Sci. World J.*, vol. 2013, 2013.
- [132] R. C. Luo, C.-C. Yih, and K. L. Su, “Multisensor fusion and integration: approaches, applications, and future research directions,” *IEEE Sens. J.*, vol. 2, no. 2, pp. 107–119, 2002.
- [133] Sonia, M. Singh, R. D. Baruah, and S. B. Nair, “A Voting-Based Sensor Fusion Approach for Human Presence Detection,” in *Intelligent Human Computer Interaction: 8th International Conference, IHCI 2016, Pilani, India, December 12-13, 2016, Proceedings*, A. Basu, S. Das, P. Horain, and S. Bhattacharya, Eds. Cham: Springer International Publishing, 2017, pp. 195–206.
- [134] E. Fotiadis, M. Garzón, and A. Barrientos, “Human Detection from a Mobile Robot Using Fusion of Laser and Vision Information,” *Sensors*, vol. 13, no. 9, pp. 11603–11635, 2013.

- [135] T. Adali, Y. Levin-Schwartz, and V. D. Calhoun, “Multimodal Data Fusion Using Source Separation: Application to Medical Imaging,” *Proc. IEEE*, vol. 103, no. 9, pp. 1494–1506, 2015.
- [136] R. Singh and A. Khare, “Fusion of Multimodal Medical Images using Daubechies Complex Wavelet Transform-A Multiresolution Approach,” *Inf. Fusion...*, vol. 2, pp. 49–60, 2012.
- [137] D. Izadi, J. Abawajy, S. Ghanavati, and T. Herawan, “A Data Fusion Method in Wireless Sensor Networks,” *Sensors*, vol. 15, no. 2, pp. 2964–2979, 2015.
- [138] X. Luo, D. Zhang, L. T. Yang, J. Liu, X. Chang, and H. Ning, “A kernel machine-based secure data sensing and fusion scheme in wireless sensor networks for the cyber-physical systems,” *Futur. Gener. Comput. Syst.*, vol. 61, pp. 85–96, 2016.
- [139] P. Angelov, X. Gu, and D. Kangin, “Empirical Data Analytics,” *Int. J. Intell. Syst.*, vol. 0, pp. 1–24, 2017.
- [140] P. Angelov, G. Xiaowei, D. Kangin, and J. Principe, “Empirical Data Analysis: A New Tool for Data Analytics,” in *IEEE International Conference on Systems, Man, and Cybernetics*, 2016, pp. 52–59.
- [141] P. Angelov, “Typicality Distribution Function - A New Density - based Data Analytics Tool,” in *IJCNN 2015 International Joint Conference on Neural Networks*, 2015, pp. 1–8.
- [142] P. Angelov, J. Victor, A. Dourado, and D. Filev, “On-line evolution of Takagi-Sugeno fuzzy models,” 2004.
- [143] P. Angelov and R. Buswell, “Evolving Rule-based Models: A Tool for Intelligent Adaptation,” vol. 2, pp. 1062–1067, 2001.
- [144] M. Zabłocki, K. Go, D. Frejlichowski, and R. Hofman, “Intelligent video surveillance systems for public spaces – a survey,” *J. Theor. Appl. Comput. Sci.*, vol. 8, no. 4, pp. 13–27, 2014.
- [145] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nat. Int. Wkly. J. Sci.*, vol. 521, pp. 436–444, 2015.
- [146] L. Shao, F. Zhu, and X. Li, “Transfer learning for visual categorization: A survey,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, 2015.
- [147] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” in *European Conference on Computer Vision*, 2014, pp. 818–833.
- [148] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems*, 2012, pp. 1–9.
- [149] P. Van Der Waerden and H. Timmermans, “Car drivers ’ characteristics and the maximum walking distance between parking facility and final destination,” *J. Transp. Land Use*, vol. 10, no. 1, pp. 1–11, 2017.

- [150] A. Asuncion and D. Newman, “UCI Machine Learning Repository,” 2007. [Online]. Available: <http://archive.ics.uci.edu/ml/datasets.html>. [Accessed: 19-Jul-2017].
- [151] Y. U. Zheng, “Trajectory Data Mining : An Overview,” vol. 6, no. 3, pp. 1–41, 2015.
- [152] P. Sadeghi-Tehran and P. Angelov, “A Real-time Approach for Novelty Detection and Trajectories Analysis for Anomaly Recognition in Video Surveillance Systems,” in *2012 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 2013, pp. 108–113.
- [153] P. C. Papageorgiou, M. Oren, and T. Poggio, “A General Framework for Object Detection,” in *International Conference on Computer Vision*, 1998, pp. 555–562.
- [154] P. Viola, O. M. Way, and M. J. Jones, “Robust Real-Time Face Detection,” vol. 57, no. 2, pp. 137–154, 2004.
- [155] M. Oualla, A. Sadiq, and S. Mbarki, “A Survey of Haar-Like Feature Representation,” in *2014 International Conference on Multimedia Computing and Systems (ICMCS)*, 2014, pp. 1–6.
- [156] N. Goix, “How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms?,” 2016.
- [157] S. P. Jenkins, “The income distribution in the UK: a picture of advantage and disadvantage,” *ISER Work. Pap. Ser.*, no. February, 2015.