# Fuzzy based binary feature profiling for modus operandi analysis

M. A. P. Chamikara[1, 2*], A. Galappaththi[1], Y.P.R.D. Yapa[1, 2], R.D. Nawarathna[1, 2], S. R. Kodituwakku[1, 2], J. Gunatilake[1, 3], A.A.C.A. Jayathilake[4] , L.H. Liyanage[5]

[1]Postgraduate Institute of Science, University of Peradeniya, Sri Lanka,

[2]Department of Statistics and Computer Science, University of Peradeniya, Sri Lanka,

[3]Department of Geology, University of Peradeniya, Sri Lanka,

[4]Department of Mathematics, University of Peradeniya, Sri Lanka,

[5]School of Computing, Engineering and Mathematics, University of Western Sydney, Australia.

Corresponding author:

M.A.P. Chamikara,
No.22, Nikaketiya, Menikhinna, 20000, Sri Lanka
Email address: pathumchamikara@gmail.com

## Abstract

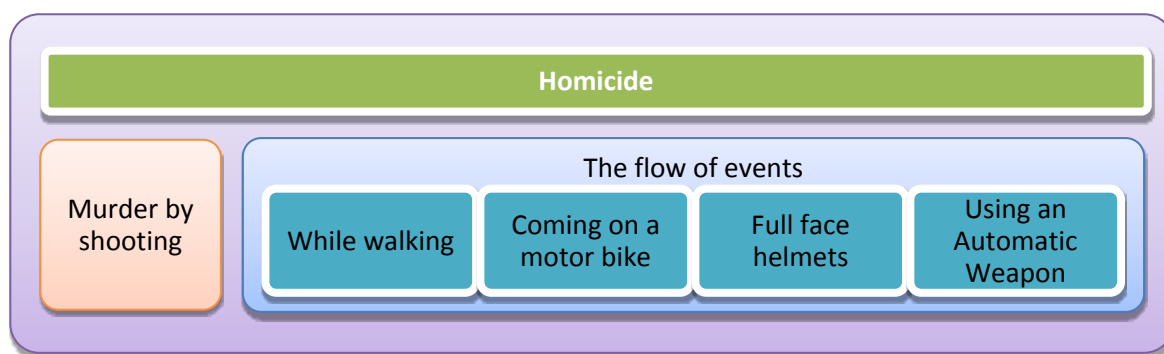
29  It is a well-known fact that some criminals follow perpetual methods of operations, known as
30  modi operandi. Modus operandi is a commonly used term to describe the habits in committing
31  crimes. These modi operandi are used in relating criminals to crimes for which the suspects have
32  not yet been recognized. This paper presents the design, implementation and evaluation of a
33  new method to find connections between crimes and criminals using modi operandi. The method
34  involves in generating a feature matrix for a particular criminal based on the flow of events of
35  his/her previous convictions. Then, based on the feature matrix, two representative modi
36  operandi are generated: complete modus operandi and dynamic modus operandi. These two
37  representative modi operandi are compared with the flow of events of the crime at hand, in order
38  to generate two other outputs: completeness probability (CP) and deviation probability (DP). CP
39  and DP are used as inputs to a fuzzy inference system to generate a score which is used in
40  providing a measurement for the similarity between the suspect and the crime at hand. The
41  method was evaluated using actual crime data and ten other open data sets. In addition,
42  comparison with nine other classification algorithms showed that the proposed method
43  performs competitively with other related methods proving that the performance of the new
44  method is in an acceptable level.

## Introduction

46  Scientists have long played a role in examining deviant behavior in society. "Deviance behaviour"
47  is a term used by scientists to refer to some form of "rule-breaking" behaviour [1]. It can be the
48  behaviour of violating a social norm or the law. Criminal behaviour is also a form of deviance, one
49  that is defined as the breaking of legal rules. Nevertheless, there is a difference between deviance
50  and crime. Deviance involves breaking a norm and evoking a negative reaction from others. Crime
51  is a deviance that breaks a law, which is a norm stipulated and enforced by government bodies
52  [1]. However, crimes affect the society negatively. Therefore, law enforcement authorities take
53  necessary actions to mitigate crimes in an environment where high crime frequencies are
54  observed each year. In this exercise the application of technology for crime analysis is being
55  widened in the world. Locard's Exchange principle states that every contact of the perpetrators
56  of a crime scene leaves a trace. The perpetrators will both bring something into the scene and
57  leave with something from the scene [2]. However, the cognitive abilities of criminals will always
58  make them minimize their risks of apprehension by conducting the perfect crime and maximizing
59  their gain [3]. Modus operandi or method of operation such as preparation actions, crime
60  methods and weapons are frequently used in criminal profiling because the past crime trends
61  show that, after criminals get used to a certain method of operation, they try to use the same
62  modus operandi in committing his/her next crime [4].

63  The criminals develop a set of actions during the performance of a series of crimes which we
64  refer to as "modus operandi" (MO). MO is developed with the crimes he/she commits and the
65  nature of trying to stick with the developed MO that has worked throughout the previous crimes
66  [5].  In any criminal career, the MO happens to evolve, no matter what the circumstances. Also,

67   it is a common behaviour that serial offenders tend to exhibit significant behaviour known as
68   his/her signature. Therefore, MOs of criminals play a major role in investigating crimes [5].  It is
69   a known fact that features such as criminal signature and physical appearance are used in crime
70   investigations in almost all the police departments around the world. Sri Lanka police also use
71   MOs of criminals to identify the suspects who have conducted crimes. Currently Sri Lanka Police
72   use a manual crime recording and investigation system.  This manual system has many problems
73   such as data redundancy, inefficiency, tediousness, inability to support crime investigation and
74   many other problems which are associated with a conventional manual system. To overcome
75   these problems, a web-based framework was proposed with geographical information support
76   containing a centralized database for crime data storage and retrieval, named SL-CIDSS: Sri Lanka
77   Crime Investigation Decision Support System [6]. The proposed system accompanies a collection
78   of data mining algorithms which effectively support the crime investigation process. Fuzzy based
79   binary feature profiling (BFPM) for modus operandi analysis is one novel algorithm which is
80   integrated with the system to provide an effective way to find the similarity between crimes and
81   criminals.

82   According to the penal code of Sri Lanka first enacted in 1882 and amended subsequently several
83   times in later years [7], Sri Lanka police classifies crimes into two categories: Grave crimes and
84   Minor offences. Until 2014, grave crimes were classified under 21 crime categories and in 2015
85   another 5 new crime categories were introduced, making it 26 categories of grave crime types.
86   Kidnapping, Fraud or mischief causing damage greater than 25000 rupees, Burglary, Grievous
87   hurt, Hurt by sharp weapon, Homicide, Rape, Robbery, Cheating by trust, Theft are 10 of the most
88   frequent crime types.  To identify the patterns involved in crimes, a collection of subtypes were
89   identified under these 26 crime types.  These subtypes have been created mainly for the purpose
90   of modus operandi analysis. Most frequent behaviors of criminals/crimes are considered as crime
91   subtypes. When a crime is logged in the Grave Crime Record (GCR) book, it is classified under one
92   of the 26 main categories. But, under the section of "nature of crime" in the GCR book, the police
93   officers record the flow of the crime incident including the subtypes.

100   **Figure 1.** *Relationship between main crime type, subtypes and crime flows*

101   A subtype is a sub category of one of the main crime types. For investigation, the nature of the
102   crime is broken into subtypes and flows according to their frequency of occurrence and
103   uniqueness. These sub categorizations have been introduced mainly to minimize the broadness
104   of main type and to improve clarity. Figure 1. depicts the relationship of the subtypes and flows

105  where there can be a flow of events to a crime recorded as one of the 26 main crime types. For
106  the simplicity and easy handling of data, the investigators have provided subtype codes and flow
107  codes. The flow of events provides a modus operandi which is most of the time unique to an
108  offender. Each subtype is provided with a code under the main type, to make the crime
109  investigation process easier.  For example, ROB/S001 denotes a subtype that is Highway robbery;
110  here ROB denotes the main type under which the corresponding subtype appears. In this case, it
111  is Robbery. Crime types are further subdivided into sub types to make the analysis and processing
112  simpler. In this manner, crime subtypes and flows have been identified under all the 26 crime
113  types. The space for adding more subtypes and flows under these crime types exists. A new
114  subtype or a flow is introduced to a particular main crime, if the same subtype or the flow
115  happens to persist for a prolonged time.

116  This paper proposes a novel method of criminal profiling using modus operandi which can be
117  used to identify associations between crimes and criminals. The method is based on a new
118  technique named, "binary feature vector profiling". Key relationships between a criminal and the
119  previous convictions are analyzed using binary feature profiling and association rule mining
120  techniques. Due to the impreciseness and vagueness of these extracted attributes, a fuzzy
121  inference system is used in making the final decision. The newly proposed method was adapted
122  into a classification algorithm in order to test its accuracy. An actual crime data set was used in
123  testing the performance of the newly proposed method and it was compared against nine well
124  established classification algorithms using ten open data sets. The results confirmed that the
125  proposed method produce competitive results compared to the other nine classification
126  algorithms.

127  The rest of the paper is organized as follows. Related work section presents a summary of the
128  work that has been conducted on modus operandi analysis as well as a brief discussion on crime
129  investigation using link analysis and association mining in general. Materials and Methods section
130  discusses the main steps of the newly proposed algorithm.  Next, Results and Discussion section
131  provides a validation and performance evaluation of the newly proposed method along with a
132  performance comparison with nine other classification algorithms. Finally, some concluding
133  remarks and future enhancements are outlined in the conclusion section.

## Related work

135  Literature shows many methods which have been developed in the area of automated crime
136  investigation. Our major concern has been laid upon the research carried out on crime
137  investigation using association mining as our research considers on developing a model to find
138  the associations between the criminals and the crimes depending on the modes operandi. C
139  Bennell and DV Canter [8] have proposed a method to use statistical models to test directly the
140  police practice of utilizing modus operandi to link crimes to a common offender. The results
141  indicated that certain features such as the distance between burglary locations, lead to high
142  levels of predictive accuracy. Craig Bennell, et. al. [9]  have tried to determine if readily available
143  information about commercial and residential serial burglaries, in the form of the offender's
144  modus operandi, provides a statistically significant basis for accurately linking crimes committed

145    by the same offenders. Benoit Leclerc, et al. [10] have reviewed the theoretical, empirical, and
146    practical implications related to the modus operandi of sexual offenders against children. They
147    have presented the rational choice perspective in criminology followed by descriptive studies
148    aimed specifically at providing information on modus operandi of sexual offenders against
149    children.

150    Clustering crimes, finding links between crimes, profiling offenders and criminal network
151    detection are some of the common areas where data mining is applied in crime analysis [11],
152    [12], [13] . Association analysis, classification and prediction, cluster analysis, and outlier analysis
153    are some of the traditional data mining techniques which can be used to identify patterns in
154    structured data. Offender profiling is a methodology which is used in profiling unknown criminals
155    or offenders. The purpose of offender profiling is to identify the socio-demographic
156    characteristics of an offender based on information available at the crime scene [14] [15].
157    Association rule mining discovers the items in databases which occur frequently and present
158    them as rules. Since this method is often used in market basket analysis to find which products
159    are bought with what other products, it can also be used to find associated crimes conducted
160    with what other crimes. Here, the rules are mainly evaluated by the two probability measures,
161    support and confidence [16], [17]. Association rule mining can also be used to identify the
162    environmental factors that affect crimes using the geographical references [18]. Incident
163    association mining and entity association mining are two applications of association rule mining.
164    Incident association mining can be used to find the crimes committed by the same offender and
165    then the unresolved crimes can be linked to find the offender who committed them. Therefore,
166    this technique is normally used to solve serial crimes like serial sexual offenses and serial
167    homicides [19].

168    Similarity-based association mining and outlier-based association mining are two approaches
169    used in incident association mining. Similarity-based association mining is used mainly to
170    compare the features of a crime with the criminal's behavioral patterns which are referred as
171    modus operandi or behavioral signature. In outlier-based association mining, crime associations
172    will be created on the fact that both the crime and the criminal have the possibility of having
173    some distinctive feature or a deviant behavior [20]. Entity association mining/link analysis is the
174    task of finding and charting associations between crime entities such as persons, weapons, and
175    organizations. The purpose of this technique is to find out how crime entities that appear to be
176    unrelated at the surface, are actually linked to each other [19]. Link analysis is also used as one
177    of the most applicable methods in social network analysis [21] in finding crime groups, gate
178    keepers and leaders [22].

179    Attribution can be used to link crimes to offenders. If two offences in different places involve the
180    same specific type, those may be readily attributed to the same offender [11].There are three
181    types of link analysis approaches, namely Heuristic-based, Statistical-based and Template- based
182    [19]. Sequential pattern mining is also a similar technique to association rule mining. This method
183    discovers frequently occurring items from a set of transactions occurred at different times [23].
184    Deviation detection detects data that deviates significantly from the rest of the data which is
185    analyzed. This is also called outlier detection, and is used in fraud detection [23] [24].

186  In classification, the data points will be assigned to a set of predefined classes of data by
187  identifying a set of common properties among them. This technique is often used to predict crime
188  trends. Classification needs a reasonably complete set of training and testing data since a high
189  degree of missing data would limit the prediction accuracy [23]. Classification comes under
190  supervised learning method [19], [25] which includes methods such as Bayesian models, decision
191  trees, artificial neural networks [26] and support vector machines. String comparison techniques
192  are used to detect the similarity between the records. Classification algorithms compare the
193  database record pairs and determine the similarity among them. This concept can be used to
194  avoid deceptive offender profiles. Information of offenders such as name, address, etc. might be
195  deceptive and therefore the crime database might contain multiple records of the same offender.
196  This makes the process of identification of their true identity difficult [23].

## Systems and Methods

198  This section provides a description about the systems and methods used in developing the fuzzy
199  based binary feature profiling for modus operandi analysis. First, an overview about how SL-
200  CIDSS captures the logics of modus operandi is explained. Then a detailed description about the
201  steps of the newly proposed algorithm is explained.

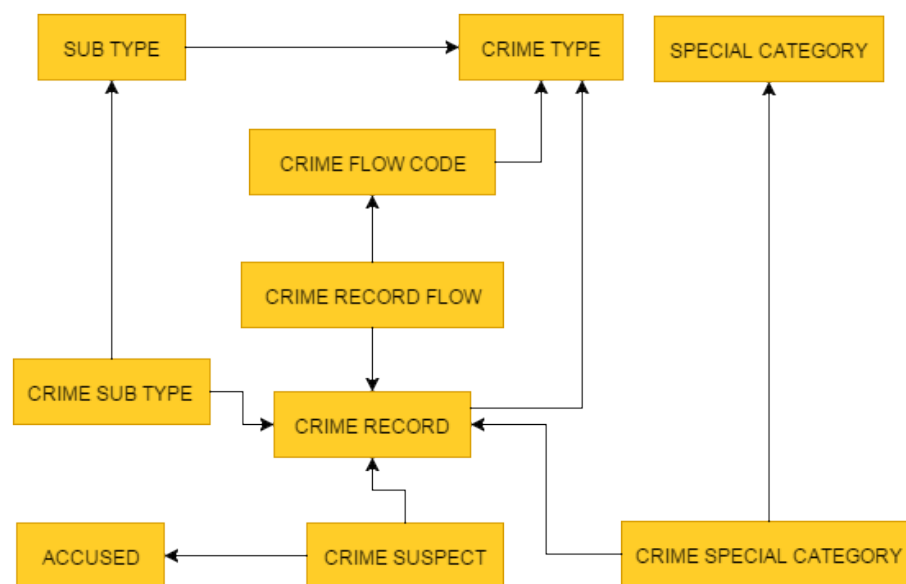



**Figure 2**. *Crime flow entity arrangement in SL-CIDSS*

212  Figure 2 shows how SL-CIDSS database captures the crime types and subtypes. A crime record
213  has a crime record flow. Typically, a crime is committed by a criminal and a particular accused
214  might commit one or more crimes. A CRIME RECORD can be of one the 26 crime types. A
215  particular CRIME RECORD will be considered under one main CRIME TYPE with the highest
216  precedence in the order of seriousness. For example, a crime incident that includes a murder and
217  a robbery will be categorized as a murder though a robbery has also taken place. But in the nature
218  of crime section, all crimes followed by the main type will be stated. Therefore, the CRIME

219  RECORD FLOW captures all the steps of the crime as a sequence of steps recorded. The crime
220  flows that have been previously registered are mapped under CRIME FLOW CODE. Also, a
221  particular CRIME RECORD instance can contain multiple SUB TYPES which are recorded as CRIME
222  SUB TYPE. The SPECIAL CATEGORY captures the crimes with special features such as crimes
223  occurring at the same location or retail shop. A crime may involve several special categories
224  which are saved in the CRIME SPECIAL CATEGORY. The ACCUSED entity records the information
225  of suspects and accused and they are related to crime through the CRIME SUSPECT entity.

226  As the first step of the newly employed method, a feature matrix is generated, resulting in a
227  binary matrix representing the crime flows. This binary feature matrix is composed of the binary
228  patterns generated on previous convictions of a particular criminal/suspect. This binary form of
229  the feature matrix provides a provision to direct application of computer algorithms with
230  methods such as Apriori based association rule mining. The reduced complexity of the binary
231  feature matrices provides an easy manipulation over the categorical and continuous valued
232  features. Figure 3 shows the steps of the proposed MO analysis algorithm.

233

234  **Step 1**: Generate the feature matrix.

235  **Step 2**: Generate the dynamic MOs (DMO) of the criminals.

236  **Step 3**: Generate the complete MO profile (CMOP) of the criminals.

237  **Step 4**: Find the deviation probability (DP) of CMOP from the crime MO under consideration
       (UMO).

238

239  **Step 5**: Find the completeness probability of UMO against DMO.

240  **Step 6**: Use the two values obtained from step 4 and 5 as inputs of a fuzzy Inference system to
       obtain the final similarity value (out of 100).

241  **Step 7**: Classify the UMO under the class with highest similarity score for validation.

242                    **Figure 3**. *Steps of the newly employed algorithm*

## 243  Generating the feature matrix

244  Table 1 shows how the feature vectors are generated and provides the way to generate modi
245  operandi of criminals as binary sequences. According to the table, events of the crime scene are
246  observed starting from its crime type. After a particular crime type is identified, the feature
247  vectors are updated with ones for each subtype and flow code that is available in the crime or
248  suspect's modus operandi. The vectors will be filled by zeros in places which the modus operandi
249  does not have any contact with. The column names to the feature matrix are generated in such
250  a way that it covers the collection of main types, sub types, crime flows and special categories at
251  hand. For example, if we consider the list of crime types, subtypes, crime flows and the special
252  category in Table 1, it results in 21-bit feature vectors as shown in the last two columns.

253    **Table 1**. *An instance of feature selection for the feature matrix generation*

| Main Semantic | Crime flow element code | Description | Suspect 1 | Suspect 2 |
|---|---|---|---|---|
| **Crime types** | HB | House Breaking | 0 | 1 |
| | HK | Hurt by Knife | 0 | 0 |
| | RB | Robbery | 1 | 0 |
| | TH | Theft | 0 | 0 |
| **Sub types** | ABD/S003 | Abduction from the legal guardian | 1 | 0 |
| | ABD/S004 | Abducting to marry | 0 | 0 |
| | ABD/S005 | Abducting for sexual harassment | 0 | 0 |
| | BGL/S004 | Use of stealth | 0 | 1 |
| | BGL/S011 | Burglary in business places | 0 | 0 |
| | ROB/S001 | Organized vehicle robbery | 1 | 0 |
| **Crime Flows** | BGL/F001 | Entering from the window | 0 | 1 |
| | BGL/F002 | Entering from the Fanlights | 0 | 0 |
| | BGL/F003 | Removing grills | 0 | 1 |
| | BGL/F004 | Breaking glasses | 0 | 0 |
| | ROB/F001 | Showing identity cards | 1 | 0 |
| | ROB/F003 | Wearing uniforms | 1 | 0 |
| | ROB/F004 | Robbery using identity cards, uniforms and chains | 0 | 0 |
| | ROB/F009 | Seizing inmates | 0 | 0 |
| | ROB/F010 | Appearing as CID officers | 0 | 0 |
| **Special Category** | Retailer 1 | Attacking/ robbing retailer 1's stores | 0 | 0 |
| | Retailer 2 | Attacking/ robbing retailer 2's stores | 0 | 0 |

254 In this manner we can produce binary MO patterns based on the crimes committed by different
255 criminals as shown in the last two columns of Table 1. According to Table 1, Suspect 1 has
256 committed a robbery with the subtypes, ABD/S003 (an abduction of a child from the legal
257 guardian), ROB/S001 (an organized vehicle robbery) and the flows, ROB/F001 (Identity cards have
258 been shown),ROB/F003 (accused has been wearing uniforms). Suspect 2 has committed a house
259 breaking with the sub type BGL/S004 (use of stealth), and the flows, BGL/F001 (Entering from the
260 window), BGL/F003 (Removing Grills).

261 Table 2 shows a feature matrix of binary patterns which is generated by considering the previous
262 convictions of suspect 1 assuming that he has conducted another robbery (conviction 2). *ct, st, fl*
263 and *sc* in Table 2 represent the abbreviations for "crime type", "sub type", "crime flow" and
264 "special category" respectively.

265 **Table 2**. *Feature matrix for Suspect 1, generated using the selected modus operandi attributes in Table 1.*

| | ct1 | ct2 | ct3 | ct4 | st1 | st2 | st3 | st4 | st5 | st6 | fl1 | fl2 | fl3 | fl4 | fl5 | fl6 | fl7 | fl8 | fl9 | sc1 | sc2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Conviction 1** | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Conviction 2** | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

266

## Generating the dynamic MOs (DMOs) of the criminals

268 Dynamic MO is a binary feature vector which is generated on bit patterns of the feature matrix
269 of a particular criminal. The main purpose of the DMO is to obtain a criminal specific crime flow
270 which captures the crime patterns which are frequently followed by a particular criminal. It is
271 named as the dynamic modus operandi as it is subject to change when the new crime flows are
272 added to the feature matrix. Therefore, this addresses the changing nature of the patterns used
273 by the criminals in committing crimes. First, a frequency threshold is generated using
274 characteristic features of the feature matrix at hand which is the matrix of all crimes committed
275 by the same criminal under consideration. The matrix shown in Table 3 is an example to a
276 situation of a feature matrix generated on the previous convictions of a criminal. For simplicity
277 let's consider a feature matrix of 10 columns.

278 **Table 3**. *Feature matrix generated on four previous convictions of a criminal*

| A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 |

279 If we consider A to J of Table 3 as crime flow features of the corresponding MOs, we can
280 understand that in the first MO the criminal has followed a crime flow of A-E-F-G-I.  The same
281 criminal has followed a crime flow of A-D-F-G-I in his second crime. Likewise the other two crime
282 flows are, A-E-F-H-I and A-D-F-G-H-I respectively.

283  The DMO of a particular criminal is generated using the Apriori method [27]. Apriori method is
284  used to find the crime entities with the frequency threshold (frt) which is generated according to
285  Equation 2. A demonstration of the generation of D in Equation 1 on the properties of feature
286  matrix is shown in Table 4.

$$D = \left\{ d \,\middle|\, d = \sum_{i=1}^{n} y_i \right\} \tag{1}$$

$$frt = \left. M_D \middle/ n \right. \tag{2}$$

287  Where,

288  $D$ = vector of distinct column frequencies of the feature matrix.

289  $y_i$ = cells in each column

290  $M_D$ = Median of D,

291  $n = \sum f$ = number of values or total frequencies,

292  c = cumulative frequency of the median class

293  h = class interval size.

294

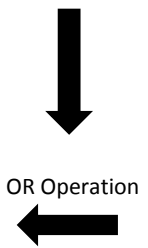
295  **Table 4.** *Column-wise addition of the feature matrix of the suspect under consideration*

| A | B | C | D | E | F | G | H | I | J | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | |
| 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | |
| 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | Summation |
| 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | |
| 4 | 0 | 0 | 2 | 2 | 4 | 3 | 2 | 4 | 0 | |

296
297  The column-wise addition of the matrix shown in Table 4 gives 4, 0, 0, 2, 2, 4, 3, 2, 4 and 0. The
298  distinct numbers are selected from the resulting vector which results in $D = [0, 2, 3, 4]$. The
299  median of D is then divided by the number of instances (rows) in the matrix as the frt, which is
300  2.5/4 = 0.625 for the above case. Therefore, frt will range from 0 to 1. This value provides an
301  insight to a fair threshold value for the Apriori method to generate the dynamic modus operandi
302  with the most frequent elements. frt is used as the frequency threshold in finding the lengthiest
303  MO with a probability of 0.625 because this value suggests that there is a moderate possibility of
304  one feature having 0.625 probability in each of MO.  This results in a dynamic modus operandi
305  (DMO) as shown in Equation 4, because the only transaction of crime attributes which provides
306  a support of 0.625 is $\sigma$ (A,F,G,I) as shown in Equation 3.

$$s = \frac{\sigma(A,F,G,I)}{|T|} = \frac{3}{4} = 0.75 \tag{3}$$

307    $DMO = [1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0]$         (4)

308

## 309   Generating the complete MO profile (CMOP) of the criminals

310    The complete MO profile (CMOP) is obtained by the OR operation between the bits of each
311    column of the feature matrix of the corresponding criminal. CMOP guarantees the provision of a
312    composite crime flow by considering all of the previous crime flow entities of a particular criminal. For
313    example, the complete profile for the feature matrix shown in Table 3 is obtained as shown in
314    Table 5.

315    **Table 5**. *OR operation on the columns to obtain the complete MO profile*

| A | B | C | D | E | F | G | H | I | J | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | |
| 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | |
| 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | |
| 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | OR Operation |
| 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | |

316    Therefore, $CMOP = [1\,0\,0\,1\,1\,1\,1\,1\,1\,0]$. CMOP contains 1s for each place for which a particular
317    crime flow entity has taken place at least once.

## 318   Finding the deviation probability (DP) of CMOP from the crime MO
## 319   under consideration (UMO)

320    First, the deviation of CMOP and UMO is obtained according to Equation 5. As the binary feature
321    vectors are commonly used to represent patterns, many methods have been invented to find
322    their similarity and distance [28]. Euclidean distance, Hamming distance, Manhattan distance,
323    Jaccard, Sorensen and Tanimoto are few of the frequently used measures in that domain [28].
324    This probability value which is named as the deviation probability (DP), is used to obtain a
325    measurement as to what extent of information is available in the UMO, extra to what is already
326    available in the CMOP of a particular criminal. Let's assume that the bit pattern to be compared
327    with the suspect's modus operandi profile under consideration is $UMO = [1\,0\,0\,0\,1\,1\,1\,0\,0\,1]$.
328    Therefore, DP provides the probability of 1s which are available in UMO but not in CMOP.

329

330    The deviation probability, DP can be given as,

331    $DP = \dfrac{\sum_{i=1}^{n} x_i - y_i}{n}$ ,for $x_i = 1, y_i = 0 \,;\, i = 1,2,...,n$         (5)

332 Where,

333 $x_i = elements\ of\ the\ UMO$

334 $y_i = elements\ of\ the\ CMOP$

335 If we consider the feature matrix on Table 3,

336 $Deviation = \ [1\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ 1] - [1\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 0]$

337 $Deviation = \ [0\ 0\ 0\ -1\ 0\ 0\ 0\ -1\ -11]$ \hfill (6)

338 Define $AD = 1$, where AD is the number of positive 1s.

339 Therefore, $DP = 1/10 = 0.1$

340 As it appears in Expression 6, it produces positive 1s for the places with the features available in
341 UMO but not in CMOP. The higher the DP, higher the amount of extra information available in
342 UMO. Hence, a DP value close to 0 indicates the absence of extra features in UMO.

### 343 Finding the completeness probability (CP) of UMO against DMO

344 For the same feature matrix which was considered in Table 3, the CP is obtained according to
345 Equation 7. Here, the UMO is compared with DMO to obtain a probability to determine what
346 extent of features in CP is available in UMO. Therefore, it is derived by the percentage of
347 attributes which are present in both UMO and DMO.

348 Let $DMO = \{x_i\}_{i=1}^{n}$ and $UMO = \{y_j\}_{j=1}^{n}$ be two binary sequences.

349 Define, $z_k = \begin{cases} 1; x_i = y_j \\ 0; otherwise \end{cases}$. Then, $CP = \frac{\Sigma_{k=1}^{n} z_k}{n}$ is the completeness probability. \hfill (7)

350 For example, if we consider $DMO = [1\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 1\ 0]$, then for the
351 $UMO = [1\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ 1]$ a CP of $3/10 = 0.3$ is generated as in the 1st, 6th and 7th positions
352 there are ones in both DMO and UMO. Higher the CP value, the more the UMO is composed of
353 crime flow entities which are available in the DMO. Therefore, a CP value close to 1 indicates
354 that the completeness of UMO compared to DMO is 100%.

### 355 Building a fuzzy inference system to obtain the final similarity score

356 The vagueness of the two measurements CP and DP generates a difficulty in calculating a
357 similarity score using crisp logic. Therefore, the two parameters CP and DP were adapted into a
358 fuzzy inference system which accepts two inputs and provides a score for the similarity between
359 a suspect and a crime. Figure 4 shows a block diagram of the proposed fuzzy inference system.
360 Mamdani fuzzy inference was used as an attempt to solve a control problem by a set of linguistic
361 rules obtained from experienced human operators [29]. First, the rule base of the fuzzy controller

362  was defined by observing the variations of CP and DP. The membership functions of the inputs
363  and outputs were then adjusted in such a way that, the parameters which seem to be wrong can
364  be fine-tuned, which is a common practice in defining fuzzy inference systems [30]. Literature
365  shows many methods used in fine tuning the fuzzy parameters.  Usage of adaptive networks [31]
366  and Neuro-fuzzy systems [32] in fine tuning the fuzzy parameters have received more attention.
367  The problem at our hand was to generate a fuzzy inference system which generates the highest
368  similarity score when the DP value goes down and CP value goes up. We conducted a manual
369  mapping procedure for the fuzzy membership functions. Therefore, the input and output space
370  of the two inputs CP and DP and the output were partitioned into 3 subsets. Namely, LOW,
371  MODERATE and HIGH. Center of gravity was used as the defuzzification strategy of the fuzzy
372  controller. Mamdani fuzzy inference was especially selected for the similarity score generation
373  procedure, for the highly intuitive knowledge base it offers due to the fact that both antecedents
374  and the consequents of the rules are expressed as linguistic constraints [33]. First, we selected
375  all of these membership functions with 50% overlap. Then the tuning procedure was conducted
376  during which we adjusted either the left and/or right spread and/or overlapping to get the best
377  possible similarity score for the given DP and CP. This procedure was conducted until the FIS
378  generated satisfactory results.

379

380

381

382

383

384

385

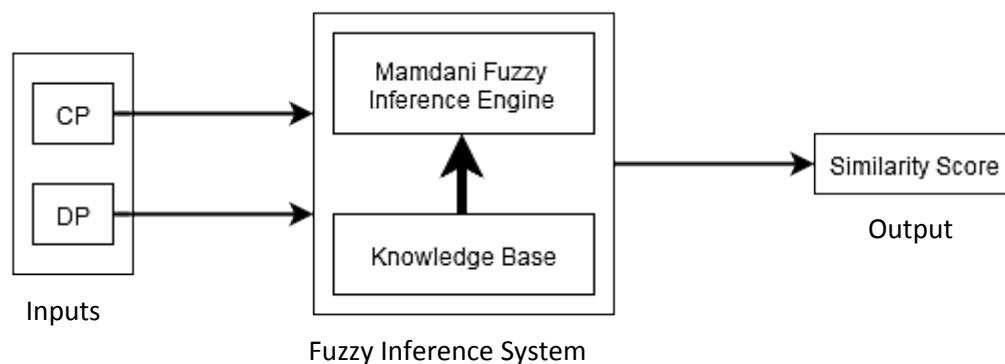



**Figure 4.** *Block diagram of the proposed fuzzy inference system.*

386

387  Figures 5 and 6 show the fuzzy inputs of the Fuzzy Inference System (FIS) which correspond to CP and DP
388  values respectively. Figure 7 depicts the fuzzy output of the FIS. As the Figures 5, 6 and 7  depict, all the
389  different levels of membership functions under each input and the output are selected to be triangular
390  and  trapezoidal  functions  as  triangular  or  trapezoidal  shapes  are  simple  to  implement  and
391  computationally efficient [34].

392

393

394

395

396

397

398

399

400

401

402

403
404

405

406

407

408

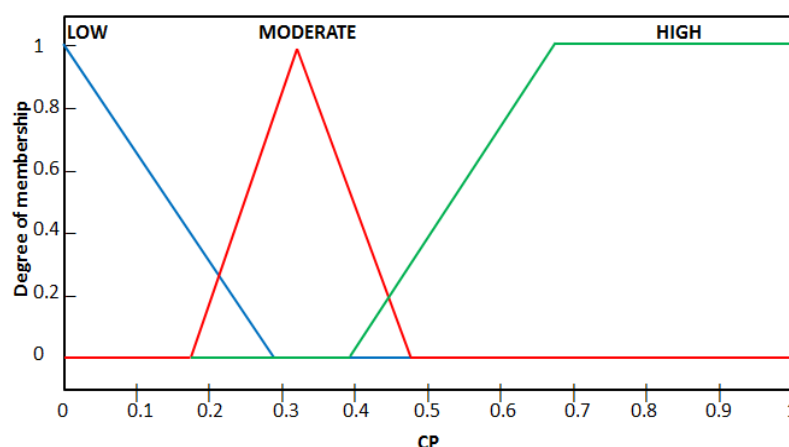
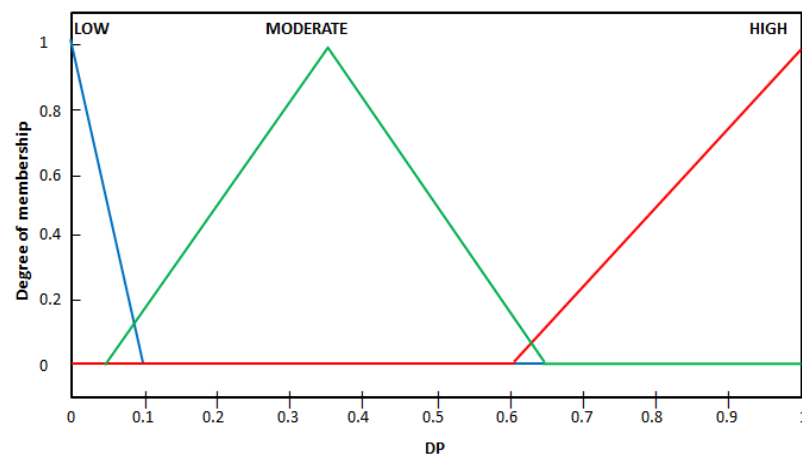



**Figure 6.** *Input fuzzy variable 2: DP*

409   .

410

411

412

413
414

415
416

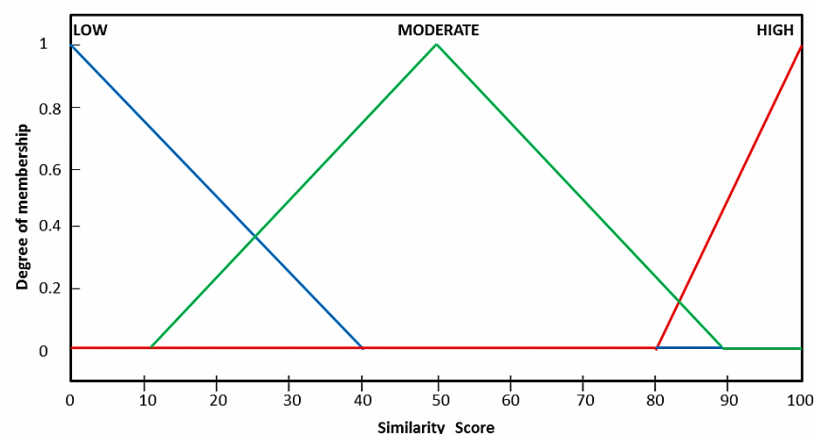



**Figure 7.** *Output fuzzy variable: similarity score*

418  As shown in Figure 7, the universe of discourse of similarity score (fuzzy output) ranges from 0 to
419  100. The defuzzified score which is generated from the FIS is considered as the measurement for
420  how close the modus operandi under consideration is to a particular suspect's profile. A higher
421  score value close to 100 provides a good indication about a high similarity between the modus
422  operandi of the crime and suspect under consideration.

423  The fuzzy rule derivation of the fuzzy controller is heuristic in nature. According to the
424  calculations of the two inputs, higher values of CP, close to 1 and lower values of DP close to 0,
425  positively affect the final similarity score. The rule base of the fuzzy model is generated
426  accordingly. The rule base provides a non-sparse rule composition of 9 combinations as
427  illustrated in Figure 8.

428

429

| DP / CP | LOW | MODERATE | HIGH |
|---|---|---|---|
| LOW | MODERATE | LOW | MODERATE |
| MODERATE | HIGH | MODERATE | LOW |
| HIGH | HIGH | MODERATE | LOW |

430
431
432
433

434        **Figure 8.** Fuzzy rule set of the rule base of the inference system

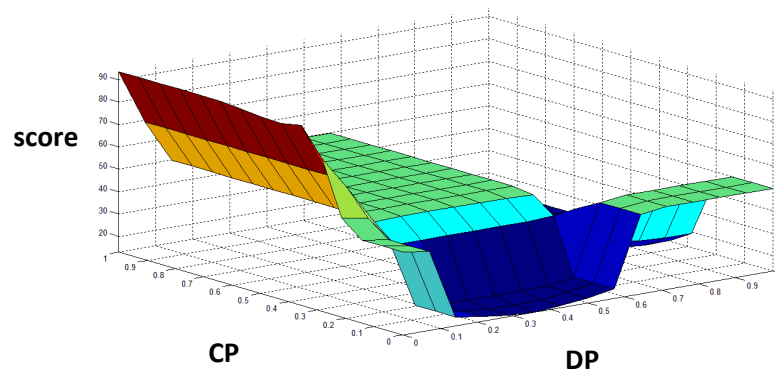
435    The rule surface of the fuzzy controller depicted in Figure 9, shows the variation of the similarity
436    score with the changes of the two inputs CP and DP. According to the figure it's perfectly visible,
437    for higher values of CP (close to 1) and for lower values of DP (close to 0), the fuzzy controller
438    generates higher values for the similarity score which are close to 100.

439

440

441

442

443

       **Figure 9.** Rule surface of the fuzzy controller.

444 ## Classification of the UMO under the class with the highest similarity

445 ## score

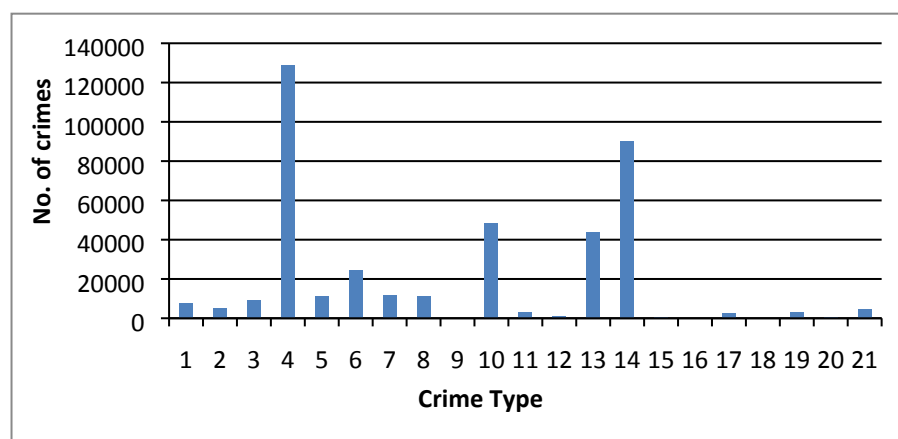
446    When the algorithm is used to find associations between modi operandi of criminals and modi
447    operandi of crimes, the similarity score which is generated from the newly proposed method can
448    be used directly. A similarity score which is close to 100 would suggest that the criminal has a
449    very high tendency to have committed the crime which is under investigation. Therefore, the
450    similarity scores can be used to classify a particular modus operandi to a most probable suspect
451    with the highest similarity score.

452    The proposed method was developed by using MATLAB 7.12.0 (R2011a) [35]. All the necessary
453    implementations were conducted using the MATLAB Script editor [36] apart from the FIS which
454    was implemented using the MATLAB fuzzy toolbox [37]. The nine classification algorithms which

455 were used for the performance comparison were classification algorithms which are already
456 packaged with the WEKA 3.6.12 tool [38].

## Results and Discussion

458 The method was tested with a crime data set obtained from Sri Lanka Police. Figure 10 shows
459 the crime frequencies in Sri Lanka by the crime types from 2005 to 2011. It shows only 21 crime
460 types because the 5 new crime types were introduced in 2015. 4th column denoting House
461 Breaking and Theft shows the highest number of occurrences. 14: Theft of property, 10: Robbery,
462 13: Cheating/ Misappropriation, 6: Hurt by Knife, 7: Homicide, 8: Rape/ Incest, 5: Grievous Hurt,
463 3: Mischief over Rs. 5000/=, 1: Abduction/Kidnapping comes next. For the validation of the
464 algorithm, 7 crime types out of these 10 types were selected for the testing data set. They are,
465 House Breaking and Theft, Theft of property, Robbery, Homicide, Rape/ Incest, Grievous Hurt,
466 Abduction/Kidnapping. 31 crime flows were selected which are common to the seven selected
467 crime types. The data set is also composed of 8 sub types and 2 special categories. Altogether
468 the data set consisted of 67 instances in which each instance is composed of 48 attribute values.
469 The data set is distributed over 20 classes (criminals).

470

477 **Figure 10**. *Frequency of different crime types from year 2005-2011*

482

483

484

485

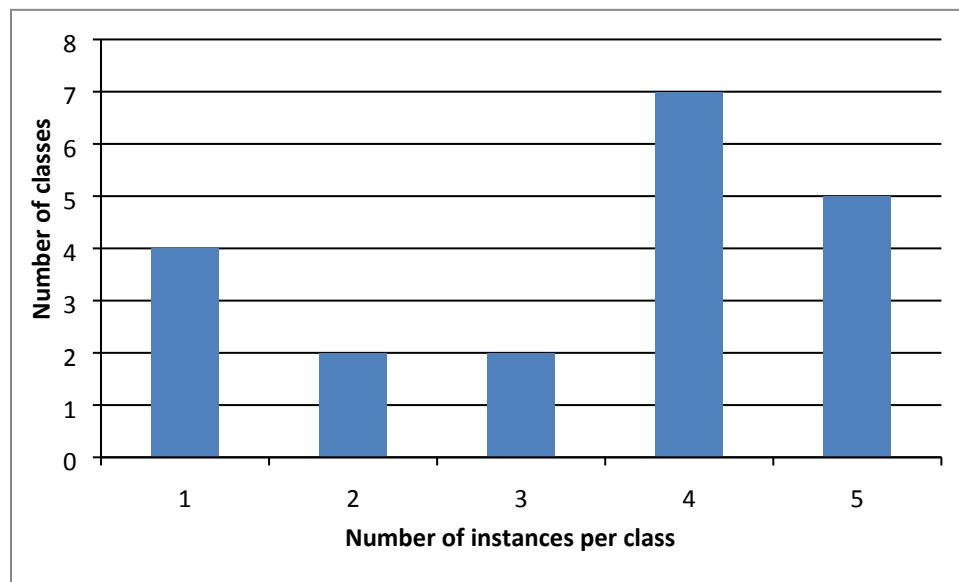
486

487

488

489

490

491

492        **Figure 11.** *Distribution of modus operandi instances over the classes of the dataset*

493    All the tests were performed in a Windows computer with Intel (R) Core (IM) i7-2670QM CPU of

494    2.20 GHz and a RAM of 8GB.  The histogram of the instance distribution over the classes is shown

495    in Figure 11.

496    10 fold cross validation [39] was used on the data set for a fair testing procedure. In 10-fold cross

497    validation, the data set is divided into 10 subsets, and the holdout method is repeated 10 times.

498    Each time, one of the 10 subsets is used as the test set and the other 9 subsets are put together

499    to form a training set. Then the average error across all 10 trials is computed [39].

500

501

502

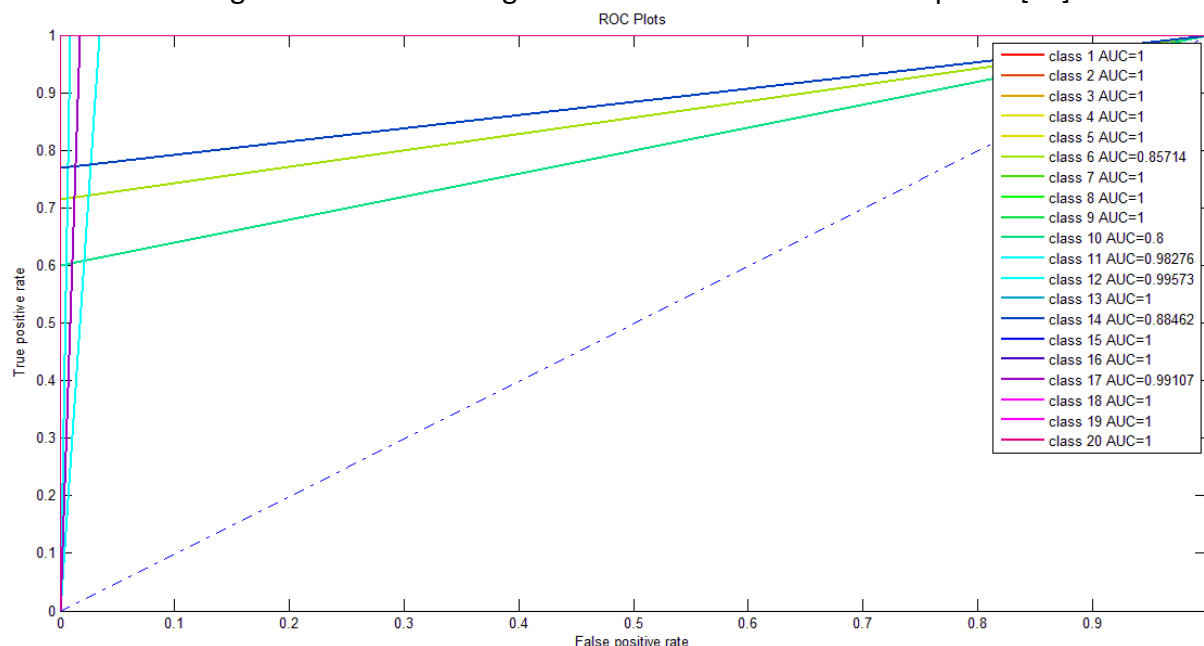
503

504

505

506

507

508

509    **Figure 12**. *ROC curves returned by the newly proposed method on the 20 classes of crime data set*

510    The test results of modus operandi classifications in Area Under Curve (AUC) [40], and time
511    elapsed for the classification are shown in Table 6. A Receiver Operating Characteristic (ROC)
512    curve is a two dimensional graphical illustration of the trade-off between the true positive rate
513    (sensitivity) and false positive rate (1-specificity). Figure 12 depicts the ROC curve plotted on the
514    classification results obtained by the newly proposed method on the crime data set. In the
515    particular instance which is shown in Figure 12, all the ROC curves related to the crime data set
516    are plotted well over the diagonal line and all of them have retuned AUC values which are either
517    equal to 1 or very close to 1, providing a very good classification.

518    To prepare the data set which was used under this research, a crime data set of around 3000
519    instances was analyzed. Due to limitations of the real crime data set, it was quite a complex task
520    to prepare a data set with a collection of sufficient modus operandi where each instance has a
521    considerable flow of crime flows. Therefore, only a sample of 67 instances could be filtered from
522    the population to generate a representative data set and it was verified by a domain expert
523    before being used in the analysis. As the number of instances was around 67, it can be considered
524    as an under-represented data sample. Another reason for the data set to become under-
525    represented was the challenge in finding classes/criminals with more than one crime committed.
526    The actual    crime data set which is used for the testing purposes is imbalanced as it is
527    apparent in Figure 11. This imbalanced nature of the data set may produce biased results. To
528    make the classification process unbiased, we used the concept of oversampling. Oversampling
529    and under-sampling are two concepts which are used in overcoming class imbalance problems in
530    input data sets.  Oversampling and under-sampling are two different categories of resampling
531    approaches, where in oversampling the small classes are incorporated with repeated instances
532    to make them reach a size close to lager classes, whereas in under-sampling, the number of
533    instances is deceased in such a way that the number of instances  reach a size close to the smaller
534    classes [41].

535    Table 6 shows the results returned by the fuzzy based binary feature profiling which was
536    conducted on the actual crime data set. As shown in the table, there is an increase in the accuracy
537    when the input data set undergoes oversampling. Since the maximum number of instances
538    available under one suspect is equal to 5, under-sampling does not provide a good accuracy. The
539    results prove that the new algorithm works well for a balanced data set as the new method
540    showed an increase in performance when the data set is subjected to an oversampling greater
541    than or equal to 5.

542

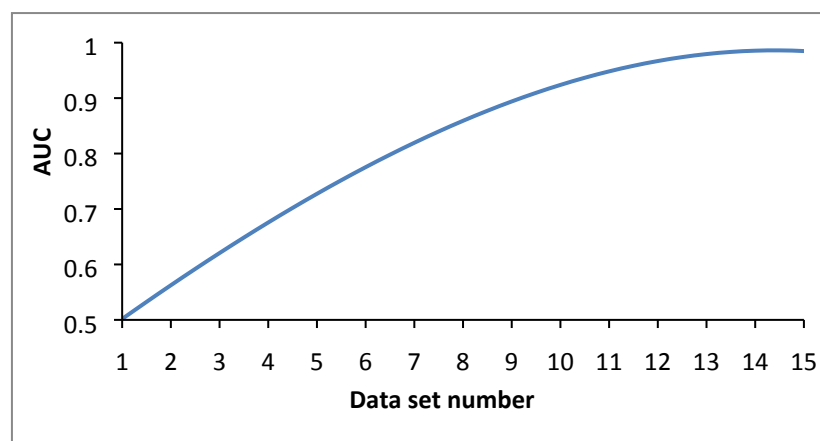
543

544    **Table 6.** *Results returned by the fuzzy based binary feature profiling for the modus operandi analysis on*
545    *actual data*

| Data set (Number) | Oversampling or Under-sampling value | AUC | Average time elapsed |
|---|---|---|---|
| 1 | 2 | 0.5417 | 0.0015 |
| 2 | 3 | 0.5562 | 0.0011 |
| 3 | 4 | 0.5965 | 0.0014 |
| 4 | N/A | 0.6937 | 0.0010 |
| 5 | 5 | 0.6612 | 0.0011 |
| 6 | 6 | 0.7063 | 0.0011 |
| 7 | 10 | 0.8033 | 0.0012 |
| 8 | 20 | 0.9339 | 0.0013 |
| 9 | 30 | 0.9661 | 0.0014 |
| 10 | 40 | 0.9637 | 0.0015 |
| 11 | 50 | 0.9756 | 0.0016 |
| 12 | 60 | 0.9626 | 0.0018 |
| 13 | 70 | 0.9365 | 0.0019 |
| 14 | 80 | 0.9391 | 0.0023 |
| 15 | 90 | 0.9671 | 0.0029 |

546

547    Figure 13 shows the change in AUC with the increase of sampling which starts from under-sampling of 2
548    and goes on to an over sampling of 90. According to the plot it can be observed that the AUC values are
549    increased when the oversampling is increased.

550

557    **Figure 13**. *Change of AUC values with oversampling*

558    The execution time of the algorithm was 0.001s when there is no oversampling or under-
559    sampling. The maximum execution time is 0.0031 when there is an oversampling of 90. According
560    to the plot shown in Figure 14, it is clear that there is an increase of execution time as the

561     oversampling size increases. But, the overall execution time is always remained under 3
562     milliseconds.

563

564

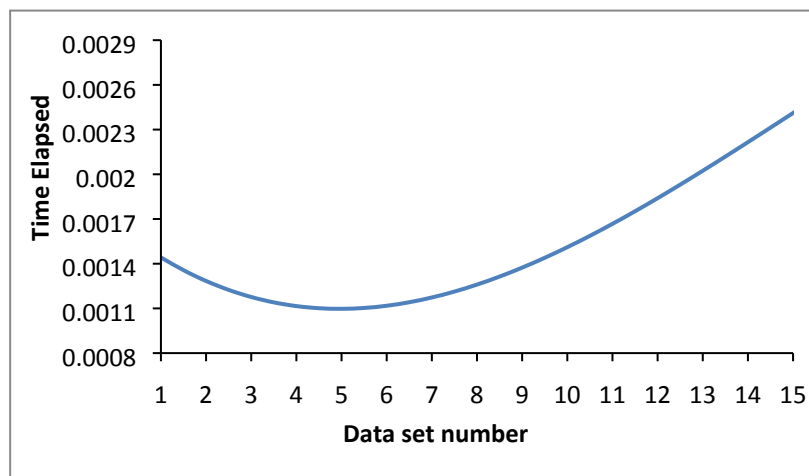
565

566

567

568

569

570

571                 **Figure 14**.  *Change of time elapsed for the 15 data sets.*

572

## Overview of the classification algorithms used for the comparison

574     It is a known fact that there is no single algorithm which can be categorized as the best to solve
575     any problem. Different classification algorithms may perform differently in different situations
576     [42]. Therefore, the newly proposed method was tested against ten other open classification data
577     sets and the performance was evaluated against the results obtained from nine other well-known
578     classification techniques, thereby assessing the quality of the newly proposed method. The nine
579     other classification algorithms include, Logistic Regression, J48 Decision Tree, Radial Basis
580     Function Network (RBFNetwork), Multi-Layer Perceptron (MLP), Naive Bayes Classifier,
581     Sequential Minimal Optimization (SMO) algorithm, KStar instance based classifier, Best-first
582     decision tree (BFTree) classifier, and Logistic Model Tree (LMT) classifier. These classifiers
583     represent four classes of classification algorithms. Namely, function based classifiers, Tree based
584     classifiers, Bayesian classifiers and Lazy classifiers.

585     Logistic Regression learns conditional probability distribution. Relating qualitative variables to
586     other variables through a logistic cumulative distribution functional form is logistic regression
587     [43]. J48 is an open source java implementation of the C4.5 decision tree algorithm [44]. A
588     decision tree consists of internal nodes that specify tests on individual input variables or
589     attributes that split the data into smaller subsets, and a series of leaf nodes assigning a class to
590     each of the observations in the resulting segments. C4.5 algorithm constructs decision trees using
591     the concept of information entropy [45]. Neural networks are flexible in being modeled virtually
592     for any non-linear association between input variables and target variables [46]. Both Radial basis

593 function networks and MLP networks are neural networks [47]. Bayesian classifiers assign the
594 most likely class to a given example described by its feature vector [48]. SMO is an
595 implementation of John Platt's sequential minimal optimization algorithm for training a support
596 vector classifier. It globally replaces all missing values and transforms nominal attributes into
597 binary one. It also normalizes all attributes by default [49] [50]. KStar (K*) is an instance-based
598 classifier which uses an entropy –based distance function [51]. BFTree uses binary split for both
599 nominal and numeric attributes [52]. LMT is a classifier for building 'logistic model trees', which
600 are classification trees with logistic regression functions at the leaves [53], [54].

601 **Table 7**. *Description of the classification data sets for performance comparison*

| Data set | Description | Number of Instances | No of Attributes |
|---|---|---|---|
| Dermatology Data Set [55] | This database has been created on a dermatology test carried out on skin samples which have been taken for the evaluation of 22 histopathological features. The values of the histopathological features have been determined by an analysis of the samples under a microscope. In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values. | 336 | 33 |
| Balance Scale Data Set [56] | This data set has been generated to model psychological experimental results. Each example is classified as having the balance scale tip to the right, tip to the left, or be balanced. The attributes are the left weight, the left distance, the right weight, and the right distance. The correct way to find the class is the greater of (left-distance * left-weight) and (right-distance * right-weight). If they are equal, it is balanced. There are 3 classes (L,B,R), five levels of Left-Weight (1,2,3,4,5), five levels of Left-Distance (1,2,3,4,5), five levels of Right-weight (1,2,3,4,5) and five levels of Right-Distance (1,2,3,4,5). | 625 | 4 |
| Balloons Data Set [57] | This data set has been generated using an experiment of stretching a collection of balloons carried out on a group of adults and children [58]. In the data set, Inflated is true if (color=yellow and size = small) or (age=adult and act=stretch). In the data set there are two main output classes, namely T if inflated and F if not inflated, two colors yellow and purple, two sizes, large and small, two act types, stretch and dip, and two age groups, adult and | 20 | 4 |

| | | | |
|---|---|---|---|
| | child. | | |
| Car Evaluation Data Set [59] | Car Evaluation Database has been derived from a simple hierarchical decision model originally developed for the demonstration of DEX by M. Bohanec and V. Rajkovic [60]. The Car Evaluation Database contains examples with information that is directly related to CAR. They are buying, maint, doors, persons, lug_boot and safety. The attribute buying is the buying price which is considered to have four levels v-high, high, med, low. Maint is the price of the maintenance which contains the four levels, v-high, high, med, low. Doors have the four levels 2, 3, 4, 5-more. Person (capacity in terms of persons to carry), lug_boot (the size of luggage boot) and safety (estimated safety of the car) have 3 levels each. | 1728 | 6 |
| Soybean Data set [61] [62] | This is a small subset of the original soybean database. The data set is distributed over four classes, D1, D2, D3 and D4. The 35 categorical variables represent different levels of qualities of the soybean vegetable. These categorical variables include, plant-stand, precip, temp, hail, crop-hist, area-damaged, severity, seed-tmt, germination, lant-growth, leaves, leafspots-halo, leafspots-marg, leafspot-size,leaf-shread, leaf-malf, leaf-mild, stem, lodging, stem-cankers, canker-lesion, fruiting-bodies, external, mycelium, int-discolor, sclerotia, fruit-pods, fruit,seed, mold-growth, seed-discolor, seed-size, shriveling and roots. The number of levels represented by each variable varies from 2 to 3. | 47 | 35 |
| Lenses Data set [63] | Lenses data set is a small database about fitting contact lenses. The data set is composed of five attributes including the class variable. The data set has three classes. Age of the patient, spectacle prescription, astigmatic, tear production rate are the attributes of the data set. The attributes contain at least of two categories and at most of three categories. | 24 | 4 |
| Nursery Data set [64] | Nursery Database has been derived from a hierarchical decision model originally developed to rank applications for nursery schools. It has been used during several years in 1980's when there has been excessive enrollment to these schools in Ljubljana, Slovenia, and the rejected applications frequently needed an objective explanation. The final decision depended on three sub problems: occupation of parents and child's nursery, family structure and financial standing, and social and health picture of the family. The model has been developed within expert system shell for decision making [65]. | 12960 | 8 |

| Tic-tac-toe Data set [66] | This database encodes the complete set of possible board configurations at the end of tic-tac-toe games, where "x" is assumed to have played first. The target concept is "win for x" (i.e., true when "x" has one of 8 possible ways to create a "three-in-a-row"). | 958 | 9 |
|---|---|---|---|
| SPECT Heart Data set [67] | The data set describes diagnosis of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) were processed to extract features that summarize the original SPECT images. The instances are described by 23 binary attributes including the class variable. | 267 | 22 |
| MONK's Problems Data set [68] | The MONK's problems have been the basis of a first international comparison of learning algorithms. The result of this comparison is summarized in "The MONK's Problems. There are three MONK's problems. The domains for all MONK's problems are the same. The data set is composed of 7 attributes and a binary class variable. | 432 | 7 |

602  As the newly proposed method accepts only binary input variables, the data sets which are used
603  for the analysis must be preprocessed into the acceptable format. For example, the "balance
604  scale" data set is composed of 4 attributes. Table 8 shows the attributes and their information of
605  the balance scale data set.

606  **Table 8**. *Attribute information of the balance data set*

| Attribute | Number of Categories | Categories |
|---|---|---|
| Class Name | 3 | L, B, R |
| Left-Weight | 5 | 1,2,3,4,5 |
| Left-Distance | 5 | 1,2,3,4,5 |
| Right-Weight | 5 | 1,2,3,4,5 |
| Right-Distance | 5 | 1,2,3,4,5 |

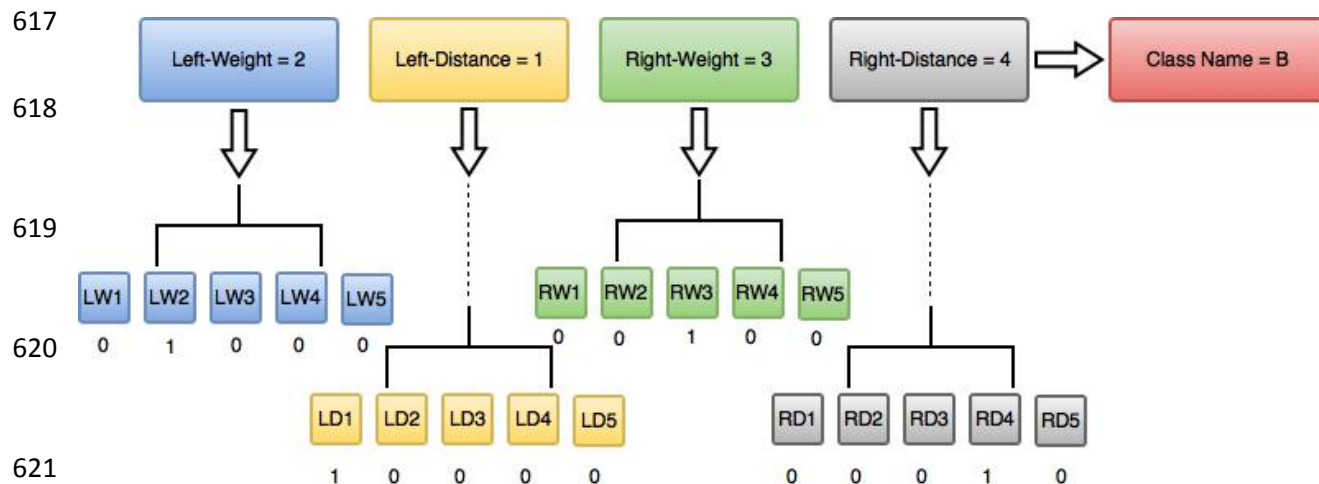
607  Therefore, the data set was adjusted as shown in figure 15, prior to using it with the proposed
608  method. Each category of a particular attribute is represented by a dummy variable. For example,
609  Left-Weight attribute results in 5 attributes in the preprocessed data set and each attribute is
610  represented using 5 binary variables as LW1, LW2, LW3, LW4 and LW5 where the presence of the
611  attribute denotes 1 and 0 otherwise. As depicted in Figure 15, if Left-Weight has a value of 2 in
612  an instance it results in 1 for the corresponding derived attribute that is LW2. Therefore, if there
613  is an instance where Left-Weight=2, Left-Distance=1, Right-Weight=3 and Right-Distance=4, Class
614  Name=B, it is represented as LW1=0,LW2=1, LW3=0, LW4=0, LW5=0, LD1=1, LD2=0, LD3=0,
615  LD4=0, LD5=0, RW1=0, RW2=0, RW3=1, RW4=0, RW5=0, RD1=0, RD2=0, RD3=0, RD4=1, RD5=0,
616  Class Name=B.

617

618

619

620

621

622    **Figure 15**. *Schematic diagram used for pre-processing of the balance dataset in such a way that it matches*
623    *the format of inputs of the newly proposed method*

624

625    The pre-processed data is then fed to the newly proposed algorithm and the nine other
626    algorithms. Performances were compared based on AUC analysis of the ROC curves, and the
627    processing time for model generation. 10 fold cross validation was used under each test for fair
628    testing procedure. For simplicity the newly proposed modes operandi analysis algorithm was
629    acronymed as BFPM (Binary feature profiling methodology).

630    As all the data sets which were used for the tests are composed of multi classes,  weighted
631    average AUC was used, where each target class is weighted according to its prevalence as given
632    in Equation 8. Weighted average was used in order to prevent target classes with smaller instance
633    counts from adversely affecting the results [69].

634    $AUC_{weighted} = \sum_{\forall c_i \in C} AUC\,(c_i) \times p(c_i)$                                    (8)

635    Table 9 shows the weighted average AUC values obtained for each data set under each
636    classification algorithm.

637    **Table 9.** *Weighted average AUC values obtained by the algorithms on classifying the data sets*

|  | BFPM | Logistic Regression | J48 | Radial Basis Function Network | Multi-Layer Perceptron | Naive Bayes Classifier | SMO | KStar | BFTree | LMT |
|---|---|---|---|---|---|---|---|---|---|---|
| **Dermatology Data set** | 1 | 0.9990 | 0.9750 | 0.9860 | 0.9980 | 0.9980 | 0.9930 | 0.9970 | 0.9690 | 0.9960 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Balance scale Data set** | 0.7945 | 0.9760 | 0.8110 | 0.9680 | 0.9770 | 0.9710 | 0.8830 | 0.9510 | 0.8130 | 0.9810 |
| **Balloons Data Set** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **Car evaluation Data set** | 0.8087 | 0.9900 | 0.9760 | 0.9740 | 1 | 0.9760 | 0.9550 | 0.9970 | 0.9940 | 0.9990 |
| **Soybean Data set** | 1 | 1 | 0.9860 | 1 | 1 | 0.9760 | 1 | 1 | 0.9740 | 1 |
| **Lenses Data set** | 0.9537 | 0.7470 | 0.8400 | 0.9170 | 0.8390 | 0.8700 | 0.7250 | 0.8870 | 0.8670 | 0.7980 |
| **Nursery Data set** | 0.9100 | 0.9880 | 0.9950 | 0.9870 | 1 | 0.9820 | 0.9640 | 0.9980 | 0.9990 | 0.9990 |
| **Tic-tac-toe Data set** | 0.9167 | 0.9960 | 0.8970 | 0.7340 | 0.9940 | 0.7440 | 0.9760 | 0.9990 | 0.9450 | 0.9920 |
| **SPECT Heart Data set** | 0.7857 | 0.8310 | 0.7560 | 0.8400 | 0.7860 | 0.8490 | 0.7070 | 0.7850 | 0.7230 | 0.8410 |
| **MONK's Problems Data set** | 0.8333 | 0.7050 | 0.9940 | 0.8130 | 0.9980 | 0.7120 | 0.7460 | 0.9970 | 0.9550 | 0.9880 |

638

639    Friedman's rank test is a nonparametric test analogous to a standard one-way repeated-
640    measures analysis of variance [70]. The Friedman's rank test results returned on the AUC test
641    data are shown in Table 10. This test returns a test statistic ($\chi^2$) value ("Chi-square") of 21.339,
642    degree of freedom of 9 and a p-value of 0.011, proving that there is an overall statistically
643    significant difference between the mean ranks of the classification algorithms. According to the
644    table, the highest mean rank is returned for MLP while the lowest mean rank is returned for SMO,
645    proving that MLP provides the best performance while SMO provides the least performance for

646 the 10 data sets tested. Therefore, it indicates that the new model provides a better performance
647 than BFTree, J48 and SMO algorithms for the 10 data sets tested.

648 **Table 10**. *Friedman's mean rank values returned on the data available in Table 9*

| Method | Mean Rank |
|---|---|
| MLP | 7.70 |
| LMT | 7.10 |
| KStar | 6.95 |
| LogisticRegression | 5.95 |
| RBFNetworks | 5.05 |
| NaiveBayesClassifier | 5.05 |
| **BFPM** | **4.95** |
| BFTree | 4.50 |
| J48 | 4.20 |
| SMO | 3.55 |

649
650 The average processing times elapsed for each algorithm to classify the data sets are given in
651 Table 11. Friedman's rank test on the data of Table 11 returned the results shown in Table 12 in
652 which the mean rank values prove better efficiency for the new method than J48,
653 LogisticRegression, SMO, RBFNetworks, BFTree, MLP and LMT. The test statistic ($\chi^2$) value ("Chi-
654 square") of 73.058, degree of freedom of 9 and a p-value of 0.000, proves that there is an overall
655 statistically significant difference between the mean ranks of the classification algorithms.

656 **Table 11**. *Average processing time for each algorithm on the classification of the ten data sets*

| | BFPM | Logistic Regression | J48 | Radial Basis Function Network | Multi-Layer Perceptron | Naive Bayes Classifier | SMO | KStar | BFTree | LMT |
|---|---|---|---|---|---|---|---|---|---|---|
| **Dermatology Data set** | 0.0027 | 0.3900 | 0.0800 | 0.3800 | 2.7300 | 0.0500 | 0.1400 | 0.2800 | 0.1100 | 2.9000 |
| **Balance scale Data set** | 0.0030 | 0.0300 | 0.0800 | 0.2200 | 0.4800 | 0 | 0.0500 | 0 | 0.0600 | 0.8400 |
| **Balloons Data Set** | 0 | 0 | 0 | 0 | 0.0200 | 0 | 0.0200 | 0 | 0.0200 | 0.0500 |
| **Car evaluation Data set** | 0.0048 | 0.4200 | 0.0300 | 0.2700 | 13.4000 | 0.0200 | 0.1900 | 0 | 0.4800 | 13.3400 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Soybean Data set** | 0.0042 | 0.0500 | 0.0700 | 0.1800 | 0.4300 | 0 | 0.0500 | 0 | 0.2300 | 0.9200 |
| **Lenses Data set** | 0.0009 | 0 | 0 | 0.0300 | 0.0800 | 0 | 0.0300 | 0 | 0.0100 | 0.0300 |
| **Nursery Data set** | 0.0013 | 8.8600 | 0.2500 | 16.4300 | 127.8600 | 0.0300 | 23.0300 | 0 | 7.9900 | 240.4600 |
| **Tic-tac-toe Data set** | 0.0091 | 0.1900 | 0.0100 | 0.1300 | 18.1800 | 0.0100 | 0.6000 | 0 | 0.6100 | 54.6700 |
| **SPECT Heart Data set** | 0.0073 | 0.0600 | 0.0100 | 0.0700 | 3.8800 | 0 | 0.0400 | 0 | 0.2300 | 2.1500 |
| **MONK's Problem Data set** | 0.0058 | 0.0800 | 0.0100 | 0.0600 | 5.5100 | 0 | 0.1300 | 0 | 0.2100 | 1.7200 |

657

658    **Table 12**. *Mean rank values returned by the Friedman's rank test on the time values available in Table 11*

| Method | Mean Rank |
|---|---|
| KStar | 2.10 |
| NaiveBayesClassifier | 2.35 |
| **BFPM** | 2.75 |
| J48 | 4.15 |
| LogisticRegression | 5.35 |
| SMO | 6.25 |
| RBFNetworks | 6.35 |
| BFTree | 6.90 |
| MLP | 9.30 |

| LMT | 9.50 |
|---|---|

659

660 Friedman's rank test results for the two measurements, AUC and time elapsed conclude that the

661 newly proposed method provides acceptable results against the nine other well established

662 classification algorithms.

## 663 **Conclusion**

664 The studies of modus operandi help crime investigation by letting the police officers to solve

665 crimes by linking suspects to crimes. Though there are many descriptive studies available under

666 modus operandi analysis, a very little amount of work is available under computer science. Many

667 of these methods have been derived using the methods based on link analysis. But, the accuracy

668 of these methods is always compromised due to the cognitive biases of the criminals.

669 A novel Fuzzy based Binary Feature Profiling method (BFPM) to find associations between crimes

670 and criminals, using modus operandi is introduced. The newly proposed method subjects not

671 only the properties of the present, but also the properties of his/her previous convictions. The

672 concept of dynamic modus operandi which is available in the proposed method considers the

673 modi operandi of all of his/her previous convictions to provide a fair rectification to the errors

674 which result due to the human cognition. Dynamic MO uses frequent item set mining to result in

675 a generalized binary feature vector. Complete MO profile also encapsulates past modi operandi

676 of a particular criminal by aggregating the modi operandi of all of his/her previous convictions to

677 one binary feature vector. This feature also guarantees a usage of criminal's past crime record

678 with more generalizability. Completeness probability measures how much information is

679 available in the new crime which is not available in the complete MO profile. Therefore, this

680 measurement provides the capability of measuring how much extra amount of information is

681 carried by the MO of the new crime. The deviation probability provides a notion about how much

682 the new MO deviates from the most frequent crime flows which are available in the dynamic MO

683 of a particular criminal. The vagueness and the impreciseness prompted the fact that it is not

684 possible to use crisp logic to generate the similarity score. Therefore, a fuzzy inference system

685 was modeled to generate the similarity score.

686 Due to the under-represented and imbalanced properties of the actual data set, the new method

687 has returned a lower performance when it was proposed to the data set without any rectification

688 on the data set. However, with the introduction of over sampling, the method returned a very

689 good performance, allowing one to arrive at the conclusion that the method could provide

690 acceptable results for a balanced data set. The method generated favorable results in providing

691 a good similarity measurement to suggest the connections between crimes and criminals. Fuzzy

692 controller of the new approach guarantees to resemble the human reasoning process by

693 confirming the usage of human operator knowledge to deal with nonlinearity of the actual

694 situation. The newly proposed method was then adapted into a classification algorithm for the

695 validation and comparison with other classification algorithms. The comparison of the new

696   method with the well-established classification algorithms confirmed the generalizability of the
697   new method.

698   The method only provides the capability to process the categorical data sets. If there are any
699   continuous variables in the data set, the values must be introduced with categories before
700   further processing. The method can be further extended to directly accept the continuous
701   attributes. As the center of gravity method is used for the defuzzification process, further
702   optimizations can be done by simplifying the defuzzification procedure. Adapting the fuzzy
703   inference engine to a Sugeno [71] type and converting the defuzzification method to a more
704   computationally efficient method such as the weighted average [72] method would provide a
705   less complex computation. This would result in even less processing time when the sophistication
706   of the data set rises.

## 707   References

708

[1] Holdaway, S., *Issues in Sociology: Crime and Deviance*.: Nelson Thornes Ltd, 1993.

[2] Chisum, W.J. , Turvey, B., "Evidence dynamics: Locard's exchange principle & crime reconstruction.," *Journal of Behavioral Profiling*, vol. 1, no. 1, pp. 1-15, 2000.

[3] Paternoster, R., Bachman, R., *Explaining Criminals and Crime: Essays in Contemporary Criminological Theory*, Ronet Bachman Raymond Paternoster, Ed.: Roxbury Publishing Company, 2001.

[4] Palmiotto, M.J., "Crime Pattern Analysis: An Investigative Tool," *Critical Issues in Criminal Investigation*, vol. 2, pp. 59-69, 1988.

[5] Douglas, J. E., Munn, C., "Modus operandi and the signature aspects of violent crime," in *Crime Classificatio Manual*, 2nd ed.: John Wiley & Sons, 2006, pp. 19-30.

[6] Chamikara, M.A.P., Galappaththi, A., Yapa, Y.P.R.D., Nawarathna, R.D., Kodituwakku, S.R., Gunathilake, J., Liyanage, L.H., "A Crime Data Analysis Framework with Geographical Information Support for Intelligence Led Policing," *Manuscript submitted to PeerJ, PeerJ PrePrints 3:e1909 https://doi.org/10.7287/peerj.preprints.1529v1*, July 2015.

[7] The 'Lectric Law Library. The 'Lectric Law Library. [Online]. http://www.lectlaw.com/files/int20.htm

[8] Bennell, C., Canter, D. V., "Linking commercial burglaries by modus operandi: Tests using regression and ROC analysis," *Science & Justice*, vol. 42, no. 3, pp. 153-164, 2002.

[9] Bennell, C., Jones, N. J., "Between a ROC and a hard place: A method for linking serial burglaries by modus operandi," *Journal of Investigative Psychology and Offender Profiling*, vol. 2, no. 1, pp. 23-41, 2005.

[10] Leclerc, B., Proulx, J., Beauregard, E., "Examining the modus operandi of sexual offenders against children and its practical implications," *Aggression and violent behavior*, vol. 14, no. 1, pp. 5-12, 2009.

[11] Oatley,G., Ewwart, B., "Data mining and crime analysis," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 2, pp. 147-153, 2011.

[12] King, R.D., Sutton, G.M., "High times for hate crimes: Explaining the temporal clustering of hate-motivated offending," *Criminology*, vol. 51, no. 4, pp. 871-894, 2013.

[13] Borg, A., Boldt, M., Lavesson, N., Melander, U., Boeva, V., "Detecting serial residential burglaries using clustering," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5252-5266, 2014.

[14] (2011) Offender Profiling. [Online]. http://www.liv.ac.uk/psychology/ccir/op.html

[15] Canter, D., Hammond, L., Youngs, D., Juszczak, P., "The efficacy of ideographic models for geograhical offender profiling," *Journal of Quantitative Criminology*, vol. 29, no. 3, pp. 423-446, 2013.

[16] Agrawal, R., Imielinkski, J., Swami, A., "Mining Association rule between sets of items in large databases," in *Proceedings of the ACM SIGMOD International Conference of Management of Data*, New York, 1993, pp. 207-216.

[17] Yi,X., Rao, F.Y., Bertino, E.,Bouguettaya, A., "Privacy preserving association rule mining in cloud computing," in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, 2015, pp. 439-450.

[18] Koperski, K., Han, J., "Discovery of spatial association rules in geographic information databases," in *Proceeding of the 4th International Symposium on Spatial Databases*, 1995, pp. 47-67.

[19] Chen, H., *Intelligence and security informatics for international security*, 1st ed.: Springer US, 2006, vol. 10.

[20] Lin, S., & Brown, D. E., "An outlier-based data association method for linking criminal incidents," *Decision Support Systems*, vol. 41, no. 3, pp. 604-615, 2006.

[21] Berry, M. J., Linoff, G., *Data Mining Techniques: For Marketing, Sales, and Customer Support*, 3rd ed.: Wiley, 2011.

[22] Chen, H., Zeng, D., Atabakhsh, H., Wyzga, W., & Schroeder, J., "COPLINK:Managing Law Enforcement Data and Knowledge," *Communications of the ACM*, vol. 46, no. 1, 2003.

[23] Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., Chau, M., "Crime Data Mining: A general framework and some examples.," *Computer*, vol. 37, no. 0018-9162, pp. 50-56, April 2014.

[24] Capozzoli, A., Lauro, F., Khan, I., "Fault detection analysis using data mining techniques for a cluster of smart office buildings," *Expert Systems with Applications*, vol. 42, no. 9, pp. 4324-4338, 2015.

[25] Chikersal, P., Poria, S. and Cambria, E., "SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning," in *Proceedings of the International Workshop on Semanti Evaluation, SemEval*, 2015, pp. 647-651.

[26] Chen, H., "Machine learning for information retrieval: neural," *Journal of the American Society for Information Science*, vol. 46, no. 3, pp. 194-216, 1995.

[27] Adamo, J. M., *Data mining for association rules and sequential patterns, Sequential and Parallel Algorithms*, 1st ed. New York: Springer Science & Business Media, 2001.

[28] S.S., Cha, S.H., Tappert, C.C., Choi, "A survey of binary similarity and distance measures," *Journal of Systemics, Cybernetics and Informatics*, vol. 8, no. 1, pp. 43-48, 2010.

[29] Mamdani, E.H., Assilina, S., "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of Man-Machine Studies*, vol. 7, no. 1, pp. 1-13, 1975.

[30] Godjevac, J., *Neuro-fuzzy Controllers: Design and Application*.: PPUR presses polytechniques, 1997.

[31] Sun, C.T., "Rule-base structure identification in an adaptive-network-based fuzzy inference system," in *Fuzzy Systems, IEEE Transactions on*, vol. 2, 1994, pp. 64-73.

[32] Abraham, A., Nath, B. and Mahanti, P.K., "Hybrid intelligent systems for stock market analysis," *Computational science-ICCS*, pp. 337-345, 2001.

[33] Zadeh, L.A., "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets and Systems*, vol. 90, pp. 117-117, 1997.

[34] MathWorks, Inc. (1994-2015) MathWorks. [Online]. http://in.mathworks.com/help/fuzzy/foundations-of-fuzzy-logic.html

[35] MathWorks. (1994-2015) MathWorks. [Online]. https://in.mathworks.com/

[36] MathWorks. (1994-2015) MathWorks Documentation. [Online]. http://in.mathworks.com/help/matlab/ref/edit.html

[37] MathWorks. (1994-2015) MathWorks Fuzzy Logic Toolbox. [Online]. http://in.mathworks.com/help/matlab/ref/edit.html

[38] Machine Learning Group at the University of Waikato. WEKA. [Online]. http://www.cs.waikato.ac.nz/ml/weka/

[39] Refaeilzadeh, P., Tang, L., Liu, H., "Cross-validation," in *Encyclopedia of database systems*, M.

Tamer Özsu Ling Liu, Ed.: Springer US, 2009, pp. 532-538.

[40] Hanley, J. A., McNeil, B. J., "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29-36, 1982.

[41] Estabrooks, A., Jo, T., Japkowicz, N., "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18-36, 2004.

[42] Wolpert, D.H., "The Lack of a Priori Distinctions Between Learning Algorithms," *Neural Computation*, vol. 8, pp. 1341-1390, 1996.

[43] Chang, Y. C. I., Lin, S. C., "Synergy of Logistic Regression and Support Vector Machine in Multiple-class Classification," in *Intelligent Data Engineering and Automated Learning - IDEAL 2004*, vol. 5, Exeter, UK, 2004, pp. 132-141.

[44] Machine Learning Group at the University of Waikato. Class J48. [Online]. http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html

[45] J.R. Quinlan, "C4.5 programs for machine learning," *Machine Learning*, vol. 16, no. 3, pp. 235-240, 1993.

[46] Bishop, C.M., *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press, 1995.

[47] Jayawardena, A. W., Fernando, D. A. K., Zhou, M.C., "Comparison of Multilayer Peceptron and Radial Basis Function Networks as Tools for Flood Forecasting," in *Destructive Water: Water-Caused Natural Disasters, their Abatement and Control*, California, 1996, pp. 173-181.

[48] Rish, I., "An empirical study of the naive Bayes classifier," , vol. 3, New York, 2001, pp. 41-46.

[49] Platt, J., "Fast Training of Support Vector Machines using Sequential Minimal Optimization," in *Advances in Kernel Methods - Support Vector Learning* , B., Burges, C., Smola, A., Schoelkopf, Ed., 1998.

[50] Keerthi, S.S., Shevade, S.K.,Bhattacharyya, C., Murthy, K.R.K., "Improvements to Platt's SMO Algorithm for SVM Classifier Design," *Neural Computation*, vol. 13, no. 3, pp. 637-649, 2001.

[51] Cleary, J.G.,Leonard, E.T., "K*: An Instance-based Learner Using an Entropic Distance Measure," in *12th International Conference on Machine Learning*, 1995, pp. 108-114.

[52] Friedman, J., Hastie, T., Tibshirani, R., "Additive logistic regression: A statistical view of boosting," *Annals of statistics*, vol. 28, no. 2, pp. 337-407, 2000.

[53] Landwehr, N., Hall, M., Frank, E., "Logistic Model Trees," *Machine Learning*, vol. 95, no. 1-2, pp. 161-205, 2005.

[54] Sumner, M., Frank, E., Hall, M., "Speeding up Logisti Model Tree Induction," in *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2005, pp. 675-683.

[55] Ilter,N., Guvenir, H.A. (1998) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Dermatology

[56] Hume, T. (1994) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Balance+Scale

[57] Pazzani, M. (1991) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Balloons

[58] M. Pazzani, "The influence of prior knowledge on concept acquisition: Experimental and computational results," *Journal of Experimental Psychology: Learning, Memory & Cognition*, vol. 17, no. 3, pp. 416-432, 1991.

[59] Bohanec, M., Zupan, B. (1997) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

[60] Bohanec,M., Rajkovic, V., "Expert system for decision making," *Sistemica*, vol. 1, no. 1, pp. 145-157, 1990.

[61] Fisher, D. (1987, Jan.) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Soybean+%28Small%29

[62] Michalski, R.S., "Learning by being told and learning from examples: an experimental comparison of the two methodes of knowledge acquisition in the context of developing an expert system for soybean desease diagnoiss," *International Journal of Policy Analysis and Information Systems*, vol. 4, no. 2, pp. 125-161, 1980.

[63] Julien, B. (1990, Aug.) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Lenses

[64] Bohanec, M., Zupan, B. (1997, June) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Nursery

[65] Bohanec,M., Rajkovic, V., "Expert system for decision making," *Sistemica*, vol. 1, no. 1, pp. 145-157, 1990.

[66] Aha, D. W. (1991, Aug.) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame

[67] Kurgan,L.A., Cios, K.J. (2001, Oct.) UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/SPECT+Heart

[68] Thrun, S. (1992, Oct.) UCI Machine Learning Repository. [Online].
https://archive.ics.uci.edu/ml/datasets/MONK's+Problems

[69] Hempstalk, K., Frank, E., "Discriminating Against New Classes: One-class versus Multi-class Classification," in *AI 2008: Advances in Artificial Intelligence: 21st Australasian Joint Conference on Artificial Intelligence*, Auckland, 2008, pp. 325-336.

[70] Howell, D.C., *Fundamental Statistics For The Behavioral Sciences Focuses*, 8th ed. Belmont: Wadsworth, Cengage Learning, 2013.

[71] Takagi, T., Sugeno, M., "Fuzzy identification of systems and its applications to modeling and control," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 1, pp. 116-132, 1985.

[72] Wu, D., Mendel, J. M., "Aggregation using the linguistic weighted average and interval type-2 fuzzy sets," *Fuzzy Systems, IEEE Transactions on*, vol. 15, no. 6, pp. 1145-1161, 2007.

709