FEATURE RANKING TECHNIQUES FOR 3D ATS DRUG MOLECULAR STRUCTURE IDENTIFICATION

SAW YEE CHING

MASTER OF SCIENCE IN INFORMATION AND COMMUNICATION TECHNOLOGY

2018

Faculty of Information and Communication Technology

FEATURE RANKING TECHNIQUES FOR 3D ATS DRUG MOLECULAR STRUCTURE IDENTIFICATION

Saw Yee Ching

Master of Science in Information and Communication Technology

2018

FEATURE RANKING TECHNIQUES FOR 3D ATS DRUG MOLECULAR STRUCTURE IDENTIFICATION

SAW YEE CHING

A thesis submitted
in fulfillment of the requirements for the degree of Master of Science in Information and Communication Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2018

DECLARATION

I declare that this thesis entitled "Feature Ranking Techniques for 3D ATS Drug Molecular Structure Identification" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : ……..

Name : SAW YEE CHING

Date : ……..……

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in term of scope and quality for the award of Master of Science in Information and Communication Technology.

Signature	:	……..
Supervisor Name	:	ASSOC. PROF. DR. AZAH KAMILAH MUDA @ DRAMAN
Date	:	……..……

DEDICATION

I would like to dedicate my work to my beloved family and friends, especially to my loving parents who has been a source of inspiration and encouragement throughout my life.

To my dearest supervisors, Associate Professor Dr. Azah Kamilah Muda @ Draman and Dr. Zeratul Izzah Binti Mohd Yusoh for being responsible, receptive and always by my side to encourage and motivate me.

ABSTRACT

Existing laboratory analysis techniques of ATS drug identification have their challenges which include the cost of training expert operators, the cost of acquired materials, and the dangers involved in operating the experiments. Furthermore, with the constantly emerging of the new ATS drugs design into the illicit market, it serves as a challenge to the comprehensive analytical method to detect and validate these compounds. This research is aimed to propose a computational intelligence approach in assisting the analysis phase of ATS drug identification process. The dataset namely ATS drugs 3D molecular structure representation dataset was analyzed. It consists of 7212 sample records associated with 1185 features. This research has investigated numerous complexities and uncertainties that have embedded in the dataset in the form of high dimensionality and existence of irrelevant and noisy features. These challenges motivated this research to tackle these problems by reduce the dimensionality of the dataset and selecting the significant subset of features from the dataset. Hence, this led to the proposal of a feature selection approach for removing the irrelevant and noisy data and selecting a feature subset which best represent the ATS drug and produce a better identification performance. The proposed feature selection approach has a simple algorithmic framework and makes use of the existing feature selection techniques to cater different variety of data issues, namely Ensemble Filter-Embedded Feature Ranking Approach (FEFR). This proposed approach is performed in two main phases. The first phase is to carry out a thorough analysis of the effectiveness and capability of various feature ranking techniques in ATS drug identification. Six feature ranking techniques were used: Information Gain (IG), Gain Ratio (GR), Symmetrical Uncertainty (SU), Support vector machine based recursive feature elimination (SVM-RFE), and Variable Importance based random forest (VI-RF). The selected feature subset by each of the selected feature ranking technique were run through five different popular classifiers: Random forest (RF), Naïve Bayes (NB), IBK, Sequential Minimal Optimization (SMO), J48, and their performances were analyzed and compared. Experiments on the dataset showed that ReliefF and VIRF performed the best among the other techniques in retaining the significant features and eliminate the irrelevant features. For the second phase, the results of these two top performers in the analysis will be selected and aggregate to gain benefit from their advantages whilst minimize their shortcomings to yield a more reliable result. All the performance is evaluated in term of the number of features selected and classification accuracy. Paired t-test also carry out to further validated the quality of the FEFR based on the classification accuracy performance metric. The results show that the feature subset selected by the FEFR feature selection approach is either superior or at least as adequate as those subsets that selected by the individual feature ranking method and the original dataset.

ABSTRAK

Teknik-teknik analisis makmal yang sedia ada untuk mengenalpasti dadah ATS mempunyai cabaran-cabaran tersendiri termasuk kos latihan pakar-pakar analisis, kos bahan-bahan yang diperlukan dan bahaya dalam mengendalikan eksperimen. Tambahan pula, dengan berkembangnya reka bentuk dadah-dadah ATS yang baru secara berterusan dalam pasaran gelap, ia sekaligus menjadikan satu cabaran dalam menyasarkan cadangan pendekatan kecerdasan berkomputer bagi mengesan dan mengesahkan sebatian-sebatian tersebut. Data set struktur molekul 3D dadah ATS perwakilan data set yang digunakan dalam experiment analisis. Ia terdiri daripada 7212 rekod sampel yang mana terdapat 1185 fitur. Kajian ini menyiasat pelbagai kompleksiti dan ketidakpastian yang berada dalam data set berbentuk dimensi yang tinggi dan kewujudan fitur yang tidak relevan dan rosak. Cabaran-cabaran ini telah mendorong kajian ini untuk menyelesaikan masalah dengan mengurangkan dimensi data set dan pemilihan subset fitur penting daripada data set. Maka, ianya membawa kepada cadangan pendekatan pemilihan fitur-fitur untuk membuang fitur yang tidak relevan dan rosak dan memilih subset fitur yang terbaik untuk mewakili dadah ATS dan menghasilkan prestasi pengenalpastian yang lebih baik. Pendekatan pemilihan fitur-fitur yand dicadangkan mempunyai kerangka kerja algoritma yang ringkas dan menggunakan teknik pemilihan fitur yang sedia ada untuk menyelesaikan pelbagai isu-isu data yang berbeza iaitu Pendekatan Kedudukan Kelompok Fitur Turas-Benam (KFTB). Pendekatan cadangan ini terbahagi kepada dua fasa. Pada fasa pertama, analisis yang teliti tentang keberkesanan dan kemampuan pelbagai teknik kedudukan fitur-fitur dalam pengenalpastian dadah ATS dilaksanakan. Terdapat enam jenis teknik fitur kedudukan yang digunakan: Information Gain (IG), Gain Ratio (GR), Symmetrical Uncertainty (SU), Support Vector Machine based recursive feature elimination (SVM-RFE), dan Variable Importance based random forest (VI-RF). Fitur subset terpilih bagi setiap teknik fitur kedudukan terpilih diuji menggunakan lima jenis pengelas berbeza iaitu: Random Forest (RF), Naïve Bayes (NB), IBK, Sequential Minimal Optimization (SMO) dan J48. Kemudian prestasi fitur subset terpilih dianalisis dan dibandingkan. Eksperimen yang dilaksanakan pada data set telah menunjukkan teknik ReliefF dan VIRF telah menghasilkan keputusan yang terbaik di antara teknik-teknik yang lain dengan mengekalkan fitur penting dan membuang fitur yang tidak relevan. Pada fasa kedua, keputusan analisis teknik-teknik dua teratas akan dipilih dan digabungkan untuk mendapatkan manfaat daripada kelebihan mereka manakala meminimumkan kekurangan bagi menghasilkan keputusan yang boleh dipercayai. Semua prestasi dinilai dari segi bilangan fitur-fitur yang terpilih dan akurasi klasfikasi. T-ujian berpasangan juga dilaksanakan untuk mengesahkan kualiti KFTB berdasarkan akurasi klasfikasi prestasi metrik. Keputusan-keputusan menunjukkan subset fitur terpilih menggunakan KFTB sama ada lebih bagus atau sekurang-kurangnya mencukupi seperti subset yang dipilih oleh kaedah fitur kedudukan individu dan dataset asal.

ACKNOWLEDGEMENTS

I would like to extend my sincere appreciation to my supervisor Prof. Madya Dr. Azah Kamilah Muda @ Draman, who has always been patient and available especially in those times that I felt overwhelmed. Your feedback and suggestions have been so helpful and I sincerely have learnt a lot from you. Thank you for believing in me particularly when I had doubts about this whole project. Also thank to Dr. Zeratul Izzah Mohd Yusoh for all her helps, support, and valuable hints.

Many thanks go to the Faculty of Information and Communication Technology for providing me a great environment to study and do research. I have learnt so much from seminars and research presentations that were held in the faculty. In addition, generous helps on study and research are provided by professors and staff in the department. Especially, providing me the facility of hardware during the period of research.

I would like to express my very great appreciation to all my friends for their care, support and encouragement to me. We shared our memorable moments together in the past two years.

Last but not least, special thanks must also go to my parents and my entire family for providing me unconditional support and encouragement throughout my time in graduate school.

TABLE OF CONTENTS

		PAGE
DECLARATION		
APPROVAL		
DEDICATION		
ABSTRACT		**i**
ABSTRAK		**ii**
ACKNOWLEDGEMENTS		**iii**
TABLE OF CONTENTS		**iv**
LIST OF TABLES		**vi**
LIST OF FIGURES		**viii**
LIST OF APPENDICES		**x**
LIST OF ABBREVIATIONS		**xi**
LIST OF PUBLICATIONS		**xiii**
CHAPTER		
1.	**INTRODUCTION**	**1**
1.1	Introduction	1
1.2	Research Background and context	2
1.3	Problem Statements	6
1.4	Research Questions	8
1.5	Research Objectives	8
1.6	Research Significance	9
1.7	Research Scopes	9
1.8	Thesis Organization	10
1.9	Summary	13
2.	**LITERATURE REVIEW**	**14**
2.1	Introduction	14
2.2	Amphetamine-type Stimulant Drug (ATS)	14
2.2.1	Abuse of ATS Drug	16
2.3	Traditional Drug Analysis	18
2.3.1	Screening Analysis	18
2.3.2	Confirmatory Analysis	22
2.4	Related Works of Traditional Methods in ATS Drug Identification	24
2.5	Related Work of Computational Intelligence Solution	26
2.5.1	Signature and Handwriting Recognition	27
2.5.2	Face Recognition	29
2.5.3	Gait Recognition	31
2.5.4	ATS Drug Identification	32
2.6	Feature Selection	34
2.7	Ensemble Feature Selection	38
2.7.1	Related Works Ensemble Feature Selection	39
2.7.2	Choice of Methods	41
2.8	Classification	46
2.9	Performance Measurement	50
2.10	Cross Validation	51
2.11	Statistical Validation	52
2.11.1	Shapiro-Wilk Test and Normal Quantile-Quantile (Q-Q) Plot	53

2.11.2 Testing for Differences in Dependent Populations 54
2.11.3 Testing for Differences Between Two Independent Populations 54
2.11.4 Differences in Three or More Independent Populations 55
2.12 Summary 55

3. RESEARCH METHODOLOGY 57
3.1 Introduction 57
3.2 Problem Situation and Solution Concept 58
3.2.1 Problem Situation 58
3.2.2 Solution Concept 59
3.3 Overview of Research Methodology 60
3.3.1 Investigation Phase 62
3.3.2 Implementation Phase 63
3.3.3 Processing Phase 64
3.3.4 Post-processing Phase 66
3.4 Operational Framework 73
3.5 Environmental Setup 74
3.5 Summary 75

4. COMPARATIVE ANALYSIS OF FILTER-EMBEDDED FEATUERE RANKING TECHNIQUES 76
4.1 Introduction 76
4.2 Overview of Experiment Design and Procedure 77
4.3 Experimental Results 78
4.3.1 Performance Analysis 78
4.4 Discussion 85
4.5 Summary 87

5. ENSEMBLE FILTER-EMBEDDED BASED FEATURE RANKING (FEFR) APPROACH 88
5.1 Introduction 88
5.2 Overview of Proposed Method 89
5.3 Examine the Quality of Selected Features by FEFR 91
5.4 Performance Analysis and Discussion 95
5.5 Statistical Validation 98
5.6 Summary 102

6. CONCLUSION AND FUTURE WORKS 103
6.1 Research Summary 103
6.2 Research Contributions 104
6.3 Conclusion 105
6.4 Research Limitations and Future Work 106
6.5 Summary 107

REFERENCES 108
APPENDIX A 120

LIST OF TABLES

TABLE	TITLE	PAGE
2.1	Colour test results with common reagents for ATS and their pre-cursors. Source from (UNODC, 2006)	19
2.2	Comparison of the general characteristics of forensic drug analysis	24
2.3	Summary and comparison of feature selection methods	37
2.4	Pseudocode for the SVM-RFE algorithm, taken from Guyon et al. (2002)	44
2.5	Pseudocode for calculating feature relevance by using Random Forests, adapted from Kocev (2011).	46
2.6	The confusion matrix structure returned by a classifier	51
3.1	Summary of investigation phase	62
3.2	Summary of the dataset	64
3.3	Default parameter of different classifiers in Weka (Witten et al., 2011)	70
3.4	Overall Research Plan	72
4.1	Performance of classification models for original ATS drugs dataset	79
4.2	Classification accuracy (ACC) values of different feature ranking techniques using random forest (RF) Classifier	80
4.3	Classification accuracy (ACC) values of different feature ranking techniques using Naïve Bayes (NB) Classifier	80
4.4	Classification accuracy (ACC) values of different feature ranking	80

techniques using IBK Classifier

4.5 Classification accuracy (ACC) values of different feature ranking techniques using SMO Classifier 81

4.6 Classification accuracy (ACC) values of different feature ranking techniques using J48 Classifier 81

4.7 Average overall classification accuracy of the ATS drug dataset based on different number of features 82

4.8 Average overall classification accuracy on ATS drug dataset based on different feature ranking techniques 84

5.1 Three Different Ensemble Methods 93

5.2 Average Overall Classification Accuracy (%) and number of features selected by Ensemble 1, Ensemble 2 and Ensemble 3 on ATS drug dataset 94

5.3 Average Overall Classification Accuracy (%) of original dataset, ReliefF, VI-RF, and FEFR on ATS drug dataset 97

5.4 Normality tests for Original dataset, ReliefF, VIRF, FEFR 101

5.5 Descriptive results for Original dataset, ReliefF, VIRF, FEFR 101

LIST OF FIGURES

FIGURE	TITLE	PAGE
1.1	Motivation of the study	5
1.2	Thesis Organization	12
2.1	Structure of β-phenethylamine	15
2.2	Basic structure of ATS	15
2.3	Different types of substituted Amphetamine	16
2.4	Crystalline methamphetamine seizures reported in East and South-East Asia, 2009-2013. Source(s): Drug Abuse Informa.on Network for Asia and the Pacific (DAINAP). UNODC Annual Report Questionnaire for the respective years and countries, 2009-2013.	17
2.5	MS Components and functions	23
2.6	Researches in forensic investigation	34
2.7	General feature selection process	35
3.1	Research design	61
3.2	Framework of proposed feature selection approach	66
3.3	Data collection for training and testing	71
3.4	Operational Framework	74
4.1	Flow chart of the experimental design	78
4.2	Performance of classification models for original ATS drugs dataset	79
4.3	Average classification accuracy based on different number of subset	83

sizes

4.4 Average classification accuracy using different feature ranking techniques 85

5.1 FEFR Algorithm 90

5.2 Flowchart of Proposed FEFR Approach 91

5.3 Experimental results using different ensemble methods based on the number of selected features 95

5.4 Experimental results using different ensemble methods based on classification accuracy (%) 95

5.5 Experimental results using different feature ranking techniques based on the classification accuracy 98

5.6 Experimental results using different feature ranking techniques based on the number of features selected 98

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	Summary of Dataset	120

LIST OF ABBREVIATIONS

ATS	-	Amphetamine-type Stimulant
UNODC	-	United Nations Office on Drugs and Crime
HPLC	-	High-performance liquid chromatography
LC	-	Liquid chromatography
GC	-	Gas chromatography
MP	-	Mobile phase
MS	-	Mass spectrometry
COSEFOS	-	Common Scheme for Evaluation of Forensic Software
IG	-	Information Gain
SU	-	Symmetrical Uncertainty
GR	-	Gain Ratio
SVMRFE	-	Support Vector Machine Recursive Feature Elimination
VIRF	-	Variable Importance based Random Forest

RF	-	Random Forest
NB	-	Naïve Bayes
SMO	-	Sequential Minimization Optimization
WEKA	-	Waikato Environment for Knowledge Analysis
FEFR	-	Filter-Embedded Feature Ranking

LIST OF PUBLICATIONS

Saw, Y.C., Yusoh, Z.I.M, M., Muda, A.K., and Abraham, A., 2017. Ensemble Filter-Embedded Feature Ranking Technique (FEFR) for 3D ATS Drug Molecular Structure. *International Journal of Computer Information Systems and Industrial Management Applications*, 9, pp.124–134.

Saw, Y.C. and Muda, A.K., 2016. An Overview Computational Intelligence Solution in Forensic Drug Analysis. *Journal of Network and Innovative Computing*, 4, pp.272–279.

Saw, Y.C., Muda, A.K, 2016, "An overview of computational intelligence technique in drug molecular structure identification," *Innovations in Bio-Inspired Computing and Applications, Springer International Publishing,* 473-480.

Saw, Y.C., Muda, A.K., and Yusoh, Z.I.M, 2016, "Comprehensive Analysis of Significant Features Determination for ATS Drug Identification," in *International Conference on Telecommunication, Electronic and Computer Engineering (ICTEC 2017)*. [Accepted]

CHAPTER 1

INTRODUCTION

1.1 Introduction

The widespread of abuse of illicit manufacture and trafficking of Amphetamine-type stimulants (ATS) poses a serious risk to the national security. ATS drug is considered as one of the most widely illicit used drug besides cannabis, cocaine, and heroin. This is due to the ready availability of the ingredients, the high flexibility of the manufacturing processes which can return a high profitability to the organized criminal groups (UNODC, 2013). Thus, these phenomena present a unique challenge to the law enforcement authorities and to the scientific staff of forensic laboratories due to the advance in the illicit manufacture of the new and unfamiliar type of ATS drug.

Today, as the increasing of the computing power and it's become more affordable, it has been successfully attracting various research from various domains, such as Bioinformatics. cheminformatics, ethology, cognitive science, etc. Therefore, this presents an alternative to mine the unknown pattern of ATS drug and extract the relevant data for knowledge discovery and decision-making process. Hence, the key concern of this research is to adopt the computational intelligence solution to cater the limitation of the current laboratory process. However, due to the ever increasing of the complexities of ATS drug molecular structure, the extracted data will be in high dimensional dataset. This may present a significant challenge to the learning algorithms as not all the features in the enormous dataset are relevant and significant for the learning algorithm. Therefore, this problem is addressed in this study to acquire the significant features that can reflect the

characteristics of the ATS drug, while eliminating the irrelevant and noisy data from the dataset. In short, this study is aimed to propose a research and development of a novel approach of feature selection methods in the ATS drug domain.

In this chapter, a brief introduction and the research background to the relevant topic will be discussed in Section 1.2. The important issues and problem that exists in the domain of this research are explained in Section 1.3. Furthermore, several specific questions that will answer by this study have been identified and presented in Section 1.4. Section 1.5 will present the objective to be achieved in this research. Next, the significance of this research, and the research scope is covered in Section 1.6 and Section 1.7 respectively.

1.2 Research Background and Context

ATS drugs, encompass a group of drugs which consists of amphetamines (amphetamine and methamphetamine) and substances of the "ecstasy"-group (MDMA, MDA, MDEA, etc.). In recent year, abuse of Amphetamine-type Stimulants (ATS) drugs globally spread in the market. According to the Trends and Patterns of Amphetamine-type Stimulants and New Psychoactive Substances report that reported by United Nations Office of Drugs and Crime (UNODC, 2015), between the year 2009 to 2013, the most common destination for ATS seized were located in East and South-East Asia and Oceania, specifically Australia, Japan, Malaysia, Russian Federation and the United Kingdom.

To date, international attention is focusing much on how to prevent or control the spreading of ATS drugs, reduction of supply of these illicit drugs and the treatment that caused by the consequences of ATS drug abuse. These areas of activity are essential however, it cannot resolve the new situation that faced today, which is the constantly

emerging of illicit manufacture of new and unfamiliar ATS drugs, or their combination, and their trafficking trends onto the illicit drug market. This phenomenon has become one of the most worrisome issues of the National law enforcement authorities. In an effort to reduce the abuse of ATS drugs, a meeting was held in London in September 1998 by UNODC's Laboratory and Scientific Section with the cooperation together with Forensic Science Service of the United Kingdom to assess the identification and analysis methods for ATS drugs and also the ring-substituted analogues in seized materials (UNODC, 2006). However, the process of experimental studies to detect the ATS drugs are slow, expensive and cannot be covered for a wider range of ATS drugs. Therefore, there is an urgent need to find a solution to cope with the limitations that present in the current experimental studies.

With the rapidly emerging in the computer technologies in recent years, computational intelligent solution becomes more affordable and ideal to cater the problems of the current experimental studies. In this case, the ATS drug will be represented in the 3D molecular structure, which gives better context perception and well suited for visualizing and handling of large numbers of objects (Vion-Dury and Santana, 1994). The 3D geometric shapes of ATS drug molecular structure is described numerically using molecular descriptor (Pratama et al., 2017). Since there are thousands of compounds present in one ATS drug element, the data set that output from the feature extraction phase will be complex and in high dimensions. This presents a significant challenge in computational intelligence techniques as it required a high computational cost for the learning process, or worse if the data contains high level of irrelevant and noisy data.

Thereby, the key concern of this study is to cater the problem of the nature of the dataset, that is concerned in improving the classification performance and to discover the significant features that exist in the ATS drug molecular structure. This objective can be

achieved by using feature selection methods as it is similar to the feature selection purpose. Specifically feature selection methods can be categorized into three main approaches, which are filter, wrapper, and embedded (Guyon, 2003). Filter approach will evaluate the relevancy of the features based on the intrinsic characteristic of the features without relying on the classifiers. In contrast, the wrapper approach does rely on the classifier's decision in selecting the relevant features. The selected feature subset will be evaluated during every repetition of the evaluation process, which may cause computationally intensive compared to the filter approach. Nevertheless, the wrapper approach provides a better performance compared to the filter approach, although it is computationally demanding. In contrary, embedded approach incorporates the feature selection process into the process of classifier construction. It is attempting to compensate the computation time taken up in the wrapper approach. Therefore, an ensemble of the filter and embedded approach in order to explore the knowledge and advantages of each approach, while mitigating their weakness to yield a more stable and accurate result.

The selected feature subset will then be validated based on the identification performance, which is the common performance measurement that is used to evaluate the quality of the selected feature subset. A high identification performance that yields from the learning algorithm indicates the selected features contain the high discriminative power to identify the sample instances. This is used to demonstrate the ability of the selected features to distinguish the class label associated with the sample data. This result will use to illustrate the solving capability of far-reaching problems, such as the deficiency in the traditional laboratory process. Figure 1.1 illustrates the motivation of the study in the ATS drug domain.

Trends leading to the problem

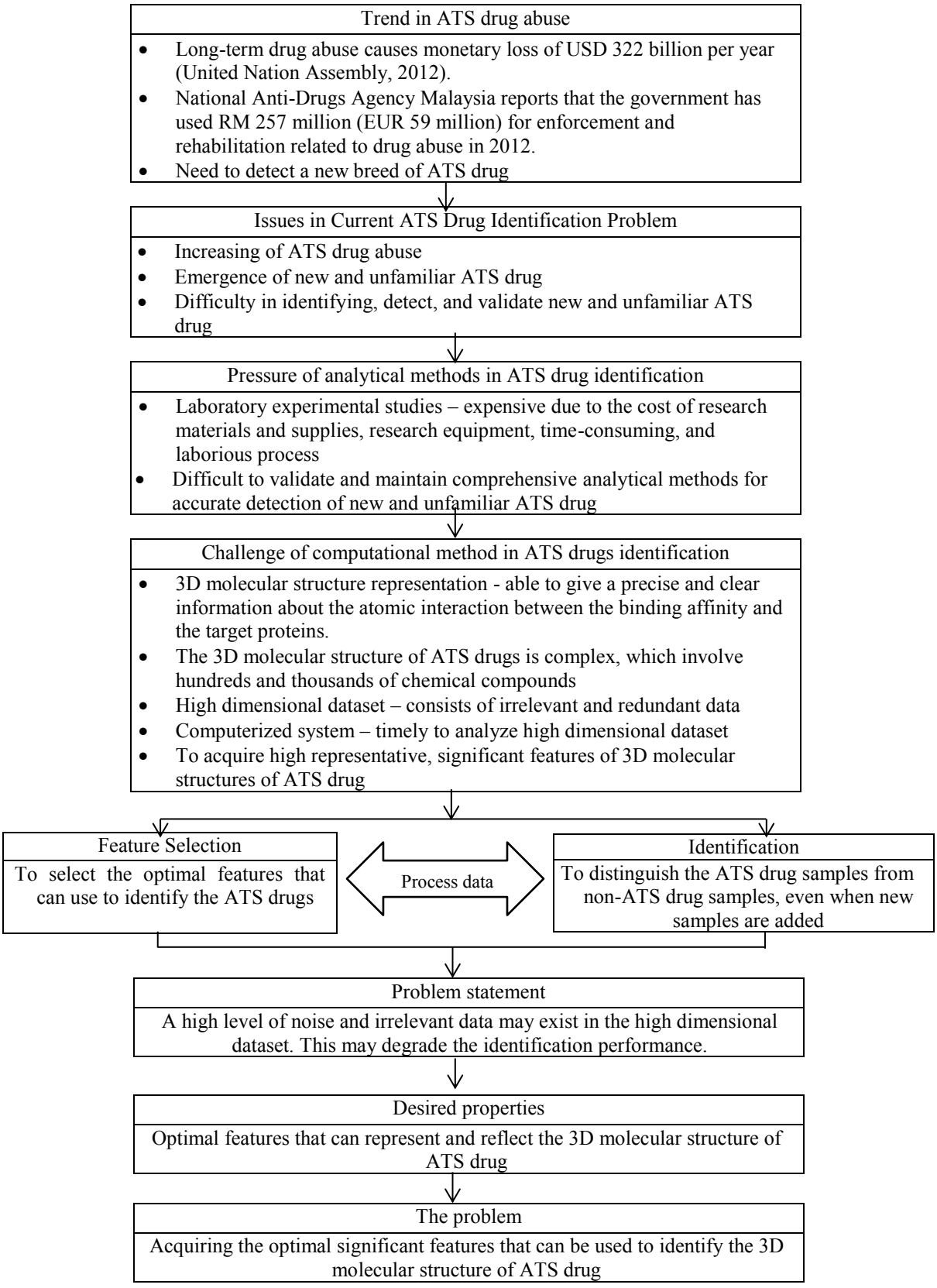

Figure 1.1: Motivation of the study