Development of Artificial Intelligence-based In-Silico Toxicity Models. Data Quality Analysis and Model Performance Enhancement through Data Generation

Development of Artificial Intelligence-based In-Silico Toxicity Models for Use in Pesticide Risk Assessment

Data Quality Analysis and Model Performance Enhancement through Data Generation

Ladan Malazizi

University of Bradford
School of Informatics

2004-2008

ABSTRACT

Toxic compounds, such as pesticides, are routinely tested against a range of aquatic, avian and mammalian species as part of the registration process. The need for reducing dependence on animal testing has led to an increasing interest in alternative methods such as *in silico* modelling. The QSAR (Quantitative Structure Activity Relationship)-based models are already in use for predicting physicochemical properties, environmental fate, eco-toxicological effects, and specific biological endpoints for a wide range of chemicals. Data plays an important role in modelling QSARs and also in result analysis for toxicity testing processes. This research addresses number of issues in predictive toxicology. One issue is the problem of data quality. Although large amount of toxicity data is available from online sources, this data may contain some unreliable samples and may be defined as of low quality. Its presentation also might not be consistent throughout different sources and that makes the access, interpretation and comparison of the information difficult. To address this issue we started with detailed investigation and experimental work on DEMETRA data. The DEMETRA datasets have been produced by the EC-funded project DEMETRA. Based on the investigation, experiments and the results obtained, the author identified a number of data quality criteria in order to provide a solution for data evaluation in toxicology domain. An algorithm has also been proposed to assess data quality before modelling. Another issue considered in the thesis was the missing values in datasets for toxicology domain. Least Square Method for a paired dataset and Serial Correlation for single version dataset provided the solution for the problem in two different situations. A procedural algorithm using these two methods has been proposed in order to overcome the problem of missing values. Another issue we paid attention to in this thesis was modelling of multi-class data sets in which the severe imbalance class samples distribution exists. The imbalanced data affect the performance of classifiers during the classification process. We have shown that as long as we understand how class members are constructed in dimensional space in each cluster we can reform the distribution and provide more knowledge domain for the classifier.

Keywords: artificial intelligent, data quality, data generation, model performance, toxicity, QSAR, classification algorithm, clustering, imbalanced dataset, endpoints

ACKNOWLEDGEMENTS

I would like to express my gratitude to all those who gave me the possibility to complete my PhD.

First I would to like to give my special thanks to my supervisor Dr. Daniel Neagu for his continuous support, guidance and encouragement throughout my work. He was not only a supervisor but a friend who I could trust and rely on when I had problems. He was always there to listen and to give me advice. He is responsible for involving me in the project in the first place. He taught me how to ask questions and express my ideas. He showed me different ways to approach a research problem and the need to be persistent to accomplish any goals. He taught me all I know and I would be forever grateful.

My special thanks also to Dr. Qasim Chaudhry for selecting me for this project and gave me countless support and encouragement. He gave me the opportunity to study my PhD. He was always there to make sure everything works smoothly within SeedCorn project from financial point of view and my study. Without financial help from his project and his countless support none of these was possible.

I want to thank the Central Science Laboratory for providing me with the funding which enabled me to study PhD in the first place and also gave me permission to use DEMETRA project's (Quality of Life and Management of Living Resources QLK5-CT-2002-00691) data for my experimental work.

I would also like to thank Professor Graves-Morris for his support and guidance throughout my study. Considering his knowledge, experience and expertise, his position as my second supervisor always gave me great comfort in my study and peace of mind.

I am grateful to the administration staff of School of Informatics, especially Rona Wilson, for her extensive support in ensuring all the related paper works are prepared and stored efficiently.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

BOOK CHAPTERS

Malazizi, L. et al. 2007. Contribution to the multi-authored book Quantitative Structure-Activity Relationship (QSAR) for Pesticide Regulatory Purposes, E. Benfenati (ed.), Elsevier, ISBN-10: 0-444-52710-9

LIST OF JOURNAL PAPERS

Malazizi, L., Neagu, D. Chaudhry, Q. 2008. A review on improving imbalanced multidimensional dataset learner performance with Density-based Class-Boost Algorithm, Expert System (submitted)

Malazizi, L. Neagu, D. Chaudhry, Q., 2007. A Data Quality Assessment Algorithm with Application in Predictive Toxicology (extended version), TASK Quarterly Journal (Scientific Bulletin of Academic Computer Center in Gdansk, Poland), ISSN 1428-6394, volume 11, number 1-2/2007 Advances in Artificial Intelligence

LIST OF CONFERENCE PAPERS

Malazizi, L. Neagu, D. Chaudhry, Q., 2008. Improving Imbalanced Multidimensional Dataset Learner Performance with Artificial Data Generation: Density based Class-Boost Algorithm, Proceedings of the 8th Industrial Conference on Data Mining (ICDM), Germany, July 2008, 165-176, ISBN 978-3-540-70717-2, *Springer LNAI 5077*, Leipzig, Germany

Malazizi, L. Neagu, D. Chaudhry, Q., 2006. A Data Quality Assessment Algorithm with Application in Predictive Toxicology, Proceedings of Symposium Advances in Artificial Intelligence and Applications (AAIA'06), Wisla, Poland, ISSN: 1896-7094, 131– 140

Malazizi, L. Neagu, D. Chaudhry, Q., 2006 Investigation, Assessment and Identification of Possible Data Quality Criteria in Predictive Toxicology, Proceedings of the 6th UK Workshop on Computational Intelligence (UKCI), University of Leeds, 229-236

Malazizi, L., Neagu, D., Chaudhry, Q., 2006. Investigating Data Quality Assessment In Predictive Toxicology, 7th Informatics Workshop for Research Students, University of Bradford, 29th March 2006, 129-132, ISBN 1-85143-2329

Malazizi, L. , Neagu, D. Chaudhry Q., 2005. Investigating Data Quality Assessment In Predictive Toxicology, 6th Informatics Workshop for Research Students, University of Bradford, 29th March 2006, 129-132, ISBN 1-85143-2329

LIST OF PRESENTATIONS

- Presentation-Meeting at Central Science Laboratory(2005)
- PhD Students Research Presentation (August 2006)
- Presentation-Meeting at Central Science Laboratory (September 2006)
- Poster Presentation at CSL –workshop (March 2006)
- Research Students Seminar at University of Bradford (August 2006)
- Presentation of my work as Erasmus exchange student at the Human Computer Interaction Group, University of Patras, Greece (Jan-March 2007)

1. INTRODUCTION

1.1 Project Description and Objectives

Problem Background:

Toxic compounds, such as pesticides, are routinely tested against a range of aquatic, avian and mammalian species as part of the registration process. The need for reducing dependence on animal testing has led to an increasing interest in alternative methods such as *in silico* modelling. These models can quantify molecular properties of a compound to predict a specific physiological effect e.g. a toxicological endpoint in a given animal species. The QSAR (Quantitative Structure Activity Relationship)-based models are already in use for predicting physicochemical properties, environmental fate, eco-toxicological effects, and specific biological endpoints for a wide range of chemicals. Most of the models are based on a particular statistical method such as regression, partial least squares and principle component analysis.

There are a number of limitations to this conventional QSAR approach. One needs to develop a separate model for each class (and sometimes for each sub-class) of chemicals. The ability of conventional QSAR approaches in handling complex data is usually limited. The outcome is often hard to generalise across different compounds or test species.

For the reasons mentioned above, the emphasis amongst molecular modelling community has shifted in recent years towards the development and use of generalised approaches that are applicable across chemical classes or test species. More powerful computational approaches have also become available recently (expert systems, Bayesian networks, machine learning etc). This has opened up new avenues for identifying relational patterns within complex datasets, and thereby overcoming the limitations associated with conventional QSARs [1].

Problem Description:

Data plays important role in modelling QSARs and also new finding in all the processes of toxicity testing. Data quality and modelling are the main issues for this project. First we need to identify the data quality problems in order to define an assessment framework. Then modelling the multidimensional and imbalance data with high accuracy would be the second step. In order to achieve these objectives, the following milestones are identified.

Project Objectives:

- Design a prototype in order to collect toxicity data from online sources. The toxicity data would be studied from structural and representational points of view.
- Study and investigate the quality of the data collected from different sources or within one source calculated by different software tools in order to identify deficiencies and highlight quality criteria applicable to the toxicity data sets.
- Define algorithm to fill in missing values, increase data size and consequently model performance in toxicity datasets.
- Define algorithm to generate artificial data in imbalanced datasets and improve model performance based on each dataset characteristics.

In this work we have used number of toxicity data files for our experimental work which are kept confidential by University of Bradford-Central Science Laboratory contract thus no quantitative information is provided in thesis analysis.

1.3 Thesis Structure

The thesis is composed of nine chapters and one appendix.

Chapter 1 (Introduction) presents the project, stating problem statement, problem description, problem's background and project objectives.

Chapter 2 (Data in Predictive Toxicology) includes the project description and objectives and also the result of comparison investigation on online toxicology data representation, structures and values presented and literature study of this issue (the work done by other people). Some of these results have been produced in first and second paper represented to the University of Bradford workshop.

Chapter 3 (Data Quality Assessment) discusses some of the data quality assessment methods in different domains and their relevance to toxicology domain. Also we discuss the missing value problem and some of the existing solutions.

Chapter 4 (A study on data quality in toxicology and new algorithm for data quality assessment process) starts with looking at DEMETRA datasets for five endpoints and highlights the deficiencies in the data values and presentation. The data has been modelled using eight different algorithms using the Weka data mining tool and comparison study between the two datasets for the same endpoints has been carried out to show how the values differences affect the model performance. The results of this work has been summarized and presented at the UKCI international conference. Also this chapter investigates the data quality assessment criterion in toxicology

domain, which has been drawn as a result of experimental work. As a result a procedural algorithm has been proposed. All the experimental work of modelling the data on five endpoints is produced. The result of the work has been presented and published at AAIA international conference, also the extended version with the result of more experimental work has been published in TQ (Task Quarterly) computer science Polish national journal.

Chapter 5 (A New Algorithm for Missing Value Generation in Toxicity Datasets) explains about using the old method of Least Square Method to generate artificial data to fill missing values in empty cells and reconstructing and modelling data. The effect of this action has been analysed in different conditions: when there are one or more version of the same dataset exists. An algorithm to show the procedural process has been proposed.

Chapter 6 (Artificial Data Generation, Data Characteristics and Model Performance) proposes an algorithm in order to generate artificial data in a way that boost the learner performance. With the use of clustering algorithms, ROC analysis and Probability Density Function, the core of potential cluster or class is identified and then artificial data added to boost the class. An algorithm has been proposed in chapter 8. The results of this work have been published in ICDM international conference.

Chapter 7 (Conclusions and Future Work) explains conclusions of the work performed for the duration of this project and also the future work and the possible extension routes have been suggested. In this chapter original contributions have also been listed.

The appendix presents details on analysis and design of the software prototype for toxicology data management.

2. DATA IN PREDICTIVE TOXICOLOGY

Predictive toxicology, the science of developing *in silico* models for toxicity prediction is an important field for chemical and pharmaceutical industry, regulatory bodies and environmental protection agencies. In this domain the use of experimental data for Quantitative Structure Activity Relationship modelling [2] is a primary task. Quantity Structure Activity Relationship is relating aspects of chemical compound structure to biological activities against various endpoints in order to predict chemical toxicity of new compounds. The sensitivity of this procedure proves how vital the data is in this domain.

2.1 Data Description in Predictive Toxicology

Public toxicity databases are a valuable source of information of available toxicity data. These databases have been scattered across public and private sources. They offer bank of chemicals and chosen endpoints that are in place for use by public, scientists, government and industry. But the main problem with databases is that they don't have a standard format and contain different types of descriptive information. A major problem with many of them is also that they don't contain chemical structure information. One of the examples of such a database is IRIS (Integrated Risk Information System) [3]. This database like most other toxicity databases is searchable and indexed by common chemical names and/or CAS number (Chemical Abstract Service Registry number). Although CAS identifiers are unique, they are subject to transcription, typing and formatting errors, and have no chemical meaning. In contrast, chemical structures have universally understood scientific content. Linkage of chemical structures with chemical toxicity information is very important issue in designing (quantitative) structure-activity relationship (Q) SAR models. This models are used for chemical compounds toxicity predictions [4, 5].

The generality, quality and usability of toxicity databases highlight the importance of the data representation from various points of view. Their data quality, structure and format, data availability and accessibility are the issues that need proper attention in order to produce reliable projects to mine information related to chemical toxicity.

The effort of environmental agencies to organize and manage toxicity databases lays on standardization of the elements of these data in order to improve their integrity and reliability. One of these organizations is National Institute of Standards and

Technology [6] which focuses on producing a common vocabulary and standardization of weights, measures, names and symbols to scientific enterprises and agreement of a data file terminologies. Another issue that raises the importance of the matter further is the use of this data for Quantitative Structure Activity Relationship modeling method or relating aspect of compound structure to biological activities in order to predict chemical toxicity of new compounds. Data analysis and integration for producing models using data mining/ machine learning techniques also rely on quality of data. The idea of developing Artificial Intelligence (AI) in-silico modeling for toxicity prediction is also main interest to regulatory bodies and environmental protection agencies that encourages a non-animal alternative to toxicity testing [7].

2.1.1 Toxicology Data Dimensions

Toxicology is the study of the harmful interactions between chemicals and biological systems. Toxic chemicals have varying degrees of activity in biological systems. The level of toxicity, however, depends on the level and type of exposure. Exposure also depends on how the toxic is introduced to the organism. If gaseous, it is likely to be inhaled or absorbed through organism's surface. Thus toxicological effects are proportional to both dose and exposure time, and can be acute (short term), sub-chronic (mid-term) or chronic (long-term) [8]. In toxicology databases compounds are listed with the number of properties, which explain all details about that specific compound and also the toxicity information. Some of the keywords used in these databases are explained below:

-Descriptor: an element that describes a specific property of the compound.

-Chemical compound: is a substance formed from two or more elements, with a fixed ratio that determines its composition.

-Endpoints: a biological effect used as an index of the effect of a chemical on an organism.

- Mechanism of Action/Mode (MOA): the way a chemical compound interacts with a living system.

-Dose: a measured amount of a chemical compound. Dose is often expressed in milligrams per kilogram (mg/kg) or parts per million (ppm).

-LD_{50}: the amount of a chemical that is lethal to one-half (50%) of the experimental animals exposed to it. LD_{50}s are usually expressed as the weight of the chemical per

unit of body weight (mg/kg). It may be fed (oral LD_{50}), applied to skin (dermal LD_{50}), or administered in the form of vapours (inhalation LD_{50}) [9]. For aquatic species, this is expressed as lethal concentration i.e. LC_{50} (mg/ litre).

Figure1 shows screen shots of four different toxicology database's user interface, which are displaying information about same chemical compound [10, 11, 12, 13, 14]. Study has shown that toxicity information related to the same compound has various representations in different databases. This variation has lots of reasons that once again prove that data is not reliable.

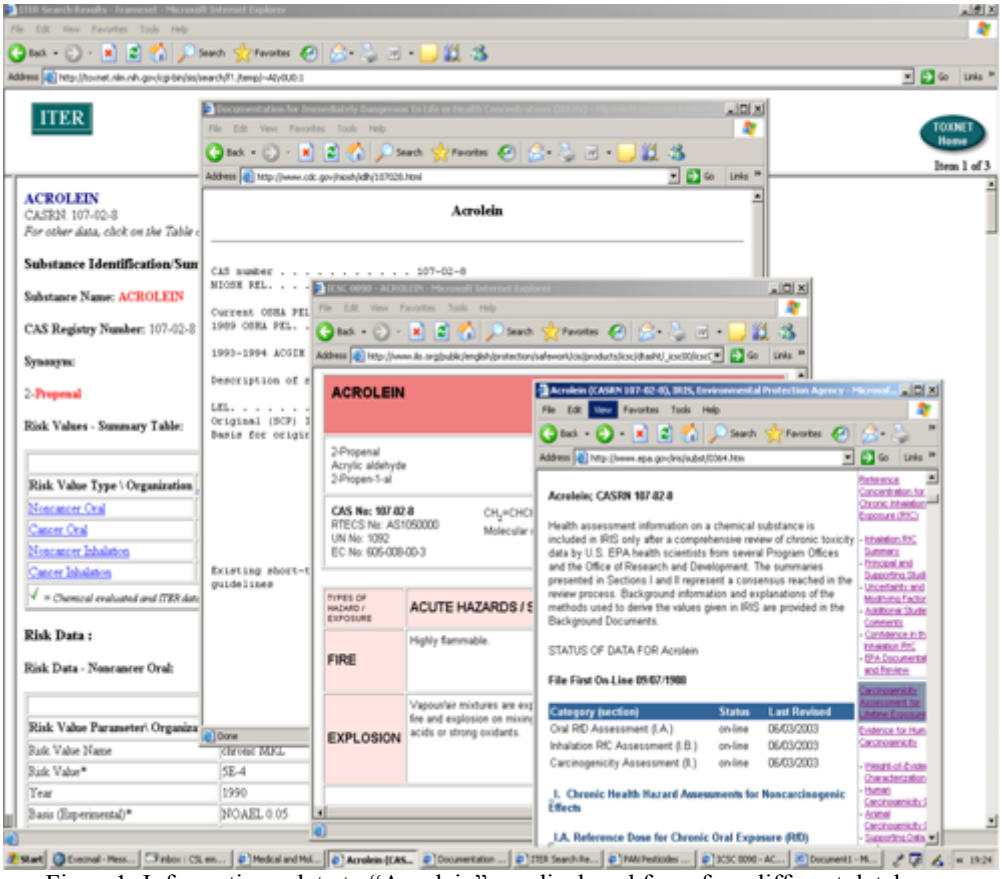


Figure1: Information relate to "Acrolein" are displayed from four different databases.

2.1.2 Descriptors Calculation

In predictive toxicology chemical compounds descriptors are calculated by different software programs. In our project we used the datasets which their descriptors have been generated by ACD and PALLAS. Followings are the description of these programs.

ACD: Advanced Chemistry Development, Inc., (ACD/Labs) is a chemistry software company offering solutions that truly integrate chemical structures with analytical

chemistry information to produce ChemAnalytics™. ACD/Labs creates innovative software packages that aid chemical research scientists worldwide with spectroscopic validation of structures, elucidation of unknown substances, chromatographic separation, medicinal chemistry, preformulation of novel drug agents, systematic nomenclature generation, and chemical patenting and publication. Combined, ACD/Labs' solutions create an analytical informatics system that provides dramatic feed-forward effects on the chemical and pharmaceutical research process. Founded in 1993, and headquartered in Toronto, Canada [15].

***PALLAS*:** Software predicting pKa, logP, logD values and metabolites based on structural formulae of compounds. In the field of industrial pharmacy perhaps the most important physicochemical characteristics of drugs and excipients are their acidity or basicity (expressed by their pKa value), their hydrophobicity and it's dependence on pH (expressed by their logP and logD values respectively).To determine precision values of pKa, logP or logD presents a great deal of work. Thus the use of computers is giving great progress to the practice. The profile in reaction kinetics depends closely on pKa values (1, 2, 3). In preparative chemistry, pKa values of the reaction products can be used to select conditions for synthesis. The knowledge of the pKa values of intermediate products is also important, although they are often very difficult to measure [16].

2.1.3 Toxicology Database Example (DSSTox –Features and Elements)

Distribute Structure-Searchable Toxicity (DSSTox) [17], is a free public accessible site, which has been launched on US Environmental Protection Agency public Internet. This website is connected to four DSSTox databases with all the associated documentations collaborating one with each other within different areas of toxicology and chemical activity. Also the main aim of this project is to adequately meet the data requirements for flexible searching, (Q) SAR model development and building of Chemical Relational Databases (CRD). This database has three major elements:

1) “Adopt and encourage the use of a common standard SDF (Structure Data File) file format for public toxicity databases that includes chemical structure, text and property information. It was already adopted as an industry standard import/export feature of chemical modeling and CRD applications. The latter refers to a computer application

that provides for storage of chemical records containing structures and text/data fields".

2) "Implement a distributed source approach that will enable decentralized; free public access to toxicity data files and that will effectively link toxicity data sources with potential users and modelers of these data from other disciplines. The DSSTox Source refers to the person or organization that compiled and currently maintains a public toxicity database for which a corresponding DSSTox SDF file has been created. The Source is considered as the owner of the data and is responsible for the file's maintenance and upgrade and would be referenced and acknowledged in any subsequent use of that file".

3) "Engage public/commercial/academic/industry groups in contributing to and expanding the DSSTox public database network: A DSSTox Central Website will serve as the hub of the DSSTox project providing general information on DSSTox standard file format, a central index of field names and links to DSSTox Sources and SDF files, CRD vendors and public tools and resources of general interest to the DSSTox community. It also connects the DSSTox user community members and to enlist their help in encouraging the DSSTox recommended standards, reporting DSSTox SDF file errors to the Sources, offering enhancements to existing DSSTox SDF files and aiding in the construction of new DSSTox SDF files "[18].

2.1.4 DEMETRA Datasets

The DEMETRA datasets have been produced by the EC-funded project QLK5-CT-2002-00691. It was focused on the development of environmental modules for evaluation of toxicity of pesticide residues in agriculture.

This project's aim has been to develop predictive models and software which give a quantitative prediction of the toxicity of a molecule, in particular molecules of pesticides, candidate pesticides, and their derivatives. The input for models is the chemical structure of the compound, and the software algorithms use "Quantitative Structure-Activity Relationships" (QSARs).

DEMETRA project has constructed a database for five toxicological end-points. A survey was conducted amongst DEMETRA partners and subcontractors to gather information about databases that contain pesticide toxicity data to determine the availability of sufficient good quality data.

Suitable datasets were defined, with pesticides and their activities. They have collected at least 100 chemicals for each end point and more than 200 chemicals for two endpoints. These datasets are named after the endpoints the values are collected for: DietaryQuail, Bee, Daphnia, OralQuail and Trout; description of the datasets is included in chapter 5 below.

2.1.5 *In silico* Analysis

Toxicity is the one of the most important parts in drug discovery. There are therefore great needs for the techniques that can identify the effects of the untested chemical compounds at the early stage of the product development on the environment. Computer based (*in silico*) techniques are the solution for this problem. These techniques are cost effective and fast without an available compound. *In silico* toxicity predictions are methods to test biological models, drugs and medical experiments using complex computer models rather than costly animal experiments. They model biochemical events relevant for toxicity testing and produce prediction from a training set of experimentally determined data (Data Driven Systems) [19].

2.1.6 *In vivo* and *in vitro* Testing

In vivo testing programs are methods that use laboratory animals. It was initiated in the early 1970s and used to assess the safety of various substances. These tests are carried out for the development of drugs, food additives, pesticides, and industrial chemicals, and in humans using several animal species. *In vivo* studies differ in duration from short-term dosing to lifetime exposure. They include studies to assess the potential of birth defects, as well as multigenerational studies for assessing adverse reproductive outcomes. They are conducted under the Good Laboratory Practices (GLP) guidelines. These guidelines, produced by the U.S. Food and Drug Administration (FDA) and other regulatory agencies, lay out the boundaries within which toxicity studies that are to be used for regulatory purposes will be conducted. Most laboratories conduct toxicology studies within the regulations of the GLP guidelines even if the studies are not going to be used for regulatory purposes [20].

2.1.7 Pesticide Risk Assessment

Pesticides are products that are used to control pests. Some examples of these are: slug pellets, ant powder, weed killers, and rat and mouse baits. Pesticides are not just chemicals; they include a very large range of different types of products. Some of them are natural which are used in farms to protect crops from insect pests, weeds and fungal diseases. However, as pesticides are used to kill unwanted pests, weeds and moulds, they can also harm people, wildlife and the environment. This is why there are strict controls in place over their sale and use. There is a legal obligation to carry out and record a Local Environment Risk Assessment for Pesticides.There are statutory controls on the advertisement, sale, supply, storage and use of pesticides in the UK. Some of these rules indicate that a pesticide should be:

- Safe to humans - the consumer, users and bystanders.
- Safe to the environment - including soil, water and non-target animals and plants.
- Effective - that they control the pest, weed, or disease and that they have no adverse effect on the crops [21].

2.1.8 QSAR Modelling

A QSAR is a model, which relates the biological activities of a series of similar compounds to one or more physicochemical or structural properties of the compounds. In this definition, "similar" means having the same mechanism of action, but not necessarily having a related chemical structure. Quantitative Structure Activity Relationships (QSARs) are mathematical models representing complex relationships between compound's chemical properties and biological activities. They are used to provide prediction of biological activities of untested or unavailable chemical compounds.

When a chemical is administered to an organism, two events must occur for a biological response to be triggered. Firstly, the compound has to be transported to the site of action (the "receptor"); secondly, it must interact with the target in an appropriate manner. Interaction with the target ("receptor") is governed largely by two factors: the size and shape of the xenobiotic, which will control how well the molecule fits the receptor site; and the nature and relative positions of appropriate functional groups on the molecule, which will affect the type and strength of the interaction with complementary groups on the receptor [2].

As it is apparent from above, data plays important role in modelling QSARs and also new finding in all the processes of toxicity testing.

Scientists nowadays use a variety of tools such as TOPKAT [22], DEREK [23] and OASIS [24] to accomplish these models. There are some data quality assurance and standard procedural issues for the data used and also for the model itself, which are as follows:

-Basic QSAR-modelling conditions: include checking data for homogeneity and representatively. Data Homogeneity: means that the used compounds for modelling should have similar chemical and biological properties.

-Representatively: the selection of the training set and the type of compounds included is very important. This set should be able to represent the chemical domain of compounds for study.

-Procedural steps: the model needs to be validated. Valid models are more suitable to predict better. They are characterized by model coefficients proportional to their significance to the modelled process and consistent with fundamental chemical, biological and toxicological knowledge [25].

Above are just some examples of the issues highlighted for QSAR modeling. Of course considering these would assure more reliable model development. There are also some general rules that need to be considered when ones want to select the data from open source databases. Some of these rules are as follows:

-Data selection from experiments that have used standardized procedure or applied Good Laboratory Practice (GLP).

-Data generated from long-term chronic exposures rather than to short-term acute exposures.

-Data measured by a single protocol and same laboratory and by the same worker.

-Free of experimental errors.

-Data generated from studies using a similar route of exposure(s) to the likely ones at the site" [26, 27].

Compiling data for QSARs modeling needs special attention. Since the base of every model is the similarity of group of compounds in structure and activities, the model at the end should be able to project this through simple testing strategies. This is achieved by selection of meaningful, interpretable descriptors.

A. The importance of descriptors selection in modelling

Modelling valuable QSARs is based on two fundamental elements: (i) datasets (collected as a result of experimental testing on selected chemicals under investigations) and (ii) descriptors (values used to describe properties of those chemicals and their effects). Properties also identify the effects and behaviours of chemicals. QSARs project the relationships between these structural descriptors and descriptors of biological effects (such as LD50: the amount of a chemical that is lethal to half (50%) of the experimental animals exposed to it) using the datasets. The reliability and validity of the produced model depends on the quality of the descriptors [20]. If one descriptor is significant in terms of correlation to the toxicity output, two are not necessarily better, particularly when they reflect the same property (are highly correlated). Thus special attention needs to be paid when descriptors are selected to model QSARs. Descriptors that are related to the biological endpoint of the compound have more impact in the produced model. Examples of such molecular properties are hydrophobicity, steric and electronic properties, molecular weight, pKa and so forth. These descriptors provide more insight details about compounds mechanical properties [25].

2.2 Data Integration

The concepts of data integration and data fusion are normally considered the same although they are different. Integration may play a similar role as fusion though it implicitly refers more to concatenation than to the extraction of relevant information [28]. Following sections clarify the meaning of these two terms further.

A. Data Integration

The fast growing technology of the Internet and e-business in recent years has caused an explosion in the amounts and types of information available to enterprises. Data is accessible from different sources and in various formats, which faces the businesses the challenge of information integration [29]. They need to use different tools and techniques to understand the data, extract information from this data and use the result for their progression and future management's decisions making. Roth et al. [29] defines information integration as "a technology approach that combines core elements from data management systems, content management systems, data warehouses and other enterprise applications into a common platform".

It requires combining and matching information in different sources and resolving a variety of conflicts. With the number of data sources growing very fast, data integration would become even more vital in the future [30].

B. Data Fusion

There are a number of definitions published from different sources, definitions such as merging, combining and integrating. Mangolini defines data fusion as a "set of methods", tools and means using data coming from various sources of different nature, in order to increase the quality of the requested information [28].

Based upon the works of Buchroithner [31] and Wald [32] the following definition was adopted best describes the concept: "data fusion is a formal framework in which are expressed means and tools for the alliance of data originating from different sources. It aims at obtaining information of greater quality; the exact definition of greater quality will depend upon the application". However, the definition for data fusion should not be restricted to data output from sensors, methods, techniques or architectures of the systems [28].

2.3 Why Data Integration in Toxicology

Integrated access to information that is spread over multiple, distributed and heterogeneous sources is an important problem in many scientific domains. In toxicology there is an overwhelming amount of data in public and commercial databases available for data analysis and knowledge discovery. The time and cost effective usage of these data is hampered by two main problems: (i) the distribution of relevant data over many heterogeneous data sources and (ii) the quantity of errors and inconsistencies within these sources. The first problem is solved by data integration approaches, which the second problem is tackled by means of data fusion or cleansing [33]. As mentioned in the first chapter data from different sources is needed to be integrated in a central repository, stored, accessed and studied for scientific findings like mining and modeling Q(SAR)s.

Regarding the presentation of the data and data structure an important step has been taken for predictive toxicology data and vocabulary standardization. XML (Extensible Markup Language) has been used to explain the data objects in order to make the intergration easier. Specifically PToxML (XML for Predictive Toxicology) has been proposed to describe chemical information.

There is also an idea of Multidimensional integration approach where the data are materialized in a local relational database. This database treats each data source as an independent dimension, which its component schema links to global schema. Although at the moment the main problem of integration is a data inconsistency so a framework is needed in order to clarify integration process for toxicity data. Several approaches to solve inconsistency between databases have been implemented. For example one is data fusion or reconciliation of data. That is different values become just one using a fusion function (i.e. average, highest a majority), depending on the data semantics. Some guidelines have been brought forward in the following chapters, which highlights the main issues about what sort of data needs to be considered and collected for study.

2.4 Benchmark Datasets

In our work besides Demetra datasets we used number of UCI [34] datasets. This is an online repository of large datasets which encompasses a wide variety of data types, analysis tasks, and application areas. The primary role of this repository is to enable researchers in knowledge discovery and data mining to scale existing and future data analysis algorithms to very large and complex data sets.
There are number of datasets which are designed for classification tasks which have been used to test our data generation algorithm in the last chapter.

2.5 Summary and Conclusions

We described toxicity data as available from online sources. This data is sometimes unreliable and possess low quality. Its presentation is also not consistent throughout different sources and that makes the interpretability and accessibility of the information difficult. The toxicity data have different dimensions. The chemical compounds also have number of properties (descriptors), which relates to their biological activities. This relationship can be modelled by QSARs, which are used for toxicity prediction of untested chemicals. There are standard procedural steps to build QSARs. Data quality and descriptor selection plays important role in this process. Data quality and QSARs are two fundamental elements in toxicology studies. Therefore we address these issues in chapter 3.

3. DATA QUALITY ASSESSMENT

Nowadays, given the development and low cost of high data storage capacities, more experimental data is available from various scientific laboratories. A modern approach to the accessibility of large amounts of data is therefore using data integration methods. In this context, data quality is one of the most important attributes for data integration.

What is data quality assessment? Data quality assessment is the process of evaluating data to meet the specific needs of the domain users. Most important measures of data quality are accuracy, completeness, consistency, timeliness, uniqueness and validity. Data quality initiatives are generally focused on improving these metrics so that data will promote reliability of the system.

Data quality efforts tend to focus on transforming data to improve the efficiency of enterprise applications. This data might comprise number of attributes and elements.

In our project we draw our attention to the importance of this issue.

We have used toxicology domain as the field of our experimental work. In predictive toxicology, the experimental data is used for Quantitative Structure Activity Relationship modelling (QSAR) [25]. QSAR is relating aspects of chemical compound structure to biological activities against various endpoints in order to predict chemical toxicity of new compounds. Currently most of toxicity data is obtained from publicly available databases such as Toxnet [10] or DSSTox [17] as collation of various experimental data from governmental or industrial bodies. But because of their limitations such as various experimental conditions, incomplete source identification or lack of standardization requirements for different measurement units, many of them may still not be fully recognized as reliable sources. Efforts are paid to organize and manage toxicity databases toward standardization and to improve their integrity and reliability by National Institute of Standards and Technology [6] which focuses on producing a common vocabulary of weights, measures, names and symbols to scientific enterprises and agreement of a data file terminologies. This effort provides procedural guidelines for experimental work but still the inconsistencies of data values within a source or from one source to another remain a subject to be addressed. These drawbacks generated a demand of methods to tackle the data quality problems [40]. In the next sections, we overview

some integration methods implemented in other domains to address the problem of low quality data. A short analysis of each method is also provided to clarify the relevance of each approach to the predictive toxicology domain.

3.1 Data Quality Assessment Methods

Data quality and information quality are terms often used synonymously, although quality data refers to attributes such as error rate, correctness and integrity but information quality is concerned with how the information is produced and interpreted and contains dimensions like: availability, completeness and documentation [41].

A number of tools and methods have been studied and introduced in order to resolve data quality issues in information integration. Some of these methods have been analyzed here. These are as follows:

A) According to Naumann [41] quality of information depend on user, which is considered as subject, the information as object and the process of accessing the information as predicate of a query. These are factors used to qualify the information based on different criteria.

For example, for the Subject (the user) there are some information quality (IQ) criteria such as: believability, concise representation, interpretability, relevancy, reputation and understandability, which are assessed by user according to his experience, sampling and continuous assessment. Criteria for the information itself include: completeness, customer support, documentation, objectivity, price, reliability, security, timeliness and verifiability, assessed based on parsing, sampling, contract and expert input. The process criterion includes: accuracy, amount of data, availability, latency and response time, assessed based on cleansing techniques, parsing and continuous assessment.

This method identifies elements that are information system processing oriented, like security, timeliness, price and latency, pointing to implementation of time and cost effective processing. Other information quality measurements such as believability, reputation and understandability are also very user dependent assessment methods which could be varied dramatically from one user to another and therefore cause problems at the end. In our opinion, for toxicology scientists, access to reliable data with accurate information is far too important that he can easily overlook the process-

based criterion. Other issues such as completeness, concise representation and reliability seem however relevant to any domain.

B) Fusionplex [42] is a system that integrates information from multiple sources and also resolves data inconsistencies by the use of data fusion methods. A feature weight is identified and added to the database related to that source with the mean of adding few more columns with features for each source. Some examples for features are: Timestamp: the time when the source information was validated, Cost: the money spent to transmit information over the network, Information Accuracy: the level of correctness attained in measurement or how this data conforms to the standards. Availability: if information is provided at random time, Clearance: the security needed to access the information.

This system retrieves all relevant data from sources and assembles them. The intermediate product is poly-instances, which include all inconsistencies.

For this method, inconsistencies are described as schematic differences between databases. For example in one database we might have “salary” as an attribute and in one “income” which both represent the same thing. Inconsistency also refers to data representation in the form such as “currency” represents US dollar in one database and Swedish krona in another. Fusionplex measures the data quality at the integration stage with fusion methods. It measures the quality based on the general information processing criteria such as accuracy and availability rather than the data itself. Two specific inconsistencies resolved with this system apply also in predictive toxicology in the form of representation of measurements and weights. However, our main problem, the values confliction from one source to another, stays still untouched. For predictive toxicology data need to be collected and integrated from different sources, but one must address the issue of suitability of this data for integration in term of quality. Some of the source features criteria currently point to validation of the source by the user, which still can be changed between various users.

C) COLUMBA: multidimensional data integration [33] is an integrated database of protein annotations. It performs the quality assessment of data by a data cleansing method. Errors in these databases are considered as syntactic and semantic. Syntax errors are mainly domain or format violations in data entries and misspellings. Syntactic cleansing such as format, domain transformation, standardization, normalization and dictionary lookup is performed by the individual parsers.

Semantic errors are considered to be very effective in the quality of data. These errors

are resolved by using redundant information, which is possible where we have another version of the same data source. The process of choosing the data is based on schema mapping rules and comparison of the data on this schema, which contains a number of tables. Then the inconsistencies are highlighted and the reliable data is integrated into a third instance of the target schema.

The limitation of the COLUMBA system lays on relying on the redundant data. Where no instances of the data are found in the system, the quality process cannot be achieved. Still this method raises some questions such as:

How redundant data is qualified? What are the sources of this data? How it can be trusted? All these questions take us back to the initial main issue: what data quality is. Some steps of this approach might be considered relevant at data cleansing stage, such as overcoming syntax and semantic errors (misspelling, standardization) which are important issues since have direct effect on the quality of the data and need consideration in any type of database management system although value of the data through this process can not be evaluated and measured for accuracy which is very important concept in toxicology domain.

D) An information quality assessment methodology (AIMQ) has been introduced by Richard Y. Wang [43] and contains three components: Product-Service-Performance (PSP), IQ Assessment and IQ Benchmark Gap Analysis. Product-Service-Performance contains four parts: these are criteria to identify the best practice of company in production and delivery.

Table1: The PSP/IQ Model

	Conforms to specifications	Meets or exceeds consumer expectations
Product quality	Sound information IQ dimensions: Free of error Concise representation Completeness Consistent representation	Useful information IQ dimensions: Relevancy Understandability Interpretability Objectivity
Service quality	Dependable information IQ dimensions: Timeliness Security	Usable information IQ dimensions: Believability Accessibility Ease of operation Reputation

The second component measures information quality according to specified dimensions, based on questionnaires. The third component consists of the IQ Benchmark Gaps Analysis techniques and the IQ Role Gaps analysis technique. IQ Benchmark Gaps compares an organization's assessment to that of a best practice organization.

This approach is mostly based on data processing. It is also user oriented in the sense that user decides according to the questionnaire if the information and processing procedures are sufficient. The evaluated outcome will differ from one user to another. Some of the criteria such as free-of-error, concise representation, believability and relevancy still need to be based on some matrixes which would be referred to at the evaluation stage and can explain for instance what sort of information is believable which still shows dependency that wouldn't be good practice for evaluating toxicity data. The other two components are entirely assessing the organizational performance in the sense of improving their products and services based on feedback from consumers.

E) A methodology for establishing and maintaining quality in data context [44] considers five levels:

-Test of completeness and emptiness: this is done for every single table or file of data used under review.

-Ranges, Means and Distributions: at this stage every element in each table is reviewed and compared with the standards of data definition document in order to assure producing meaningful value within the standard framework.

-Derived Relationships: this analysis investigates the meaning of the data in relationships when appears for the same element in different tables.

-Meaning and Interpretation: the real meaning of data is interpreted within all the collected data.

-Hypothesis and Discovery: the result of the prior analysis is studied, the source of inconsistencies identified and relevant actions taken to overcome the weaknesses.

The approach of establishing and maintaining quality in data context within organization could be a first step forward toward a reliable toxicity database management system. It is an approach of inside out observation and cleansing. It is also a good applicable practice for any system although it has no use at the integration stage when the organization is forced to collect data from various sources. It makes the internal database free of error but has no control over incoming data from

distributed sources, since the data sources have different file formats, structure and represent different information and doesn't match the internal database schema.

F) Data quality in predictive toxicology: identification of chemical structures and calculation of chemical properties:

Helma [5] highlights some of data inefficiencies and errors in toxicology databases, also draws some rules from a case study, which was carried out in order to emphasize how some of the elements in experimental work could be assessed under quality assurance. In most toxicology databases instead of chemical structure, compounds are identified by CAS Registry number, and because of formatting or typing errors sometimes compounds cannot be identified in sources of data.

Since this data is studied for QSAR modeling it is essential to identify the properties of compound, which may have an impact to the endpoints, and consequently to the prediction of the toxic effect and the model. Some recommendations have been given for the retrieval of structures from external databases and the calculation of chemical descriptors which are based on how data should be recorded in one organization database with the emphasis on accuracy of the chemical structures and systematic problems.

What has been highlighted by Helma emphasizes the idea of data representation rules in any source of toxicity database. Some of these rules are drawn from standard agencies in place for collecting, storing and processing such data. If every organization follows the same rule, soon all the toxicology databases would have common representations, attributes and elements, which automatically increase quality and reliability of their data.

3.2 Possible Approaches for Measuring Data Quality in Predictive Toxicology

As it has been mentioned with integration process data need to be collected from various sources. This data need to be filtered in order to extract the quality data. At the moment there are lots of publicly available data sources. But there are no guidelines on how to measure the quality of these sources. These sources may conflict with each other at three different levels:

- Schema level; the sources are in different data models or have different schemas within the same data model.

- Representation level: the data in the sources is represented in different natural languages or different measurement systems.
- Data level: there are factual discrepancies among the sources in data values that describe the same object [45].

In toxicology we are more concerned about the sources at data level. High data quality has been defined as data that is fit for use by data consumers and is treated independent of the context in which data is produced and used. Data quality in general has been characterized by quality criteria or dimensions such as accuracy, completeness, consistency and timeliness. However there is no general agreement on data quality dimensions. Following is a table about these dimensions collected from different experts in this field which are good guidelines on this issue [45]. All these elements could be considered when we select the data source. We need to examine the source against this criterion and then make the selection. Some of these elements might not apply to the type of data we wish to collect but some of them like: timeliness, completeness or accuracy is very important in our case as mentioned above. Based on the information above we can look at Table 3 which identifies the ways on how some of the elements on source selection can be measured [46]. Since scientists of molecular data are the best people who can comment on this, one approach could be to choose number of users to measure quality of some of the reputable sources based on this table. Then this information could be stored in some sort of database and used when required. This could be in the form of a table with the list of sources and quality measurements above as fields in the table. We could use ranking criteria (for example between 0-1) or percentage for each field and fill the table with values we collect. Sometimes the source itself provides some information about itself for example how often the data is updated or how old the data is. In other cases users themselves could process this.

Table2: some of the highlighted dimensions for source selection

Source Specific	Ease of understanding	User grade from 1-10 based on representation of the data.
	Reputation	User grade from 1-10 based on personal preferences and professional experience.
	Reliability	Ranking from 1-10 based on accuracy of experimental method with which the data is produced.
	Timeliness	Update-frequency measured in days. For instance in toxicology the laboratory standards may change over time and also the results of one laboratory is so dependant on the environmental factors of where the laboratory is based and when the data has been collected.

The information in this table may be starting point in research for automation of a dataset development as a feasible computer science objective.

3.3 Data Cleaning Methods

Data cleaning techniques and procedures for noise removal could also enhance data quality. These methods could be applied to the data at different stages to address variety of data quality problems. At data collection stage, this could be in the form of removing duplicated records, missing values, spelling errors and outdated codes [47]. These techniques could also be used at data analysis stage. The purpose of data cleaning at this stage is to remove data errors in order to increase the quality for better classification models produced by machine learning and data mining algorithms. These techniques are based on outlier detection. Examples of some of these techniques are: cluster based, distance based and density based outlier detection. In this thesis our proposed algorithm for quality assessment addresses some of the criteria for cleaning data (section 4.3). Some common criteria are also such as: check for invalid values, out of range values, null values and missing values.

3.4 The Missing Values Problem

The subject of missing values in databases has long been studied and discussed in different domains. This is a big problem for data storage and processing. Incomplete data reduces quality and reliability of the models. It appears in different forms in databases. It is indicated for example by “0”, “N/A”, “NIL”, dashes or just empty cells. The missing values may occur for different reasons such as complexity of the computational measurement for that specific parameter (when data has to be generated by computer automatically) or, in some surveys; because of ambiguity (i.e. the interviewee finds it difficult to answer a given question). But the significance of the missing data and its effect on data mining is not always clear in final analysis. Even some well established data analysis tools assume that there is not particular importance for the missing value and that is why the system or the user may replace it or just omit it. This proves how important is to trust and understand the available data. How this problem could be overcome has always been an issue. In some domains, when the data is processed, the incomplete data is simply ignored, deleted at data cleaning stage or not considered in analysis. Various machine learning and data mining tools deal with the problem differently.

3.4.1 Methods to Overcome Missing Values Problem

A) Omit records: one very simple way is to omit the records with empty cells, which as a result reduces the sample size and has dramatic affect on the analysed data parameters and statistical characterization of the data. It is used especially in cases when high proportion of data is missing.

B) Calculation of the average (mean): in cases when there is just one attribute value missing, the average value of all the other attributes in that row is calculated and considered for the missing value. The danger of this approach is the decrease of the data characteristics variability and balancing out the values. Also when there are outliers in the datasets, in the case when the related missing values falls in that category, this will affect the mean value calculated parameter. All of these will affect the final result of the mining task especially when the procedure is based on classification strategies [48].

C) Single imputation: for cases where a big proportion of data is missing, the method of statistical computation of missing values could be applied. With this strategy missing data in each cell is estimated based on available data in another relevant cell, which satisfies certain matching criteria. A good example is missing income value estimated by comparison with existed records from another survey for a person living in the same area and having the same educational background and age. The disadvantage of this method is that artificial data, which replace the missing values, is exactly the same as existing data. This results in reduced data values variation characteristics and increases specific class or cluster of information that in turn affect the quality of data mining analysis results [49].

D) Multiple imputations: is another method, which instead of filling empty cells with specific values, it replaces them with a set of applicable values. The multiply imputed datasets are then analysed and results are combined [50].

E) Expectation Maximization (EM) algorithm: it firstly uses imputation for missing values and then re-estimates the missing data values and iterates until convergence. The method of iteration has been widely applied to missing data problems [50].

3.5 Imbalanced Datasets

In data mining, classification learning is a supervised learning scheme that uses knowledge gained through the training process of classified instances for

classification of unseen examples. One of the main issues for classifier during this process is the samples distribution of classes or class balance. Imbalanced or skewed dataset [51], affect the performance of classification algorithms. The over represented classes provide enough information for training the classifier because of their sufficient number of samples against the under represented class. Real world scientific applications often face this problem for a number of reasons [52].

Various approaches and methods have been proposed to tackle imbalanced data problem. One of these methods is one-sided selection [53] in which the border-line/negative examples or the ones overlapping in two class dimensional space are removed.

Another method is DataBoost-IM approach [54]. According to this method the hard examples from minority and majority class are identified. Then the synthetic samples are generated using the hard samples and added to the original dataset. The class distribution and the total weights of the different classes in the new training set are re-balanced at the last stage.

Guided re-sampling technique [55] is another solution which first determines the subcomponents within each class. The element in each subcomponent is re-sampled until each subcomponent has the same number of examples as biggest subcomponent. Then the between-class imbalance is eliminated by randomly selecting and duplicating members of the minority class.

SMOTEBoost [56] is another method which increases the learner performance in classification of minority class with creating synthetic instances by operating in the feature space rather than data space. Using this method a new minority class sample is created in the neighbourhood of the minority class target.

There are also some methods which down-size the majority class in order to equalize the distribution of two classes [57] [58]. All these methods concentrate on the two-class problem with minority and majority class.

3.6 Summary and Conclusions

Data quality is an important issue in scientific domains. There are some approaches to overcome the problem. Naumman [41] introduces an information quality framework based on the user, information and the process of accessing this information. Another example is Fusionplex [42], which is a system that integrates information from

multiple sources and also resolves data inconsistencies by use of fusion methods. COLUMBA [33] is another system that performs the quality check by data cleansing procedures. The Information Quality Assessment Methodology introduced by Richard Y. Wang [43], overcome quality issue by defining number of criteria in its components. The methodology for establishing and maintaining quality in data context [44] is another strategy, which assess the data at different levels. Helma [5] also introduces some method for measuring quality in predictive toxicology. Data cleaning is also used to enhance the quality of the data. There is also an issue of missing values in datasets, which reduces the quality and reliability of the data. There are some methods in use to overcome this problem. Some of these methods are such as: omit records, calculate average, single imputation, multiple imputations and expectation maximisation.

Another issue is modelling this data which raises the problem of how subcomponents of this data have been structured that identify the data as balanced or imbalanced. The imbalanced data affect the performance of classifiers during the classification process. Number of methods has been proposed to overcome imbalanced data problem such as; re-sampling [55], DataBoost-IM [54] and SMOTEBoost [56].

In this chapter we identified the main weaknesses for data characterization and impact in model development. In chapter eight we propose an algorithm in order to overcome the problem and improve model performance.

4. A STUDY ON DATA QUALITY IN TOXICOLOGY AND NEW ALGORITHM FOR DATA QUALITY ASSESSMENT PROCESS

Measuring the quality of available information is an important issue in many scientific domains, and even more so if provides a basis for further model development. An example is predictive toxicology that relies on data from public and commercial databases for analysis and modelling towards Quantitative Structure Activity Relationship (QSAR) models development. Much work has been done on QSAR modelling, but in many cases little attention has been paid to the quality of toxicology data, since there is not a clear definition of what quality is or criteria to base the quality assessment on. This chapter presents the result of investigation into online databases and the contradiction in the data presentation and values and also number of quality issues from experimental work on toxicological data (DEMETRA datasets) and also proposes some quality criteria as a result of the study.

4.1 Inconsistencies in Online Databases Data Presentation and Values

In this study we compare values for same attributes of chemical compounds presented for same carcinogenicity tests by three different sources, also study the values presented for aquatic studies by two other databases. The following online databases have been studied in order to find the deficiencies in data values and representation:

-IRIS (Integrated Risk Information System): this data base provides toxicology data in support of human health risk assessment. It is compiled by the U.S Environmental Protection Agency and contains 543 chemical records on values on cancer and non-cancer health effects that may result from lifetime oral or inhalation exposure to specific chemical compound in the environment.

-ITER (International Toxicity Estimates for Risk): this database contains data in support of human health risk assessment. It is compiled by Toxicology Excellence for Risk Assessment and contains 617 chemical records with key data from different sources. It provides data related to non-cancer oral, cancer oral, non-cancer inhalation and cancer inhalation.

-CCRIS (Chemical Carcinogenesis Research Information System): it is a scientifically evaluated and referenced data bank, developed and maintained by the National Cancer Institute. It contains 8937 chemical records. Data is represented for

carcinogenicity studies, tumor promotion studies, mutagenicity and tumor inhibition studies.

These databases can be accessed via Toxnet web site. [10]

For information relate to aquatic species testing, following databases have been chosen:

-ECOTOX: The ECOTOXicology database (ECOTOX) is a source for locating single chemical toxicity data for aquatic life, terrestrial plants and wildlife. ECOTOX was created and is maintained by the U.S.EPA, Office of Research and Development (ORD), and the National Health and Environmental Effects Research Laboratory's (NHEERL's) Mid-Continent Ecology Division. This site is accessible via: http://www.epa.gov/ecotox[17]

-USGS: The Columbia Environmental Research Centre provides leadership and scientific information for the U.S. Geological Survey by addressing national and international environmental contaminant issues, and assessing effects of habitat alterations on aquatic and terrestrial ecosystems. This includes large-river floodplains, coastal habitats, wetlands, and lakes. This site is accessible via: http://www.cerc.usgs.gov [59]

For this work following issues have been considered:

a) We selected common chemical compounds in first three databases (IRIS, ITER and CCRIS) relate to carcinogenicity and non- carcinogenicity studies. The dose used and the time of exposure which logically needed to be in the same measurements. The type of exposure also needed to be the same. For this purpose following chemical compounds were chosen to study: PHENOL with CASRN: 108-95-2.

These databases represent the data for testing on various types of rats and mouse.

b) In the last two aquatic databases (ECOTOX and USGS), we have selected information provided for testing on: PENTACHLOROPHENOL, CASRN: 87865 tested on DAPHNIA MAGNA.

c) We compared the values and recorded them in the table.

The results of our study are listed below:

Results of PENTACHLOROPHENOL, CASRN: 87865 tested on DAPHNIA MAGNA from two aquatic databases are as follows:

Table4: Results from ECOTOX online database

Scientific name	End point	Effect	Effect Measurement	Trend Effect%	Duration/Exp Type	Conc(ug/L)
Daphnia magna	LC50	MOR	MORT	---------	48 h	F55
Daphnia magna	LC50	MOR	MORT	INC ----------	48 h	A 1230,1120-1340
Daphnia magna	LC50	MOR	MORT	----------	48 h	F 320
Daphnia magna	NR-ZERO	MOR	MORT	NEF -------- 0	48 h	A`400

Table5: Results from USGS online database

CHEMICAL	PERCENT	DESCRIP	SPECIES	SIZE
PENTACHLOROPHENOL	86	TECHNICAL MATERIAL	DAPHNIA MAGNA	1ST INSTAR
PENTACHLOROPHENOL	96	TECHNICAL MATERIAL	DAPHNIA MAGNA	1ST INSTAR
Chemical Name:	**PENTACHLOROPHENOL**			
Common Use:	HERBICIDE			
Measurement Units	UG			
CAS Number:	87-86-5			

HARDNESS	TYPE	TEST_UNT	TOX_UNT	LC50_24H
40	STATIC	EC	UG	
40	STATIC	EC	UG	

FROM_24H	TO_24H	LC50_48H	FROM_48H	TO_48H
		410	319	527
		240	138	307

DIET	TEMP	PH
	17	7.4
	17	7.4

Results for PHENOL with CASRN: 108-95-2 on rats from IRIS, ITER and CCRIS for oral Carcinogenicity Studies is as follows:

Table6: Results from IRIS online database for Carcinogenicity studies on rats

	IRIS
Species	F344 rats (male)
Dose	0, 2500, or 5000 ppm-0, 260, and 585 mg/kg-day
Effect	Decrease in body weight, decrease in water consumption
Period	103 weeks
Route	oral
Study	National Toxicology Program

Additional information:

NCI (1980): dose related decreases in body weight as compared with the controls were observed in male by 15% in high dose. Water consumption was reduced by approximately 10% at the high dose. In aged rat assessment found statistically significant increases in chronic kidney inflammation in high-dose (5000 ppm). There were no significant changes at low dose.

There were no dose-related trends in cancer incidence in male rats but the study reported several tumours for which statistically significant increases were seen in low-dose males only.

Table7: Results from ITER online database for Carcinogenicity studies on rats

	ITER
Species	F344 rats (male)
Dose	0, 2500, or 5000 ppm-0, 260, and 585 mg/kg-day
Effect	Decrease in body weight
Period	2 years
Route	oral
Study	National Toxicology Program (1980)

Additional information:

ATSDR: Health Canada, RIVM and U.S.EPA have evaluated the carcinogenicity data for phenol. Under the current EPA guidelines phenol would be characterised as group D, not classifiable as to human carcinogenicity. Health Canada did not assign a specific classification but indicated that available data support the likelihood that phenol is at most weakly carcinogenic. RIVM noted that available data in animals suggest that phenol can act as a tumour promoter.

Table8: Results from CCRIS online database for Carcinogenicity studies on rats

	CCRIS
Species	F344 rats (male)
Dose	0; 2500; 5000 ppm
Effect	Negative
Period	Not specified
Route	oral
Study	National Toxicology Program, National Cancer Institute

Analysis: the results show that there are inconsistencies in data presentation. In IRIS, ITER and CCRIS data is represented in three different ways.

IRIS presents the data in long document explanatory form. Although seem to be supported by the same laboratories as ITER and also information are much detailed compare to other two databases.

ITER presents the data in the form of summary table from different laboratories. For some of those laboratories data doesn't exist.

In CCRIS data is represented for each individual species in very short form. Some information is missing like the duration of the exposure of the compounds.

It is very difficult to go through pages of information and select the numerical figures which are a big draw back in presenting information.

The data in all three databases is in textual form which is very difficult to analyze. Non numerical data presentation can not be used for experimental modeling.
Although it seems that data in IRIS and ITER are represented by same laboratories still information are not the same. (as noted in additional information section)
Data presented in ECOTOX and USGS have different form, used different attributes to explain the measurements (some are similar) (tables 4 and 5). The values are different as expected in both databases. Although the data in ECOTOX doesn't explain the test condition like: PH and Temperature which would have effect on result so the comparison can not be very consistent since the test condition is not clear in ECOTOX.

4.2 Detailed Investigations and Experiments on DEMETRA Data

The contribution of our investigations at this stage is to highlight some usual problems of data quality used for toxicity prediction. Our current objective is the study of inconsistencies in data values and their effect on downstream QSAR modelling.
Given the current facilities available for complex calculations, it seems that high confidence is implicitly awarded to data downloaded from online resources. The same applies to data generated by specialist software. We used the opportunity to study the DEMETRA data sets on some issues on data quality for large databases. We started with identification of descriptors sharing the same name and duplicated as generated by various software used by research laboratories involved in the project. We addressed the differences in data source values and also differences in performance of models developed from the same data sources. Data on five toxicity endpoints are provided by the DEMETRA project [60] for four different species: Bee, Daphnia, Trout, OralQuail and DietaryQuail. For each dataset, values for six compound descriptors calculated by two specialist programs: ACD [61] and Pallas [62], have been considered. These programs calculate pKa, logP, logD values and also metabolites based on structural formulae of compounds. In the field of industrial pharmacy perhaps the most important physicochemical characteristics of compounds are their acidity or basicity (expressed by their pKa value), hydrophobicity and its dependence on pH (expressed by their logP and logD respectively) [62]. Calculating accurate values of pKa, logP, logD and other chemical descriptors requires a great deal of work and use of specialized software.

For this work the number of chemical compounds presented in each data set varies from 105 for Bee endpoint to 252 for Trout. Our aim was to highlight the variation of values for each descriptor produced from one program to another and also to compare any further quantitative differences between specific descriptors calculated by one program with the value for the same descriptor and chemical compound generated by the other one. Then we compare the accuracy of basic classification models built using input data presented for each endpoint by descriptors calculated by ACD and Pallas. Ten tables were investigated, two for each endpoint. The aim of this experiment was to identify how the predictive models' quality is affected by hidden parameters such as source of data, subjective input characterization in running feature extraction algorithms etc. Comparisons of models performance will address variations, contradictions, reliability and deficiency issues.

4.2.1 Data Pre-Processing

Follow up experiments are related to value comparisons for different endpoint presented by software (Pallas, ACD). There are number of Excel files presented by Demetra project explaining the descriptors values for same chemical compounds tested on different species, which we were going to work with. They are endpoints for: Bee, Trout, Daphnia, DietryQuail and OralQuail. For each endpoint we have number of files which have been calculated by different programs but we are only concern about the values which are produced by Pallas and ACD.

This experiment is divided into number of subsections, each section apply to each endpoint.

For all endpoints we consider the same number of chemicals with same descriptors. Since there are some empty cells in each file we will clean the data before training any model.

Following tasks have been carried out on each endpoint:

- Data cleaning: all the uncompleted rows (or with zero values) have been deleted.
- Same chemicals have been selected and inserted in new file.
- Same descriptors have been selected plus ID attribute (field) and saved in a file with added SD extension in their name which stands for selected descriptors. (ex: B_2DPALLAS_v2_SD.xls).
- After the selection we have two files for each endpoint to start with. One file produced from Pallas and one from ACD. For Bee endpoint we have 95 selected

compounds which are common between both files. For Trout we have 262 selected chemical compounds. For Daphnia we have 244 compounds. For OralQuail we have 104 chemical compounds and for DietaryQuail we have 107 chemical compounds.

Following file: "data (5DS+ClassEC)_v2.xls" (from Demetra) also categorizes the chemicals into classes which should be used for modeling. A column has been added in each file to contain the class category of chemical compound.

- There are number of pdf files presented by Demetra project which specify the criterion for selecting test and training data sets for each endpoint which are as follows: TestData_Trout.pdf, TestData_Daphnia.pdf, TestData_OralQuail.pdf, ToxicityTestData_DietaryQuail.pdf, TestData_Bee.pdf
- We used Weka [63] and SPSS [64] for modeling. In Weka number of supervised learning algorithm have been selected and used for modeling.
- For training models on Weka, each file has been converted in Weka format and saved as arff file.
- In SPSS, Linear Regression has been chosen for modeling.

4.2.1.1 Procedure

The following files have been under consideration: (for Bee endpoint): B_2DPALLAS_v2.xls and Bee1_1v1_ACD2D_v1.xls, (for Daphnia): D1.1v1_ACD2Dv1.xls and D_2DPallas_v2.xls, (for Trout): T3.1v2_2DPallas_v1.xls and T3.1v2_2DACDv1.xls, (for OralQuail): OQ1_1v1_ACD2Dv1.xls and OQ_2DPallas_v1.xls and (for DietryQuail): DQ_2DPallas_SD.xls, DQ1_1v1_ACD2D_v1.xls. Common descriptors from both files were selected.

- Selected descriptors from first file calculated by Pallas are:

LogP(Pallas), LogDpH3 (Pallas), LogDpH5 (Pallas), LogDpH7 (Pallas), LogDpH7.4 (Pallas), LogDpH9 (Pallas).

- Following descriptors are selected from second file calculated by ACD:

LogP (ACD), LogDpH3 (ACD), LogDpH5 (ACD), LogDpH7 (ACD), LogDpH7.4 (ACD), LogDpH9 (ACD).

- From all the files, number of chemical compounds has been deleted in order to make sure both files have same number of chemicals. There were number of chemicals in each file which had missing values for some of the descriptors, those have been deleted originally.

- The files have been converted to CSV format and then into arff to be compatible with Weka data mining tool.
- Following guidelines provided from DEMETRA (mentioned above) test data and training data has been specified (85% training, 15% testing). Chemicals which their IDs was specified by document for each endpoint have been selected as testing data and saved in a file with added "testing" extension. Same thing was performed for training data.
- The training set and testing set files have also been converted in arff format for modeling in Weka.
- For Daphnia following chemicals with ID: 59, 92, 409 were not added to testing set because of missing values, they were deleted at initial stage.
- For Oral Quail the chemicals with following IDs: 48, 95, 51, 103, 125, 139, 282, 346 have different CAS number (in DEMETRA classification file and toxicity file).
- For Trout endpoint the same problem exists for chemicals with following IDs: 51, 92, 123, 125, 139, 140, 171, 228, 230, 282, 305, 324, 332, 345, and 346.
- For some of the chemicals the value for all descriptors are same (ex: Trout: IDs: 268,270,279).
- For Trout endpoint, legend for Pallas is different from other endpoints, although the descriptors were selected accordingly (ex: Pallas04=LogDpH7 but in other files Pallas05=LogDpH7).
- At the initial stage when classes from DEMETRA files assigned to chemical compounds in Bee1_1v1_ACD2D_v1.xls and B_2DPALLAS_v2.xls, an inconsistency was discovered. In file produced by ACD, chemical compound with ID=450=Allethrin has been given CAS no: 584-79-2, in file produced by Pallas, ID=450=28434-00-6=s-bioallethrin which in Toxnet this is different name for same chemical which have the same CAS: 284-79-2. Also for curiosity files produced by Dragon (another program for chemicals descriptors calculation) [65] for the same endpoint was checked. In Dragon the CAS number for compound "Diquet" is: 828-00-2 but in DEMETRA file for the same compound the CAS number is: 85-00-7. The result is shown in table 8.

Table8: Results from CCRIS online database for Carcinogenicity studies on rats

Demetra	5DS+ClassEC)_v2.xls: 452=Diquet=85-00-7	ACD=450=Allethrin=585-79-2 Pallas=450=s-bioallethrin=28434-00-6
Dragon	452=Diquet=828-00-2	450=s-bioallethrin=2764-72-9
Toxnet	Diquet=2764-72-9 & 85-00-7	s-bioallethrin=Allethrin=584-79-2

4.2.2 Comparison of Global Parameters and Source Value Difference

Considering there are two files (specified as training) one presents the values for chemical compound calculated by ACD and the other one by Pallas.

1) Using Excel functions we calculated the parameters Mean, Min, Max and STDEVP for each column (LogP, LogDpH3, LogDpH5, LogDpH7, LogDpH7.4, LogDpH9) in these two files separately. At the second stage we compared each descriptors value for these parameters with its corresponding from the other file (ex: values for parameters from LogP for ACD were compared for same parameters for LogP for Pallas). Lastly the results were presented as graph to visualize the differences.

2) For second experiment, using Excel two extra columns were added to each original column in files calculated by Pallas. Data from file calculated by ACD was copied to first file for comparison purposes. One added column in Pallas present the values for the same descriptors which have been calculated by ACD and second column present the subtraction of two values (Values presented by Pallas minus values presented by ACD for the same chemical compound). Then at the next stage the parameters Mean, Min, Max and STDEVP were calculated for value difference. A graph was used to visualize these parameters.

3) Modelling: Files prepared for training and testing were used for modelling in Weka using following algorithms:

BayesNet (Bayes): is a probabilistic graphical model that represents a set of variables and their probabilistic independencies.

Multilayerperceptron (Function): a network composed of more than one layer of neurons, with some or all of the outputs of each layer connected to one or more of the inputs of another layer.

IBK (Lazy): K-nearest neighbour classifier which can select appropriate value of K based on cross-validation. Can also do distance weighting.

ClassificationViaRegression (Meta): it uses regression learner to learn a model to predict each binary target.

ZeroR (rules): algorithm for building and using a 0-R classifier which predicts the mean for a numeric class or the mode for a nominal class.

LMT (Tree): algorithm for logistic model tree structure.

J48 (Tree): algorithm for generating a pruned or unpruned C4.5 decision tree.

JRip (Rules): this algorithm implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER).

For modelling first a training data (from ACD and Pallas for same endpoint) was tested against test set using these algorithms and then for second time the training set was modelled with 10-fold Cross Validation testing method using same algorithms. The accuracy of each model(one from modelling against testing set and one from modelling with Cross Validation testing method) was recorded in separate table to compare which model suits which endpoint. Other parameters from modelling can also be recorded.

4.2.3 Results of Global Value Comparison

The result of the experiments for each endpoint has been recorded in the form of tables of values and also by graph for visualization. The result of this comparison is shown by separate graph for each column of the table (attribute/descriptor). The summary of the result has also been recorded in one single table for all the descriptors at the end of the section. These summary results have been shown graphically by two different graphs. There is also an analysis of the results for each endpoint which clarifies the findings. For simplicity in the analysis section we just refer to the summary table and summary graphs.

4.2.3.1 Bee Endpoint

As first experiment four following parameters have been calculated for each column in each file (in ACD and Pallas): Minimum, Maximum, Average (mean), Standard Deviation (for population) using Excel functions and recorded. The result for each descriptor was separated in different table to show the difference of these parameters for descriptors which values calculated with ACD and Pallas. The result was showed on charts to visualize the comparison. Table 9 shows the recorded parameters. For example second column in the table headed “PALLAS001” records the values for Mean, Max, Min and STDEVP calculated for the first descriptor (attribute) for all the chemical compounds in files produced by PALLAS program for Bee endpoint and so on. As it shown for instance in top table the minimum value for the first attribute PALLAS001 in PALLAS is -0.952306 for a chemical compound and in the ACD file the lowest value for the first attribute ACD001 in ACD file is -1.4202 and so on for

other parameters. These tables highlights the statistical differences between two files produced for same chemical compounds with same attributes (descriptors).

Table9: Results of the experiment for calculation of values Mean, Min, Max, and STDEVP for Bee endpoint

B_2DPALLAS_v2	PALLAS001	PALLAS002	PALLAS003	PALLAS005	PALLAS006	PALLAS007
Mean	3.121244223	2.748138773	2.702774168	2.5439506	2.509456989	2.390566365
Max	8.16996	8.16996	8.16996	8.16996	8.16996	8.16996
Min	-0.952306	-3.78509	-4.75384	-5.6052	-5.99314	-7.5105
STDEVP	1.978863877	2.293025338	2.467585066	2.748947509	2.807788628	3.016086395

Bee1_1v1_ACD2D_v1	ACD001	ACD002	ACD003	ACD004	ACD005	ACD006
Mean	3.239792632	2.935269474	2.817703158	2.686745263	2.675383158	2.640947368
Max	8.2665	8.1404	8.1404	8.1404	8.1404	8.1678
Min	-1.4202	-3.4969	-4.2068	-5.0751	-5.2504	-5.8599
STDEVP	2.060131802	2.306282759	2.462253314	2.618868261	2.649345297	2.763805686

As it has been mentioned previously the six attributes which are shown in table nine as PALLAS001, PALLAS002, PALLAS003, PALLAS005, PALLAS006, PALLAS007 and also in ACD file are ACD001, ACD002, ACD003, ACD004, ACD005, ACD006 are LogP, LogDpH3, LogDpH5, LogDpH7, LogDpH7.4, LogDpH9. For clarity and to distinguish between two files PALLAS and ACD we replace the names with PALLAS and ACD followed by numbers for the descriptors.

But in the following graphs for each endpoint we have shown the comparison between each parameter (max, mean, min, stdevp) calculated for each descriptor (column) by ACD and PALLAS. Following graph is the representation of the calculated parameters for the descriptor (LogP) for all chemical compounds which shows the value for both Pallas and ACD.

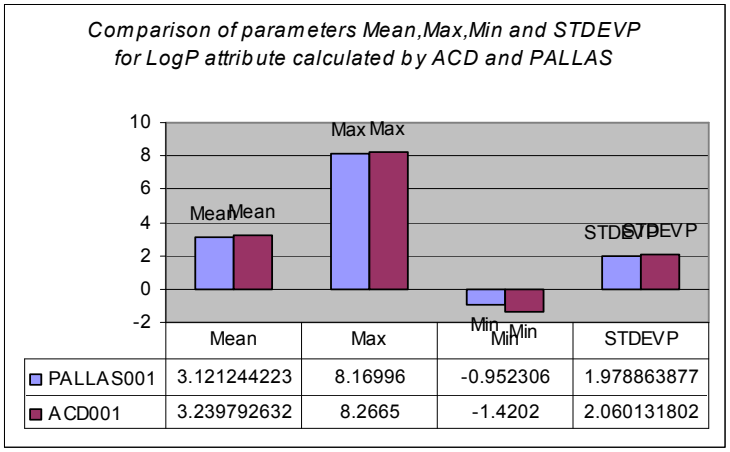
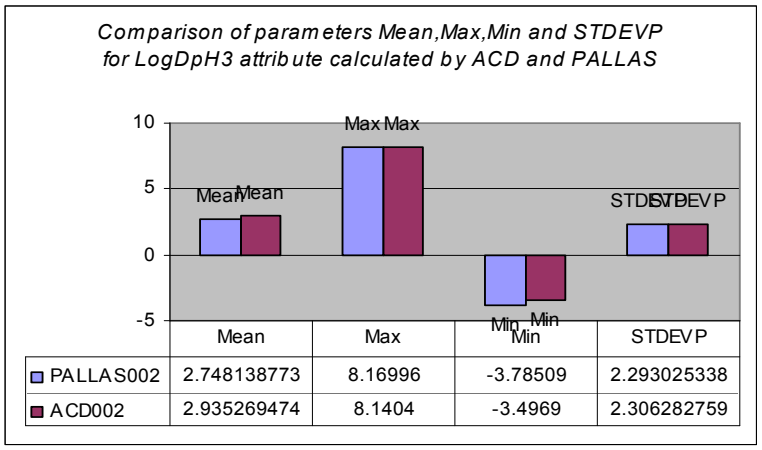


Figure2: the result of the mean, max, min and STDEVP for LogP attribute for all chemical compounds for Bee endpoint

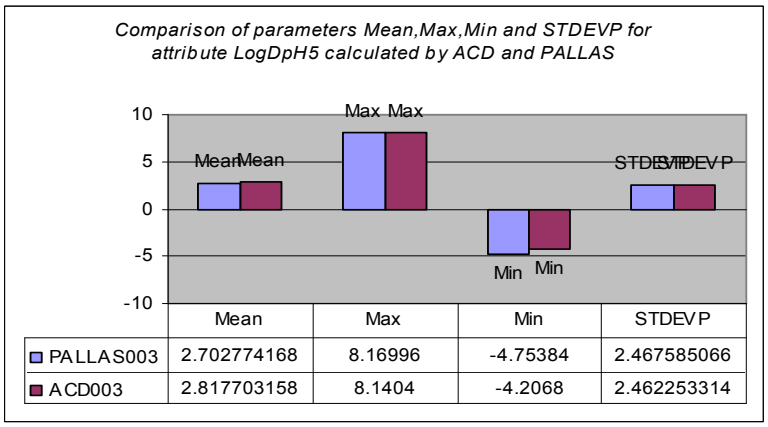
Following graph is the result of the calculated parameters for descriptor (LogDpH3) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS002	2.748138773	8.16996	-3.78509	2.293025338
ACD002	2.935269474	8.1404	-3.4969	2.306282759

Figure3: the result of the mean, max, min and STDEVP for LogDpH3 attribute for all chemical compounds for Bee endpoint

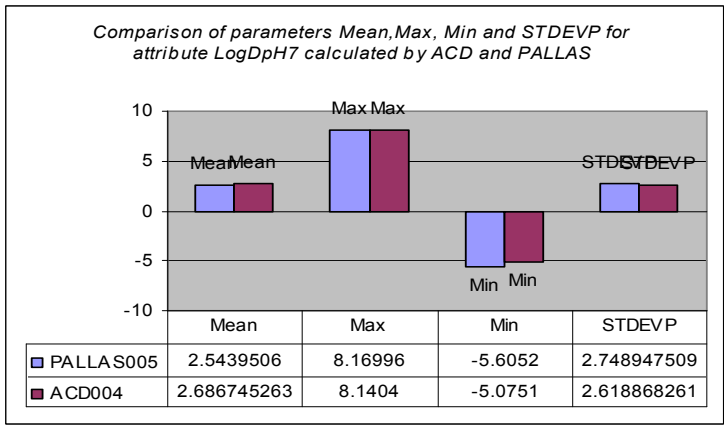
Following graph is the result of the calculated parameters for descriptor (LogDpH5) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS003	2.702774168	8.16996	-4.75384	2.467585066
ACD003	2.817703158	8.1404	-4.2068	2.462253314

Figure4: the result of the mean, max, min and STDEVP for LogDpH5 attribute for all chemical compounds for Bee endpoint

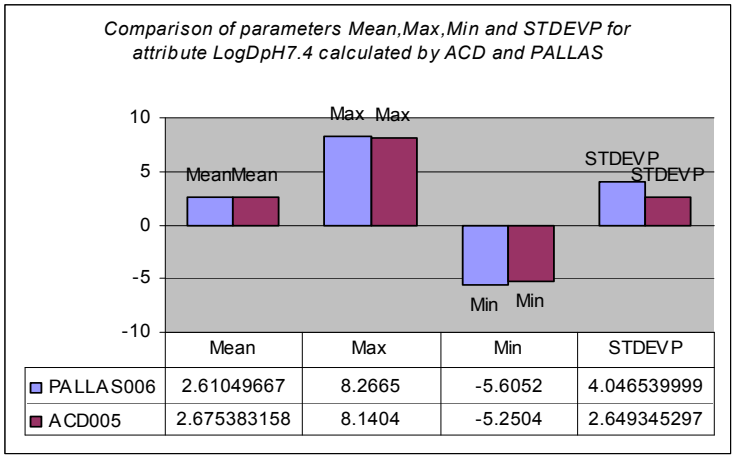
Following is the result of the calculated parameters for descriptor (LogDpH7) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS005	2.5439506	8.16996	-5.6052	2.748947509
ACD004	2.686745263	8.1404	-5.0751	2.618868261

Figure5: the result of the mean, max, min and STDEVP for LogDpH7 attribute for all chemical compounds for Bee endpoint

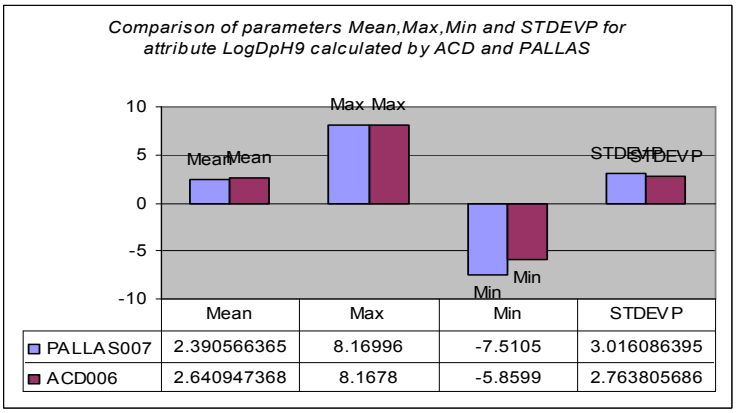
Following is the result of the calculated parameters for descriptor (LogDpH7.4) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS006	2.61049667	8.2665	-5.6052	4.046539999
ACD005	2.675383158	8.1404	-5.2504	2.649345297

Figure6: the result of the mean, max, min and STDEVP for LogDpH7.4 attribute for all chemical compounds for Bee endpoint

Following is the result of the calculated parameters for descriptor (LogDpH9) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS007	2.390566365	8.16996	-7.5105	3.016086395
ACD006	2.640947368	8.1678	-5.8599	2.763805686

Figure7: the result of the mean, max, min and STDEVP for LogDpH9 attribute for all chemical compounds for Bee endpoint

As it has been shown in previous graphs for this endpoint there are value differences for each calculated parameter (min, max, mean, stdevp) between two files(ACD, PALLAS). In the second experiment we have calculated these differences and recorded in the table and also visualise by graphs. Following shows the results of this experiment.

For instance if for “Mean” parameter for LogP in PALLAS file for Bee endpoint we have value: 2.390566365 and for same parameter in ACD file we have 2.640947368, the difference between these two values would be: -0.250381004 when the value produced by ACD is deducted from value presented by PALLAS. In following tables

we calculated these differences to clarify the problem. We have also shown the results by two graphs for better visualisation. Both graphs produce the same results.

Table10: Results of the experiment for calculation of value difference between each descriptor presented by ACD and Pallas for Bee endpoint

Pallas value minus ACD value	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Mean	-0.118548408	-0.187130701	-0.114928989	-0.142794663	-0.165926168	-0.250381004
Max	2.21882	2.22028	2.21881	3.58602	3.80667	3.81636
Min	-3.00158	-3.00158	-3.00158	-3.00158	-3.00158	-3.27919
STDEVP	0.848747337	0.88311137	0.867948411	0.988557764	1.021915587	1.13706038

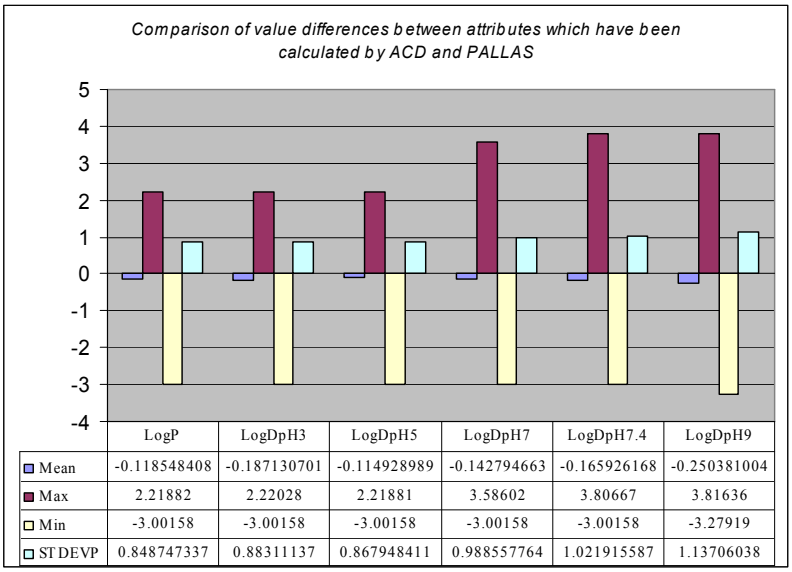


Figure8: the result of the value difference between attributes calculated by ACD and PALLAS for Bee endpoint

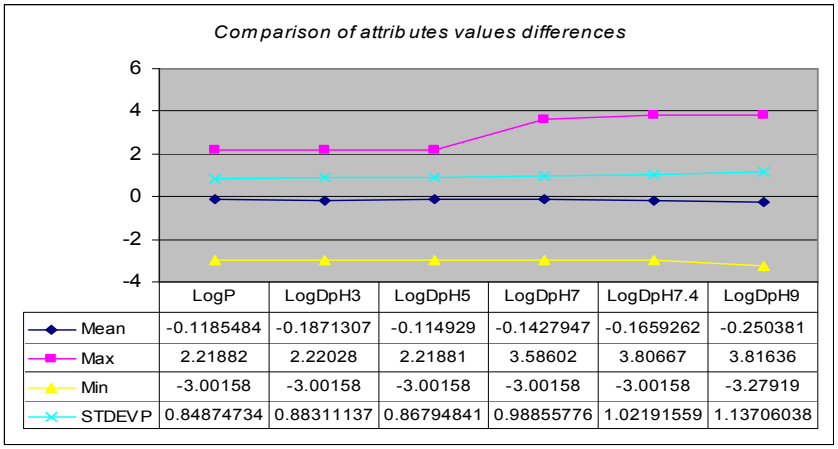


Figure9: the result of the attribute values difference for Bee endpoint

For comparison purposes and also for clarity the previous results of all the values of max and min for descriptors and also max and min descriptors value difference put in

one table.The table also shows the ID for descriptors which hold the minimum and maximum value for that descriptor.

Table11: Results of the experiment for calculation of value difference, Min, Max, and also ID of the chemicals presented by ACD and Pallas for Bee endpoint

Pallas	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	-0.952306	-3.78509	-4.75384	-5.6052	-5.99314	-7.5105
Max	8.16996	8.16996	8.16996	8.16996	8.16996	8.16996
Min Value Difference	-3.00158	-3.00158	-3.00158	-3.00158	-3.00158	-3.27919
Max Value Difference	2.21882	2.22028	2.21881	3.58602	3.80667	3.81636
ID of Min	192	382	457	373	373	373
ID of Max	146	146	146	146	146	146
ACD	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	-1.4202	-3.4969	-4.2068	-5.0751	-5.2504	-5.8599
Max	8.2665	8.1404	8.1404	8.1404	8.1404	8.1678
Min Value Difference	-3.00158	-3.00158	-3.00158	-3.00158	-3.00158	-3.27919
Max Value Difference	2.21882	2.22028	2.21881	3.58602	3.80667	3.81636
ID of Min	373	382	382	373	373	373
ID of Max	146	248	248	248	248	146

Analysis: as it shown in table 11**,** the minimum value difference for LogP ranges from -3.00158 to 2.21882 which show inconsistencies for value calculation from one program to another. The graphical results are also shown in figures8 and 9. For logDpH3 and LogDpH5 the difference is in the same range and suddenly the graph has big picks on maximum value difference for LogDpH7 and LogDpH7.4 and LogDpH9 up to 3.8163. The minimum value difference still shows the gap between calculated values which is almost same for all descriptors and ranges from -3.0015 to -3.279 (LogDpH9) which shows a pick at the end for the last descriptor. The ID for chemical compound which hold the minimum value and maximum value for each decriptor is also different in some cases. For instance the ID for chemical compound which has the minimum value for LogP descriptor is 192 in files produced by PALLAS program and is 373 in files produced by ACD. These two IDs belong to two different compounds.

4.2.3.2 Daphnia Endpoint

As first experiment four following parameters have been calculated for each column in each original file (in ACD and Pallas): Minimum, Maximum, Average (mean), Standard Deviation (for population) using Excel functions and recorded in the bottom of the each column. The result for each descriptor was separated in different table to show the difference of these parameters for descriptors which values calculated with

ACD and Pallas. The results are shown on charts to visualize the comparison. Following are the results achieved from this experiment.

Table12: Results of the experiment for calculation of values Mean, Min, Max, and STDEVP for Daphnia presented by ACD and Pallas

D_2DPALLAS_v2_SD	PALLAS001	PALLAS002	PALLAS003	PALLAS005	PALLAS006	PALLAS007
Mean	3.111978528	2.676990389	2.66961504	2.46706606	2.42831149	2.30689498
Max	11.6915	11.6915	11.6915	11.6915	11.6915	11.6915
Min	-2.701	-6.54063	-6.54052	-7.85046	-7.89228	-9.10505
STDEVP	2.200465279	2.599263541	2.65732754	2.82991807	2.86838325	3.02949894
D1_1v1_ACD2Dv1_SD	ACD001	ACD002	ACD003	ACD004	ACD005	ACD006
Mean	3.270184836	2.966811066	2.90584713	2.7529582	2.72891475	2.65871311
Max	13.676	13.676	13.676	13.676	13.676	13.676
Min	-2.3559	-5.4966	-5.8715	-6.499	-6.6644	-6.8685
STDEVP	2.254639469	2.484495301	2.55814327	2.65603901	2.67889429	2.74335268

Following is the result of the calculated parameters for the descriptor (LogP) for all chemical compounds which shows the value for both Pallas and ACD.

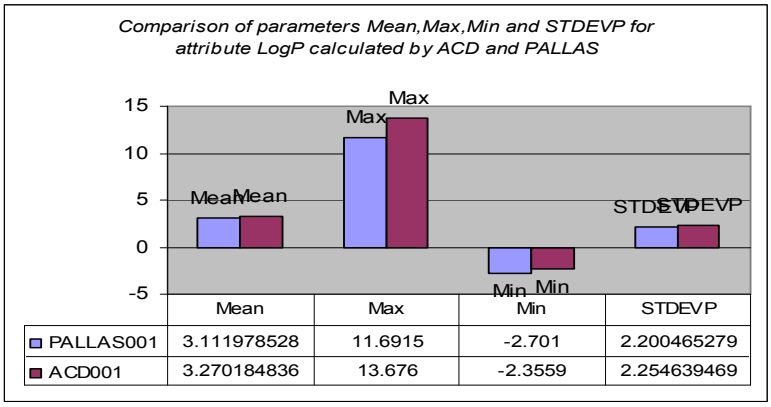


Figure10: comparison of parameters Mean, Max, Min and STDEVP for LogP for all chemicals for Daphnia endpoint

Following is the result of the calculated parameters for the descriptor (LogDpH3) for all chemical compounds which shows the value for both Pallas and ACD.

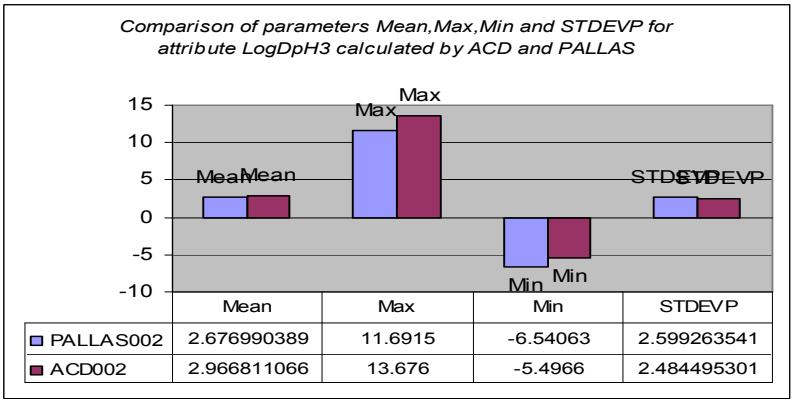
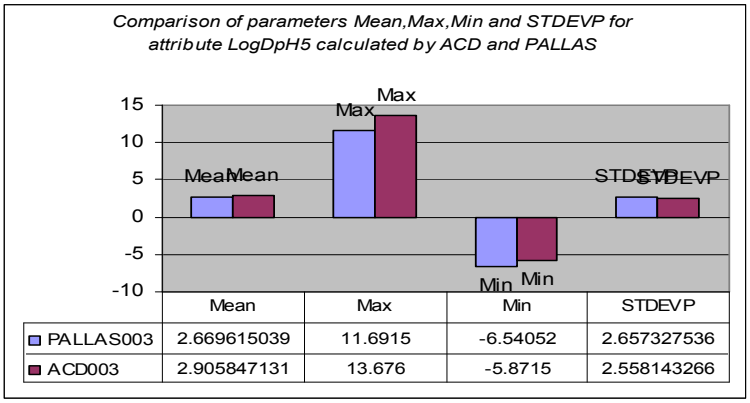


Figure11: comparison of parameters Mean, Max, Min and STDEVP for LogDpH3 for all chemicals for Daphnia endpoint

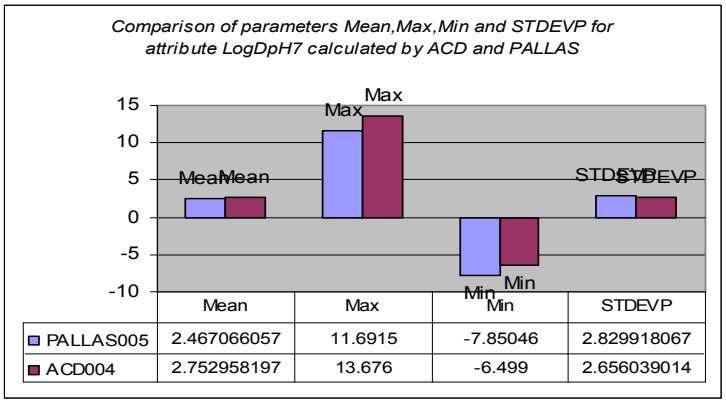
Following is the result of the calculated parameters for the descriptor (LogDpH5) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS003	2.669615039	11.6915	-6.54052	2.657327536
ACD003	2.905847131	13.676	-5.8715	2.558143266

Figure12: comparison of parameters Mean, Max, Min and STDEVP for LogDpH5 for all chemicals for Daphnia endpoint

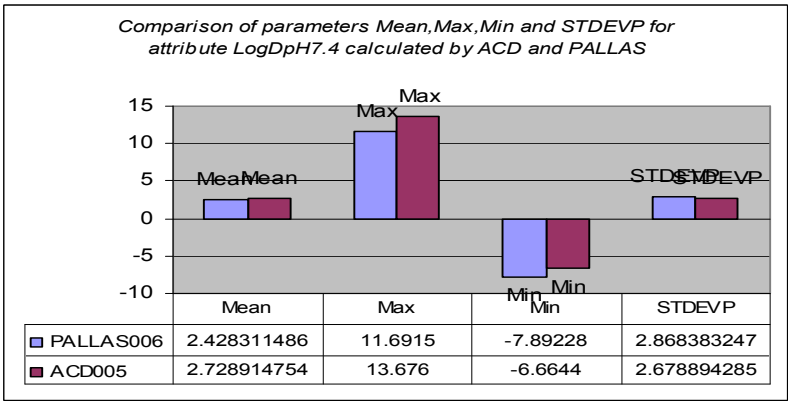
Following is the result of the calculated parameters for the descriptor (LogDpH7) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS005	2.467066057	11.6915	-7.85046	2.829918067
ACD004	2.752958197	13.676	-6.499	2.656039014

Figure13: comparison of parameters Mean, Max, Min and STDEVP for LogDpH7 for all chemicals for Daphnia endpoint

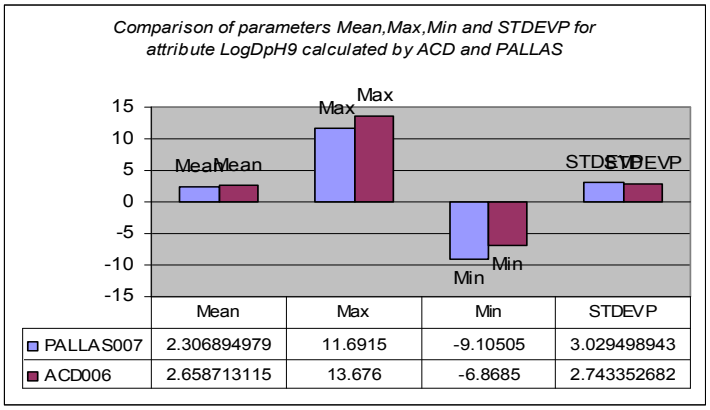
Following is the result of the calculated parameters for the descriptor (LogDpH7.4) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS006	2.428311486	11.6915	-7.89228	2.868383247
ACD005	2.728914754	13.676	-6.6644	2.678894285

Figure14: comparison of parameters Mean, Max, Min and STDEVP for LogDpH7.4 for all chemicals for Daphnia endpoint

Following is the result of the calculated parameters for the descriptor (LogDpH9) for all chemical compounds which shows the value for both Pallas and ACD.



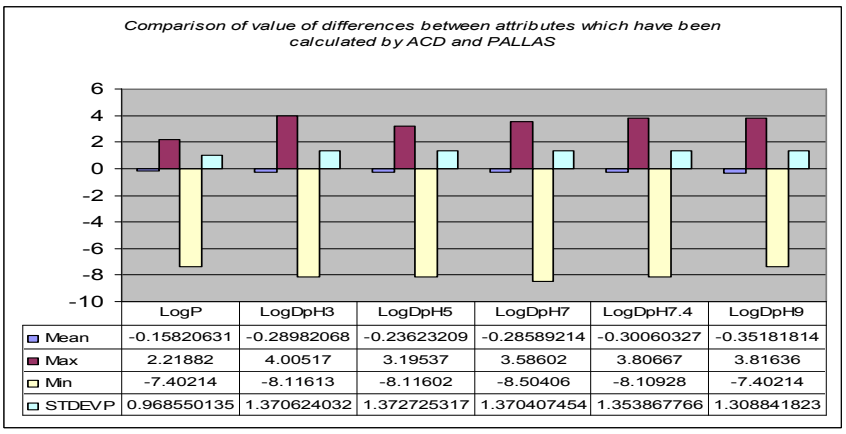
	Mean	Max	Min	STDEVP
PALLAS007	2.306894979	11.6915	-9.10505	3.029498943
ACD006	2.658713115	13.676	-6.8685	2.743352682

Figure15: comparison of parameters Mean, Max, Min and STDEVP for LogDpH9 for all chemicals for Daphnia endpoint

Following shows the result of the second experiment on data, which presents the statistical parameters for value differences between each descriptor that, has been presented by ACD and Pallas (visualize by two different graphs).

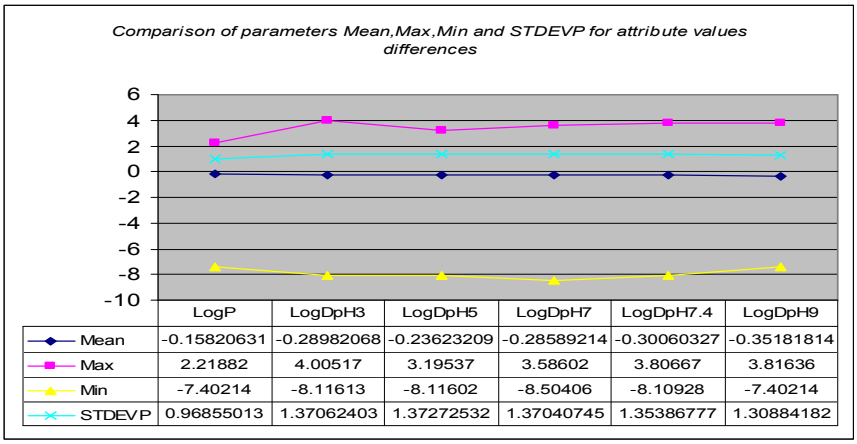
Table13: Results of the experiment for calculation of value difference of the chemicals presented by ACD and Pallas for Daphnia endpoint

Pallas value minus ACD value	**LogP**	**LogDpH3**	**LogDpH5**	**LogDpH7**	**LogDpH7.4**	**LogDpH9**
Mean	-0.158206308	-0.289820677	-0.236232093	-0.28589214	-0.300603268	-0.351818136
Max	2.21882	4.00517	3.19537	3.58602	3.80667	3.81636
Min	-7.40214	-8.11613	-8.11602	-8.50406	-8.10928	-7.40214
STDEVP	0.968550135	1.370624032	1.372725317	1.370407454	1.353867766	1.308841823



	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Mean	-0.15820631	-0.28982068	-0.23623209	-0.28589214	-0.30060327	-0.35181814
Max	2.21882	4.00517	3.19537	3.58602	3.80667	3.81636
Min	-7.40214	-8.11613	-8.11602	-8.50406	-8.10928	-7.40214
STDEVP	0.968550135	1.370624032	1.372725317	1.370407454	1.353867766	1.308841823

Figure16: comparison of value difference for all chemicals for Daphnia endpoint



	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Mean	-0.15820631	-0.28982068	-0.23623209	-0.28589214	-0.30060327	-0.35181814
Max	2.21882	4.00517	3.19537	3.58602	3.80667	3.81636
Min	-7.40214	-8.11613	-8.11602	-8.50406	-8.10928	-7.40214
STDEVP	0.96855013	1.37062403	1.37272532	1.37040745	1.35386777	1.30884182

Figure17: comparison of value difference for all chemicals for Daphnia endpoint showing by different graph

For comparison purposes all the values of max and min for descriptors and also minimum and maximum value difference put in one table (table 14). The table also shows the ID for descriptors which hold the minimum and maximum value for that descriptor.

Table14: Results of the experiment for calculation of value difference, Min, Max, and also ID of the chemicals presented by ACD and Pallas for Daphnia endpoint

Pallas	**LogP**	**LogDpH3**	**LogDpH5**	**LogDpH7**	**LogDpH7.4**	**LogDpH9**
Min	-2.701	-6.54063	-6.54052	-7.85046	-7.89228	-9.10505
Max	11.6915	11.6915	11.6915	11.6915	11.6915	11.6915
Min Value Difference	-7.40214	-8.11613	-8.11602	-8.50406	-8.10928	-7.40214
Max Value Difference	2.21882	4.00517	3.19537	3.58602	3.80667	3.81636
ID of Min	346	51	51	417	417	143
ID of Max	418	418	418	418	418	418
ACD	**LogP**	**LogDpH3**	**LogDpH5**	**LogDpH7**	**LogDpH7.4**	**LogDpH9**
Min	-2.3559	-5.4966	-5.8715	-6.499	-6.6644	-6.8685
Max	13.676	13.676	13.676	13.676	13.676	13.676
Min Value Difference	-7.40214	-8.11613	-8.11602	-8.50406	-8.10928	-7.40214
Max Value Difference	2.21882	4.00517	3.19537	3.58602	3.80667	3.81636
ID of Min	143	143	143	143	143	143
ID of Max	90	90	90	90	90	90

Analysis: as it shown in table 14 the value difference for LogP calculated by ACD and Pallas ranges from Min –7.40214 to Max 2.2188. Considering the numerical value the difference of 7.40214 is extremely high and unreliable. This value belong to compound with the ID=90. This means there is difference of 7.40214 between the value for LogP calculated for this compound by ACD and PALLAS. The results also showed graphically in figures16 and 17. The variation for compounds with ID=39, 415 and 419 is also very high, ranging from –2.389 to –3.300. The ID for compounds holding the minimum and maximum values for each descriptor, are also different. For instance in files produced by ACD the compound with biggest LogP value has ID, 90

but in PALLAS that is calculated for compound with ID, 418. For other descriptors there are also same inconsistencies. This creates doubts in reliability to produced values by these two programs.

4.2.3.3 Trout Endpoint

As first experiment four following parameters have been calculated for each column in each file (in ACD and Pallas): Minimum, Maximum, Average (mean), Standard Deviation (for population) using Excel functions and recorded in the bottom of the each column. The result for each descriptor was separated in different table to show the difference of these parameters for descriptors which values calculated with ACD and Pallas. The results are shown on charts to visualize the comparison. Following are the results achieved from this experiment. The analysis follows at the end of this section.

Table15: Comparison of parameters Mean, Min, Max, and STDEVP of the chemicals presented by ACD and Pallas for Trout endpoint

T3.1v2_2DPallas_v1	PALLAS001	PALLAS002	PALLAS003	PALLAS004	PALLAS005	PALLAS006
Mean	3.264245559	2.837395973	2.876839175	2.781595156	2.760571741	2.676192422
Max	8.68196	8.68196	8.68196	8.68196	8.68196	8.68196
Min	-2.701	-6.54063	-6.54052	-6.5299	-6.51344	-9.10505
STDEVP	2.024651218	2.418369574	2.401885891	2.477602322	2.501244884	2.628349429
T3.1v2_2DACDv1	ACD001	ACD002	ACD003	ACD004	ACD005	ACD006
Mean	3.401836641	3.137187786	3.086464122	2.963532443	2.945440458	2.878457252
Max	13.676	13.676	13.676	13.676	13.676	13.676
Min	-2.3559	-5.4966	-5.8715	-6.499	-6.6644	-6.8685
STDEVP	2.053151135	2.269363202	2.333381275	2.425906057	2.443537865	2.506445732

Following is the result of the calculated parameters for the descriptor (LogP) for all chemical compounds which shows the value for both Pallas and ACD.

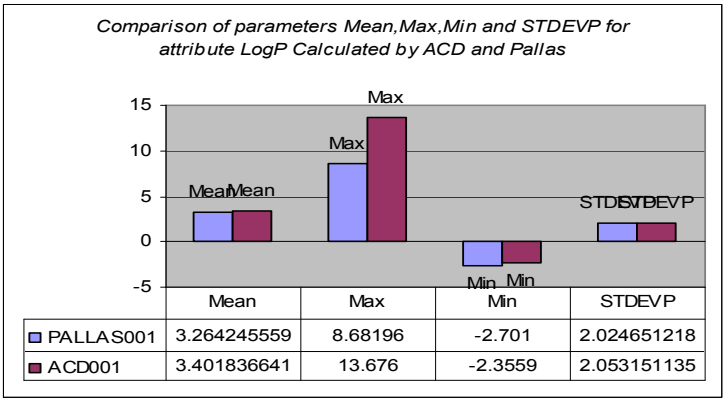
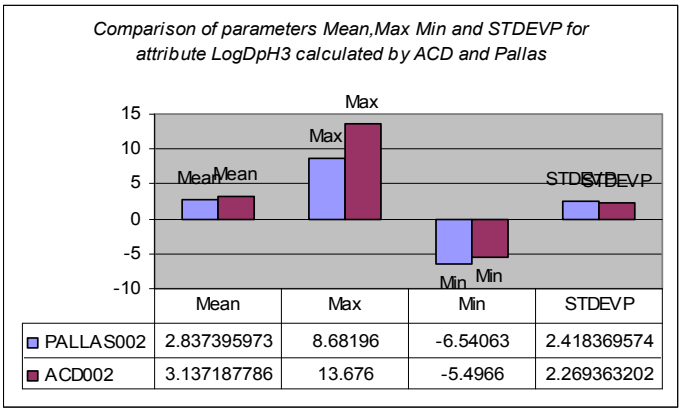


Figure18: comparison of parameters Mean, Max, Min and STDEVP for LogP for Trout endpoint

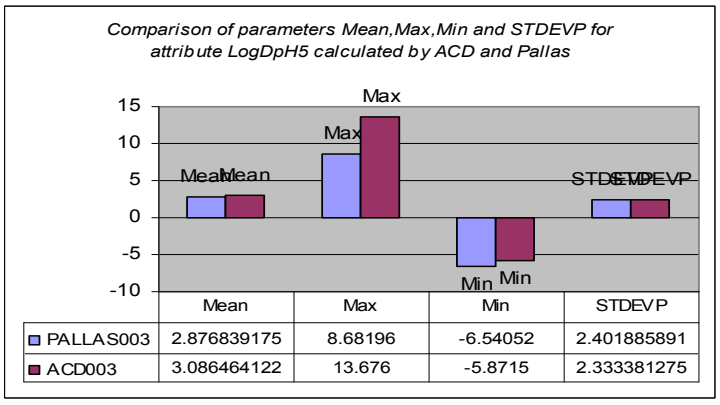
Following is the result of the calculated parameters for the descriptor (LogDpH3) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS002	2.837395973	8.68196	-6.54063	2.418369574
ACD002	3.137187786	13.676	-5.4966	2.269363202

Figure19: comparison of parameters Mean, Max, Min and STDEVP for LogDpH3 for Trout endpoint

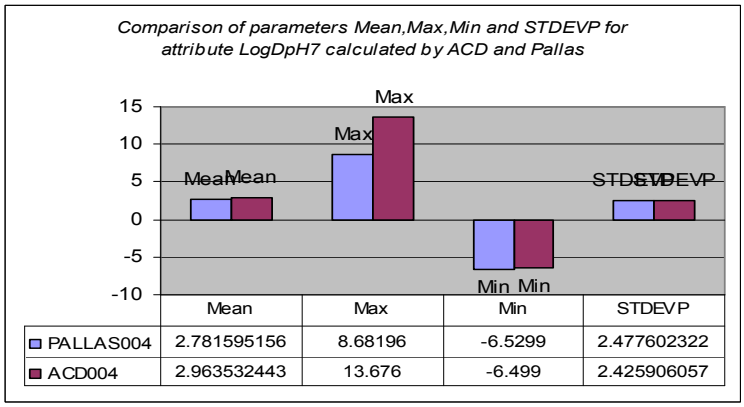
Following is the result of the calculated parameters for the descriptor (LogDpH5) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS003	2.876839175	8.68196	-6.54052	2.401885891
ACD003	3.086464122	13.676	-5.8715	2.333381275

Figure20: comparison of parameters Mean, Max, Min and STDEVP for LogDpH5 for Trout endpoint

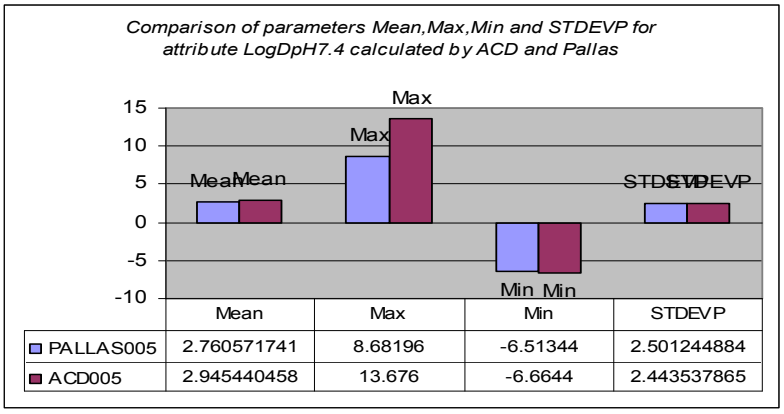
Following is the result of the calculated parameters for the descriptor (LogDpH7) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS004	2.781595156	8.68196	-6.5299	2.477602322
ACD004	2.963532443	13.676	-6.499	2.425906057

Figure21: comparison of parameters Mean, Max, Min and STDEVP for LogDpH7 for Trout endpoint

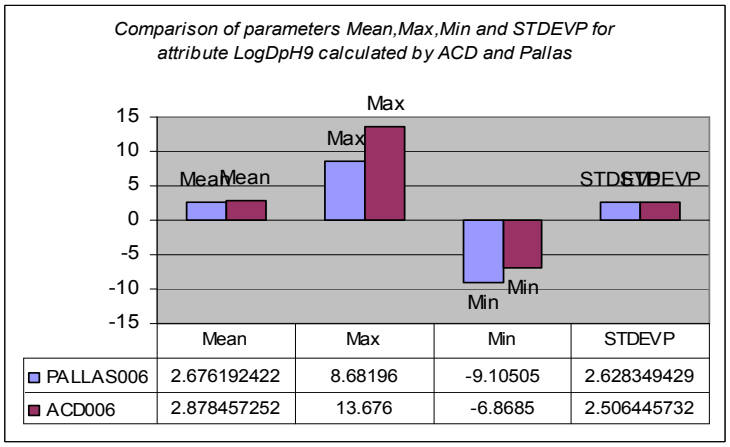
Following is the result of the calculated parameters for the descriptor (LogDpH7.4) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS005	2.760571741	8.68196	-6.51344	2.501244884
ACD005	2.945440458	13.676	-6.6644	2.443537865

Figure22: comparison of parameters Mean, Max, Min and STDEVP for LogDpH7.4 for Trout endpoint

Following is the result of the calculated parameters for the descriptor (LogDpH9) for all chemical compounds which shows the value for both Pallas and ACD.



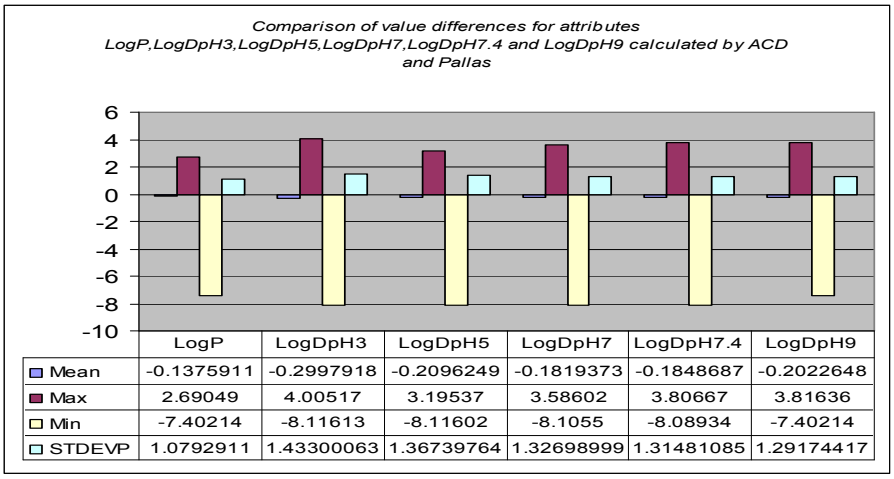
	Mean	Max	Min	STDEVP
PALLAS006	2.676192422	8.68196	-9.10505	2.628349429
ACD006	2.878457252	13.676	-6.8685	2.506445732

Figure23: comparison of parameters Mean, Max, Min and STDEVP for LogDpH9 for Trout endpoint

Following shows the result of the second experiment on data which presents the statistical parameters for value differences between each descriptor that has been presented by ACD and Pallas (visualize by two different graphs).

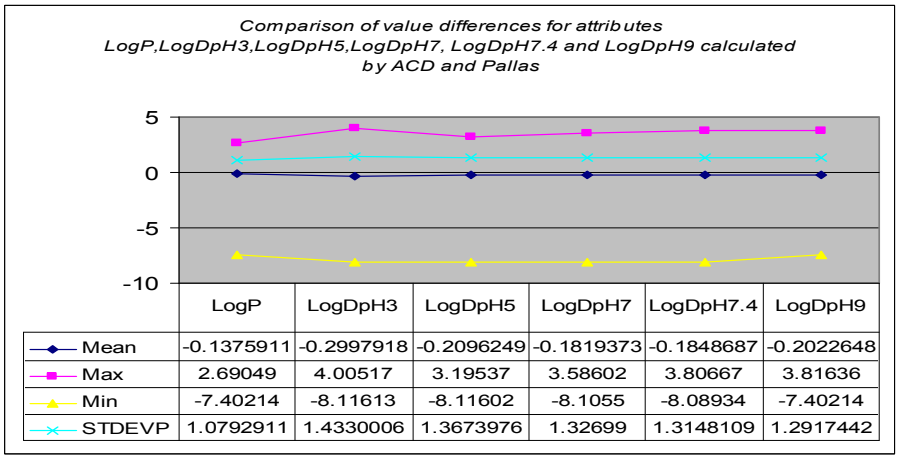
Table16: Results of the experiment for calculation of value difference between each descriptor presented by ACD and Pallas for Trout endpoint

Pallas value minus ACD value	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Mean	-0.137591082	-0.299791814	-0.209624947	-0.181937287	-0.184868717	-0.20226483
Max	2.69049	4.00517	3.19537	3.58602	3.80667	3.81636
Min	-7.40214	-8.11613	-8.11602	-8.1055	-8.08934	-7.40214
STDEVP	1.079291098	1.43300063	1.367397639	1.326989991	1.314810855	1.291744174



	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Mean	-0.1375911	-0.2997918	-0.2096249	-0.1819373	-0.1848687	-0.2022648
Max	2.69049	4.00517	3.19537	3.58602	3.80667	3.81636
Min	-7.40214	-8.11613	-8.11602	-8.1055	-8.08934	-7.40214
STDEVP	1.0792911	1.43300063	1.36739764	1.32698999	1.31481085	1.29174417

Figure24: comparison of value difference for attributes for trout endpoint



	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Mean	-0.1375911	-0.2997918	-0.2096249	-0.1819373	-0.1848687	-0.2022648
Max	2.69049	4.00517	3.19537	3.58602	3.80667	3.81636
Min	-7.40214	-8.11613	-8.11602	-8.1055	-8.08934	-7.40214
STDEVP	1.0792911	1.4330006	1.3673976	1.32699	1.3148109	1.2917442

Figure25: comparison of value difference for attributes for trout endpoint by different graph

For comparison purposes all the values of max and min for descriptors and also max and min descriptors value difference put in one table (table 17).The table also shows the ID for descriptors which hold the minimum and maximum value for that descriptor.

Table17: Results of the experiment for calculation of value difference, Min, Max, and also ID of the chemicals presented by ACD and Pallas for Trout endpoint

Pallas	**LogP**	**LogDpH3**	**LogDpH5**	**LogDpH7**	**LogDpH7.4**	**LogDpH9**
Min	-2.701	-6.54063	-6.54052	-6.5299	-6.51344	-9.10505
Max	8.68196	8.68196	8.68196	8.68196	8.68196	8.68196
Min Value Difference	-7.40214	-8.11613	-8.11602	-8.1055	-8.08934	-7.40214
Max Value Difference	2.69049	4.00517	3.19537	3.58602	3.80667	3.81636
ID of Min	346	51	51	51	51	143
ID of Max	93	93	93	93	93	93
ACD	**LogP**	**LogDpH3**	**LogDpH5**	**LogDpH7**	**LogDpH7.4**	**LogDpH9**
Min	-2.3559	-5.4966	-5.8715	-6.499	-6.6644	-6.8685
Max	13.676	13.676	13.676	13.676	13.676	13.676
Min Value Difference	-7.40214	-8.11613	-8.11602	-8.1055	-8.08934	-7.40214
Max Value Difference	2.69049	4.00517	3.19537	3.58602	3.80667	3.81636
ID of Min	143	143	143	143	143	143
ID of Max	90	90	90	90	90	90

Analysis: as it shown in table 17 and also in figures24 and 25**,** the value comparison shows that there are high variations. For LogP the value difference starts from –7.4021 to 2.69049. This variation continues across table for all other descriptors. For this endpoint also the ID for compounds which hold the minimum value and maximum value for descriptors are different in two files. For instance the ID for a compound with minimum value for LogP presented by ACD is 143 and in PALLAS are 346.

4.2.3.4 DietaryQuail Endpoint

As first experiment four following parameters have been calculated for each column in each file (in ACD and Pallas): Minimum, Maximum, Average (mean), Standard Deviation (for population) using Excel functions and recorded in the bottom of the each column. The result for each descriptor was separated in different table to show the difference of these parameters for descriptors which values calculated with ACD and Pallas. The results are shown on charts to visualize the comparison. Following are the results achieved from this experiment.

Table18: Calculation of values Mean, Min, Max, and STDEVP for DQ endpoint

DQ_2DPallas	PALLAS001	PALLAS002	PALLAS003	PALLAS005	PALLAS006	PALLAS007
Mean	3.57314711	3.23550601	3.20153731	3.07774728	3.05740324	3.01022853
Max	8.16996	8.16996	8.16996	8.16996	8.16996	8.16996
Min	-2.244	-2.36733	-2.51701	-3.92561	-3.99634	-4.04894
STDEVP	2.13345441	2.3971412	2.37181452	2.42087622	2.42777757	2.4655906
DQ_1v1_ACD2D_v1	ACD001	ACD002	ACD003	ACD004	ACD005	ACD006
Mean	3.55633271	3.34428879	3.28468037	3.12487664	3.10740467	3.0653486
Max	8.5027	8.5012	8.5008	8.4633	8.4118	8.1678
Min	-1.4136	-1.4184	-2.5887	-5.6348	-5.8515	-5.9949
STDEVP	2.06740742	2.12813686	2.2116162	2.45152799	2.48180499	2.52838837

Following is the result of the calculated parameters for the descriptor (LogP) for all chemical compounds which shows the value for both Pallas and ACD.

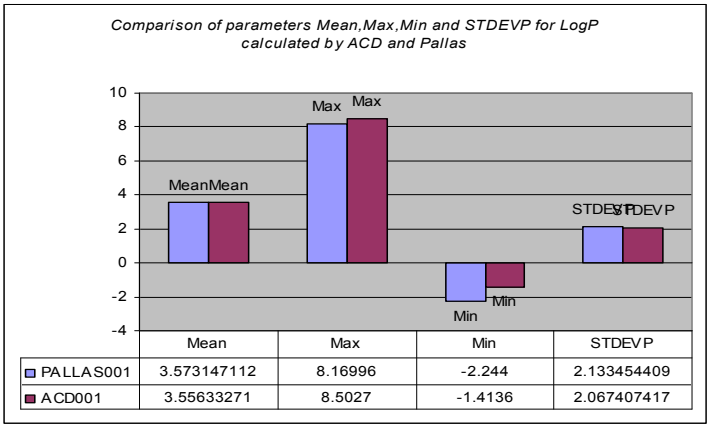
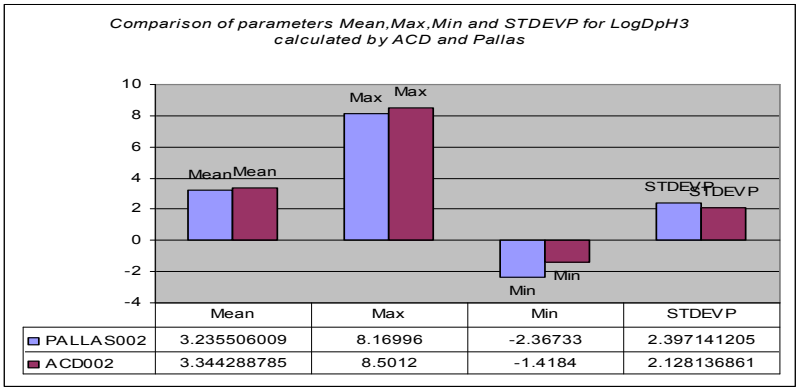


Figure26: comparison of parameters Mean, Max, Min and STDEVP for LogP for DQ endpoint

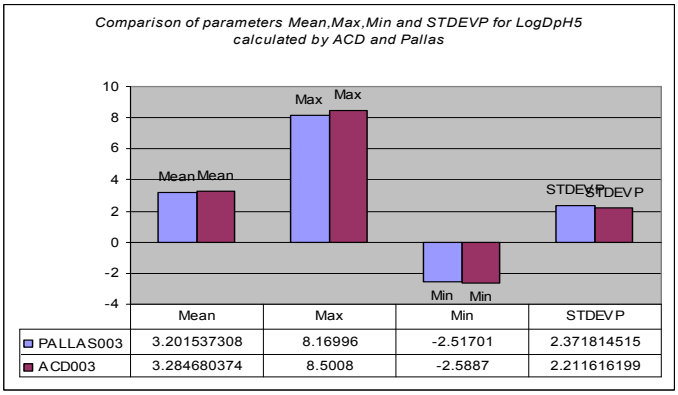
Following is the result of the calculated parameters for the descriptor (LogDpH3) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS002	3.235506009	8.16996	-2.36733	2.397141205
ACD002	3.344288785	8.5012	-1.4184	2.128136861

Figure27: comparison of parameters Mean, Max, Min and STDEVP for LogDpH3 for DQ endpoint

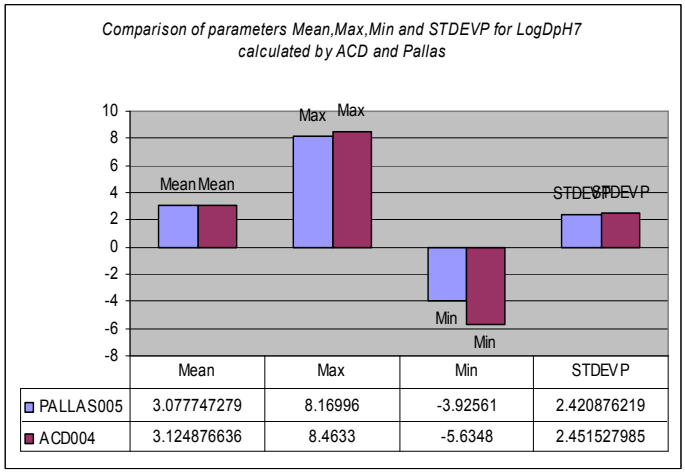
Following is the result of the calculated parameters for the descriptor (LogDpH5) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS003	3.201537308	8.16996	-2.51701	2.371814515
ACD003	3.284680374	8.5008	-2.5887	2.211616199

Figure28: comparison of parameters Mean, Max, Min and STDEVP for LogDpH5 for DQ endpoint

Following is the result of the calculated parameters for the descriptor (LogDpH7) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS005	3.077747279	8.16996	-3.92561	2.420876219
ACD004	3.124876636	8.4633	-5.6348	2.451527985

Figure29: comparison of parameters Mean, Max, Min and STDEVP for LogDpH7 for DQ endpoint

Following is the result of the calculated parameters for the descriptor (LogDpH7.4) for all chemical compounds which shows the value for both Pallas and ACD.

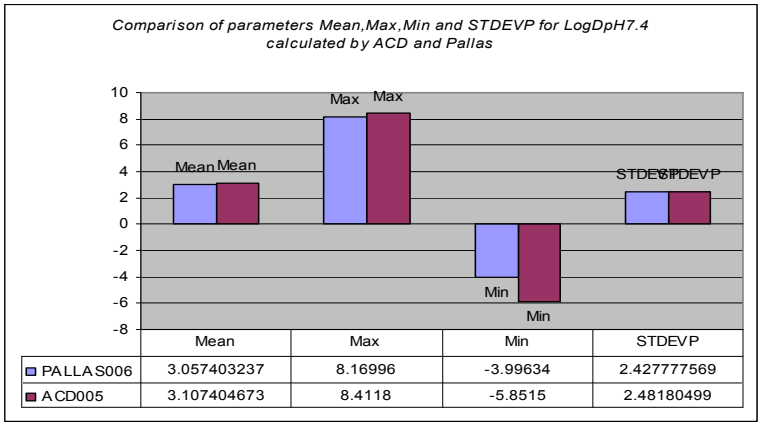


Figure30: comparison of parameters Mean, Max, Min and STDEVP for LogDpH7.4 for DQ endpoint

Following is the result of the calculated parameters for the descriptor (LogDpH9) for all chemical compounds which shows the value for both Pallas and ACD.

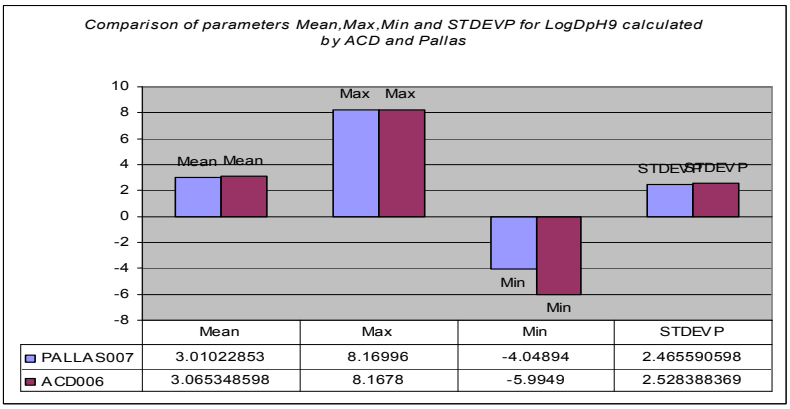
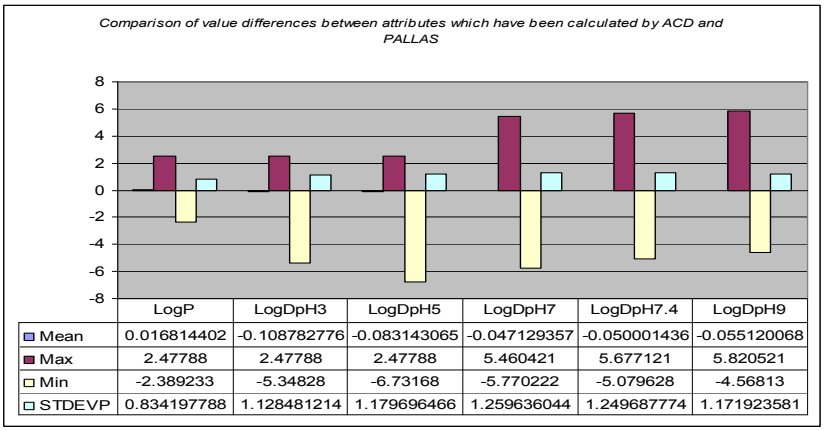


Figure31: comparison of parameters Mean, Max, Min and STDEVP for LogDpH9 for DQ endpoint

Following shows the result of the second experiment on data which presents the statistical parameters for value differences between each descriptor that has been presented by ACD and Pallas (visualize by two different graphs).

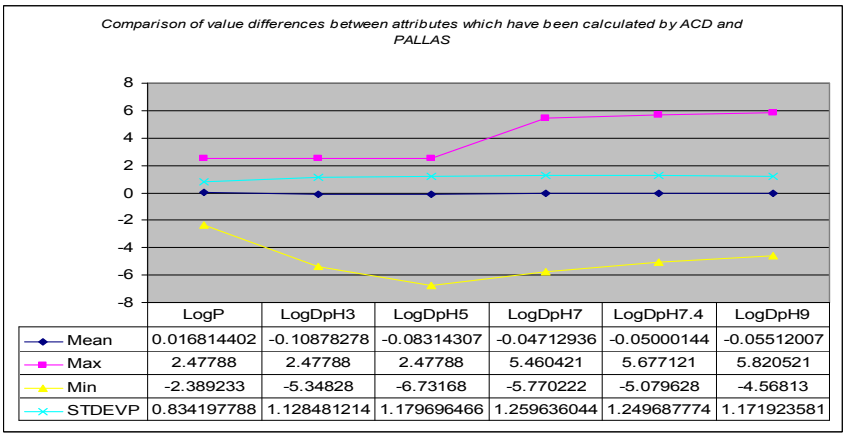
Table19: Results of the experiment for calculation of value difference between each descriptor presented by ACD and Pallas for DQ endpoint

Pallas value minus ACD value	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Mean	0.016814402	-0.108782776	-0.083143065	-0.047129357	-0.050001436	-0.055120068
Max	2.47788	2.47788	2.47788	5.460421	5.677121	5.820521
Min	-2.389233	-5.34828	-6.73168	-5.770222	-5.079628	-4.56813
STDEVP	0.834197788	1.128481214	1.179696466	1.259636044	1.249687774	1.171923581



	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Mean	0.016814402	-0.108782776	-0.083143065	-0.047129357	-0.050001436	-0.055120068
Max	2.47788	2.47788	2.47788	5.460421	5.677121	5.820521
Min	-2.389233	-5.34828	-6.73168	-5.770222	-5.079628	-4.56813
STDEVP	0.834197788	1.128481214	1.179696466	1.259636044	1.249687774	1.171923581

Figure32: comparison of value difference for all chemicals for DQ endpoint



	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Mean	0.016814402	-0.10878278	-0.08314307	-0.04712936	-0.05000144	-0.05512007
Max	2.47788	2.47788	2.47788	5.460421	5.677121	5.820521
Min	-2.389233	-5.34828	-6.73168	-5.770222	-5.079628	-4.56813
STDEVP	0.834197788	1.128481214	1.179696466	1.259636044	1.249687774	1.171923581

Figure33: comparison of value difference for all chemicals for DQ endpoint by different graph

For comparison purposes all the calculated values of max and min for descriptors and also maximum and minimum value difference between descriptors, put in one table (table 20). The table also shows the ID for descriptors which hold the minimum and maximum value for that descriptor.

Table20: Results of the experiment for calculation of value difference, Min, Max, and also ID of the chemicals presented by ACD and Pallas for DQ endpoint

Pallas	**LogP**	**LogDpH3**	**LogDpH5**	**LogDpH7**	**LogDpH7.4**	**LogDpH9**
Min	-2.244	-2.36733	-2.51701	-3.92561	-3.99634	-4.04894
Max	8.16996	8.16996	8.16996	8.16996	8.16996	8.16996
Min Value Difference	-2.38923	-5.34828	-6.73168	-5.77022	-5.07963	-4.56813
Max Value Difference	2.47788	2.47788	2.47788	5.460421	5.677121	5.820521
ID of Min	442	337	230	230	230	230
ID of Max	146	146	146	146	146	146
ACD	**LogP**	**LogDpH3**	**LogDpH5**	**LogDpH7**	**LogDpH7.4**	**LogDpH9**
Min	-1.4136	-1.4184	-2.5887	-5.6348	-5.8515	-5.9949
Max	8.5027	8.5012	8.5008	8.4633	8.4118	8.1678
Min Value Difference	-2.38923	-5.34828	-6.73168	-5.77022	-5.07963	-4.56813
Max Value Difference	2.47788	2.47788	2.47788	5.460421	5.677121	5.820521
ID of Min	447	447	230	447	447	447
ID of Max	411	411	411	411	411	146

Analysis: as it shown in table 20 and figures32 and 33, there are sudden picks for LogDpH5 which carries for LogDpH7, LogDpH7.4 and LogDpH9. The difference between Min value calculated for ACD and Pallas for LogP and LogDpH3 is very high. There are also differences between the compounds which hold the maximum and minimum values for descriptors across two files. For instance in the PALLAS file the ID for compound with minimum LogP value is 442 but in ACD is 447.

4.2.3.5 OralQuail Endpoint

As first experiment four following parameters have been calculated for each column in each file (in ACD and Pallas): Minimum, Maximum, Average (mean), Standard Deviation (for population) using Excel functions and recorded in the bottom of the each column. The result for each descriptor was separated in different table to show the difference of these parameters for descriptors which values calculated with ACD and Pallas. The result was showed on charts to visualize the comparison. Following are the results achieved from this experiment.

Table21: Results of the experiment for calculation of parameters Mean, Min, Max, and STDEVP of the chemicals presented by ACD and Pallas for OQ endpoint

OQ_2DPallas_v1	PALLAS001	PALLAS002	PALLAS003	PALLAS005	PALLAS006	PALLAS007
Mean	2.882384377	2.356942575	2.225951613	2.050985316	2.038297758	2.036627778
Max	8.16996	8.16996	8.16996	8.16996	8.16996	8.16996
Min	-2.82609	-6.54063	-6.54052	-6.5299	-7.05401	-7.78253
STDEVP	2.207122552	2.68692086	2.701264404	2.740502259	2.741228452	2.716455172
OQ1_1v1_ACD2Dv1	**ACD001**	**ACD002**	**ACD003**	**ACD004**	**ACD005**	**ACD006**
Mean	3.052807692	2.684603846	2.597591346	2.466474038	2.454674038	2.419223077
Max	13.676	13.676	13.676	13.676	13.676	13.676
Min	-1.6559	-4.7181	-3.763	-4.6224	-5.0095	-5.2056
STDEVP	2.198203905	2.351511638	2.372459659	2.487731672	2.503263771	2.497835524

Following is the result of the calculated parameters for the descriptor (LogP) for all chemical compounds which shows the value for both Pallas and ACD.

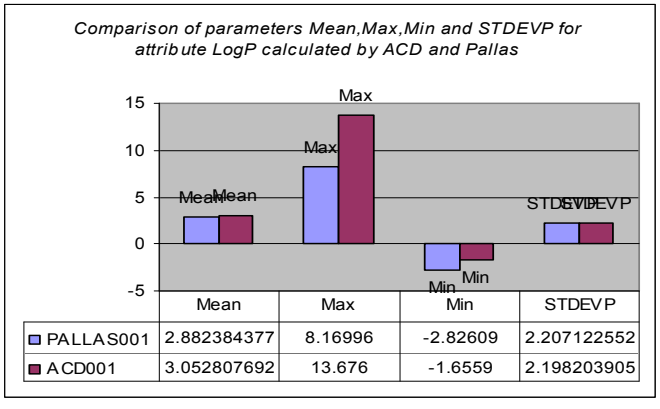
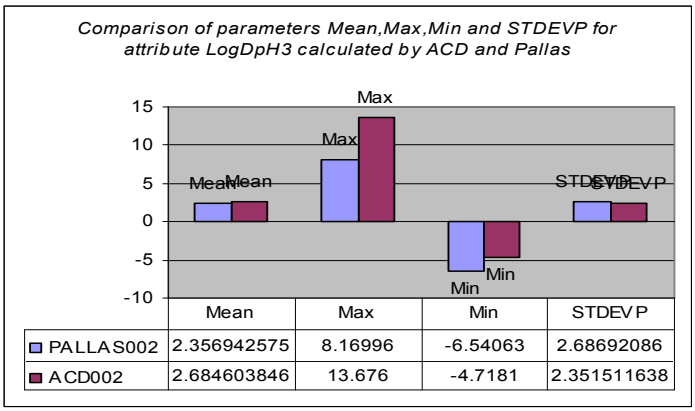


Figure34: comparison of parameters Mean, Max, Min and STDEVP for LogP for OQ endpoint

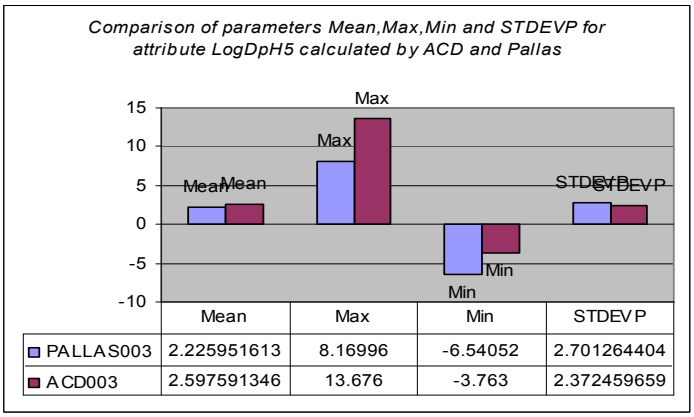
Following is the result of the calculated parameters for the descriptor (LogDpH3) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS002	2.356942575	8.16996	-6.54063	2.68692086
ACD002	2.684603846	13.676	-4.7181	2.351511638

Figure35: comparison of parameters Mean, Max, Min and STDEVP for LogDpH3 for OQ endpoint

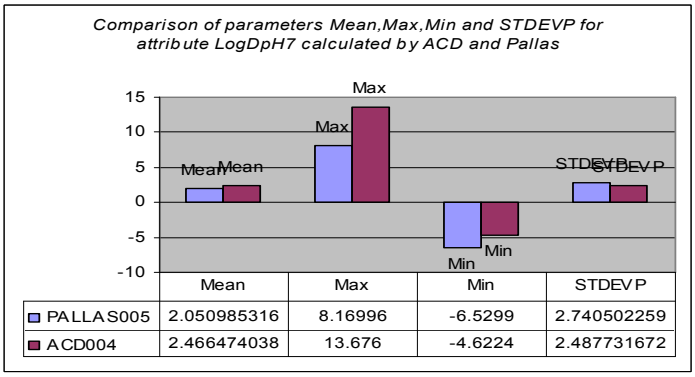
Following is the result of the calculated parameters for the descriptor (LogDpH5) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS003	2.225951613	8.16996	-6.54052	2.701264404
ACD003	2.597591346	13.676	-3.763	2.372459659

Figure36: comparison of parameters Mean, Max, Min and STDEVP for LogDpH5 for OQ endpoint

Following is the result of the calculated parameters for the descriptor (LogDpH7) for all chemical compounds which shows the value for both Pallas and ACD.



	Mean	Max	Min	STDEVP
PALLAS005	2.050985316	8.16996	-6.5299	2.740502259
ACD004	2.466474038	13.676	-4.6224	2.487731672

Figure37: comparison of parameters Mean, Max, Min and STDEVP for LogDpH7 for OQ endpoint

Following is the result of the calculated parameters for the descriptor (LogDpH7.4) for all chemical compounds which shows the value for both Pallas and ACD.

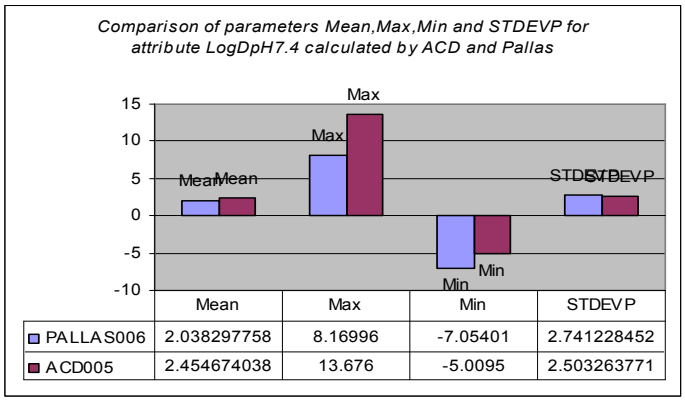


Figure38: comparison of parameters Mean, Max, Min and STDEVP for LogDpH7.4 for OQ endpoint

Following is the result of the calculated parameters for the descriptor (LogDpH9) for all chemical compounds which shows the value for both Pallas and ACD.

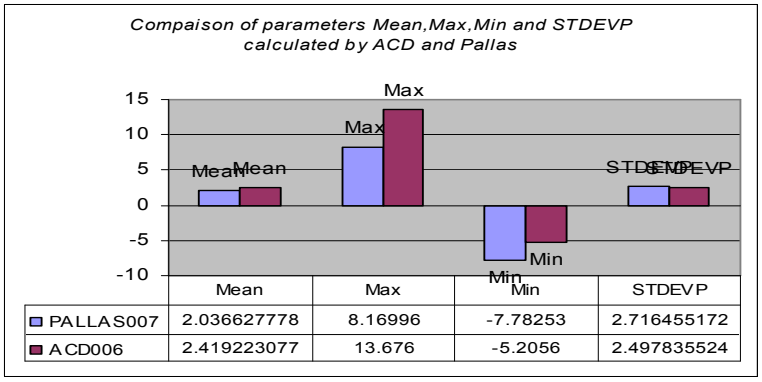
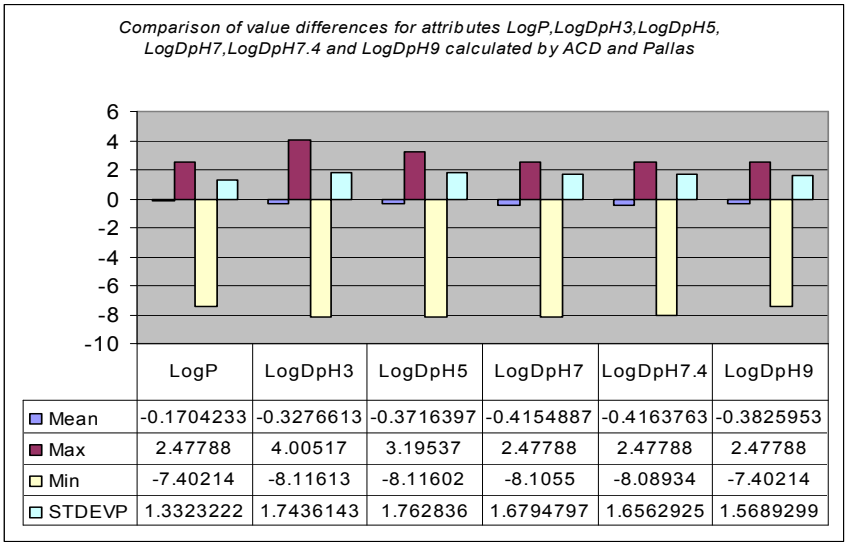


Figure39: comparison of parameters Mean, Max, Min and STDEVP for LogDpH9 for OQ endpoint

Following shows the result of the second experiment on data which presents the statistical parameters for value differences between each descriptor that has been presented by ACD and Pallas (visualize by two different graphs).

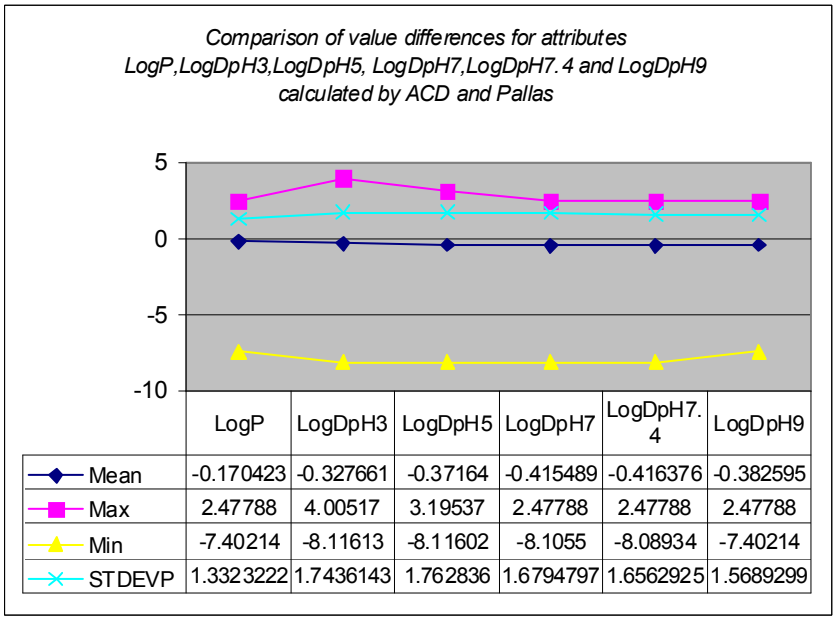
Table22: Results of the experiment for calculation of value difference between each descriptor presented by ACD and Pallas for OQ endpoint

Pallas value minus ACD value	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Mean	-0.170423315	-0.327661271	-0.371639733	-0.415488722	-0.416376281	-0.382595299
Max	2.47788	4.00517	3.19537	2.47788	2.47788	2.47788
Min	-7.40214	-8.11613	-8.11602	-8.1055	-8.08934	-7.40214
STDEVP	1.332322235	1.743614319	1.762835989	1.679479728	1.656292526	1.568929912



	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Mean	-0.1704233	-0.3276613	-0.3716397	-0.4154887	-0.4163763	-0.3825953
Max	2.47788	4.00517	3.19537	2.47788	2.47788	2.47788
Min	-7.40214	-8.11613	-8.11602	-8.1055	-8.08934	-7.40214
STDEVP	1.3323222	1.7436143	1.762836	1.6794797	1.6562925	1.5689299

Figure40: comparison of value difference for attributes for OQ endpoint



	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Mean	-0.170423	-0.327661	-0.37164	-0.415489	-0.416376	-0.382595
Max	2.47788	4.00517	3.19537	2.47788	2.47788	2.47788
Min	-7.40214	-8.11613	-8.11602	-8.1055	-8.08934	-7.40214
STDEVP	1.3323222	1.7436143	1.762836	1.6794797	1.6562925	1.5689299

Figure41: comparison of value difference for attributes for OQ endpoint by different graph

For comparison purposes all the values of max and min for descriptors and also max and min descriptors value difference put in one table (table 23). The table also shows the ID for descriptors which hold the minimum and maximum value for that descriptor.

Table23: Results of the experiment for calculation of value difference, Min, Max, and also ID of the chemicals presented by ACD and Pallas for OQ endpoint

Pallas	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	-2.82609	-6.54063	-6.54052	-6.5299	-7.05401	-7.78253
Max	8.16996	8.16996	8.16996	8.16996	8.16996	8.16996
Min Value Difference	-7.40214	-8.11613	-8.11602	-8.1055	-8.08934	-7.40214
Max Value Difference	2.47788	4.00517	3.19537	2.47788	2.47788	2.47788
ID of Min	433	51	51	51	372	372
ID of Max	146	146	146	146	146	146
ACD	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	-1.6559	-4.7181	-3.763	-4.6224	-5.0095	-5.2056
Max	13.676	13.676	13.676	13.676	13.676	13.676
Min Value Difference	-7.40214	-8.11613	-8.11602	-8.1055	-8.08934	-7.40214
Max Value Difference	2.47788	4.00517	3.19537	2.47788	2.47788	2.47788
ID of Min	347	347	347	372	372	372
ID of Max	90	90	90	90	90	90

Analysis: as it shown in figures40 and 41 and also in table 23, the value difference calculated for each descriptor for all chemical compounds are very high. There are also differences between values for chemicals with lowest and highest values for descriptors across both files. For instance the ID for chemical compound with minimum value for LogP is 347 in ACD file and is 433 in PALLAS.

4.2.4 Comparison of Model Performance

Original data sets (prepared for training and testing) were used to develop Weka models based on the following algorithms: ClassificationViaRegression, BayesNet, MultilayerPerceptron, IBK, ZeroR, LMT, J48 and JRip [58]. For performance of models study, two case studies have been considered. Firstly models obtained from training data (separated inputs from ACD and Pallas for same endpoint) were tested against test data sets. Secondly 10-fold Cross Validation has been used on training data. The accuracy of each model (one from modeling against testing set and one from modeling with Cross Validation testing method) was recorded to identify which model suits which endpoint. Other parameters from modeling can also be recorded. We compare classification accuracy for models obtained as described above, once using training set against test set and once using 10-fold Cross Validation with 8 algorithms on all the endpoints.

4.2.5 Descriptor Swap (LogP, LogDpH3, LogDpH5, LogDpH7)

For this work first the value difference from previous experiment were considered. If the value was big, then number of descriptors was swap between files produced by ACD and Pallas. Then the produced files were modelled using all the algorithms used

in previous work using Weka. At the last stage results were compared with the ones collected from the previous work (containing files with their original descriptors values modelled using Weka). First time LogP was swapped between two files for all endpoint and data was modelled. Second time for just two endpoints (with large descriptors value difference) three more descriptors were also swapped and the data was modelled and result was recorded. The idea of this task was to see how much the variation of values could affect the model performance. Table 23 and 24 shows the result of modelling PALLAS and ACD dataset for trout endpoint in its original form (with no swapping) and then after LogP swap, LogP and LogDpH5 and LogP, LogDpH5 and LogDpH7 swap between two files (ACD and PALLAS). The algorithms were trained against test in this instance.

Table23: Model performance after descriptor swap for Trout dataset produced by Pallas

Endpoint Trout	Algorithm accuracy against test set (%)							
Pallas	**BN**	**MLP**	**IBK**	**CVR**	**J48**	**JRip**	**ZeroR**	**LMT**
Original Model	42.5	47.5	40	40	42.5	45	40	42.5
LogP swap	58.6957	54.3478	39.1304	56.5217	52.1739	50	43.4783	56.5217
LogP & LogDpH5 swap	58.6957	45.6522	58.6957	43.4783	52.1739	56.5217	43.4783	47.8261
LogP,LogDpH5 & LogDpH7 swap	56.5217	56.5217	43.4783	56.5217	58.6957	47.8261	43.4783	56.5217

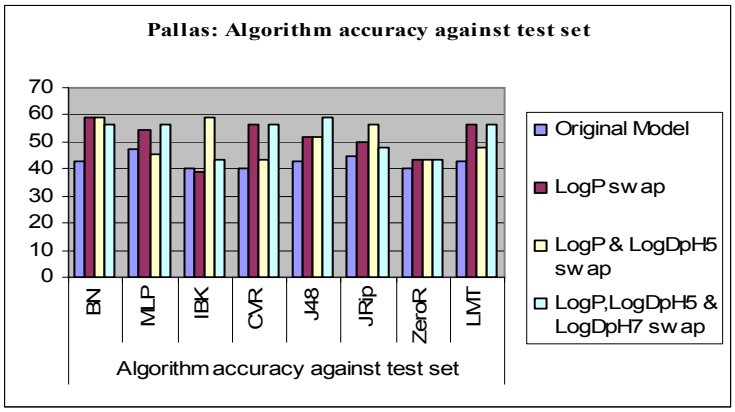


Figure42: Algorithm accuracy after descriptor swap for Trout dataset produced by Pallas

Table24: Model performance after descriptor swap for Trout dataset produced by ACD

Endpoint Trout	Algorithm accuracy against test set (%)							
ACD	**BN**	**MLP**	**IBK**	**CVR**	**J48**	**Jrip**	**ZeroR**	**LMT**
Original Model	47.5	50	35	45	40	37.5	40	45
LogP swap	63.0435	54.3478	47.8261	58.6957	47.8261	43.4783	43.4783	50
LogP & LogDpH5 swap	56.5217	50	58.6957	39.1304	54.3478	60.8696	43.4783	41.3043
LogP,LogDpH5 & LogDpH7 swap	56.5217	56.5217	47.8261	50	50	54.3478	43.4783	47.8261

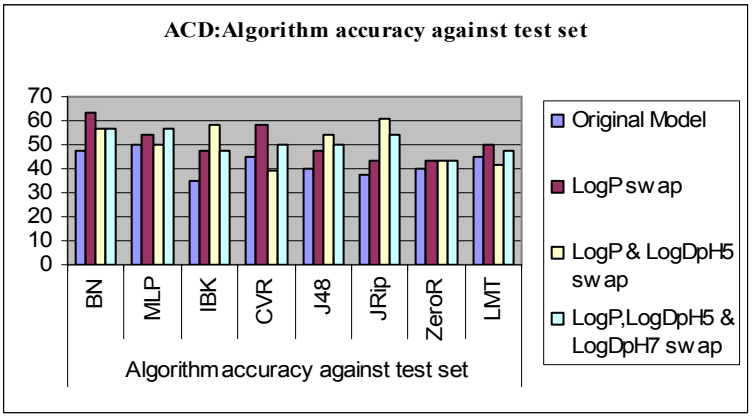


Figure43: Algorithm accuracy after descriptor swap for Trout dataset produced by ACD

Following tables 25 and 26 shows the result of modelling both datasets (ACD, PALLAS) after and before swapping. The data was modelled using 10-fold Cross Validation in this instance. The results are shown graphically in Figures 44 and 45.

Table25: Model performance after descriptor swap for Trout dataset produced by Pallas using 10-fold Cross Validation

Endpoint Trout	Algorithm accuracy tested by 10-fold Cross Validation (%)							
Pallas	**BN**	**MLP**	**IBK**	**CVR**	**J48**	**JRip**	**ZeroR**	**LMT**
Original Model	44.1176	47.549	38.2353	43.1373	48.5294	43.1373	44.6078	45.5882
LogP swap	57.4074	52.3148	44.4444	54.6296	50.9259	50.9259	44.9074	49.537
LogP & LogDpH5 swap	46.5686	49.5098	34.8039	48.5294	45.098	47.0588	44.6078	50
LogP,LogDpH5 & LogDpH7 swap	56.9444	52.3148	48.1481	51.3889	53.2407	57.4074	44.9074	46.2963

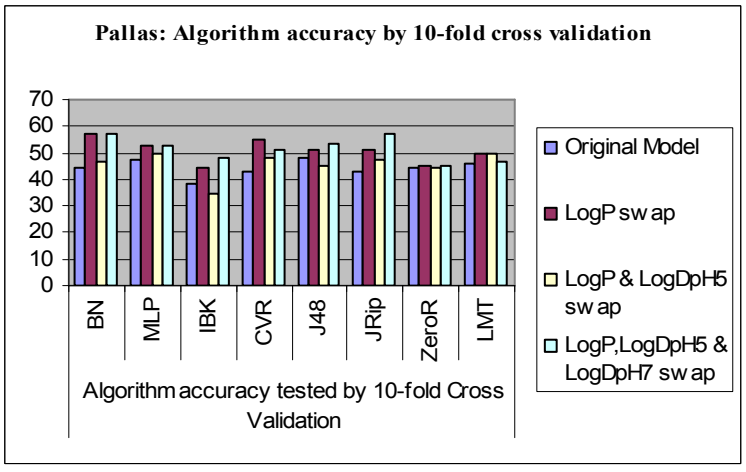


Figure44: Algorithm accuracy after descriptor swap for Trout dataset produced by Pallas using 10-fold Cross Validation

Table26: Model performance after descriptor swap for Trout dataset produced by ACD using 10-fold Cross Validation

Endpoint Trout	Algorithm accuracy tested by 10-fold Cross Validation (%)							
ACD	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
Original Model	42.6471	47.0588	42.1569	50.4902	48.0392	48.5294	44.6078	46.0784
LogP swap	56.0185	53.7037	47.6852	55.0926	54.6296	55.0926	44.9074	53.2407
LogP & LogDpH5 swap	45.5882	46.5686	38.7255	47.549	45.5882	45.098	44.6078	48.5294
LogP,LogDpH5 & LogDpH7 swap	58.3333	55.5556	46.7593	56.4815	53.7037	51.8519	44.9074	53.2407

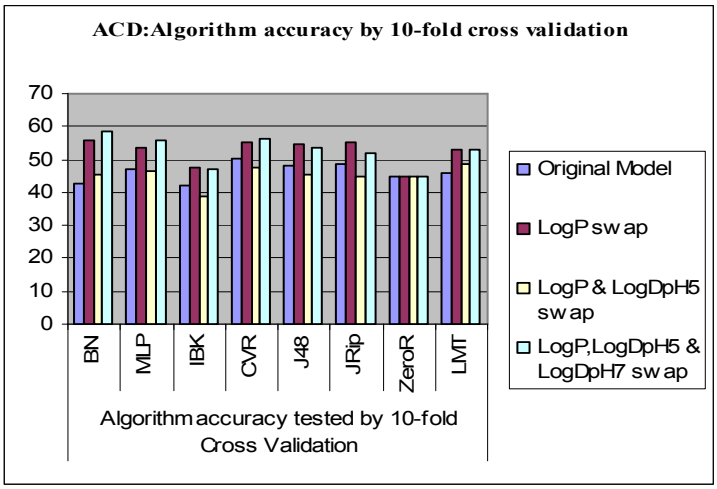


Figure45: Algorithm accuracy after descriptor swap for Trout dataset produced by ACD using 10-fold Cross Validation

Analysis: for this endpoint, modelling training set against test set the results have been improved dramatically. This shows correlation between swapped descriptors values in another file is more with rest of the descriptors. The results for 10-fold Cross Validation method have also been improved. For instance using BN algorithm with 10-fold Cross Validation, the result of the original modelling shows 42.6471% accuracy (table 26) but after LogP swap the results have increased to 56.0185%. In very few cases there is reduction in classification accuracy. For example in table 26 the result recorded for J48 algorithm for original file in ACD is 48.0392% but after swapping LogP & LogDpH5 the result decreased to 45.5882%.

These value ranges show more correlation could be used as guidelines for descriptors value bias. The results are more improved in the case of LogP, LogDpH5 and LogDpH7 swap in both cases.

Table27: Algorithms accuracy after descriptor swap for LogP in ACD and Pallas datasets for Bee endpoint

Endpoint	Algorithm accuracy against test set (%)							
Bee	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
(Pallas)-Before	31.25	37.5	18.75	37.5	37.5	37.5	31.25	37.5
(Pallas)-After	31.25	31.25	37.5	37.5	31.25	25	31.25	37.5
(ACD)-Before	31.25	37.5	31.25	31.25	25	31.25	31.25	37.5
(ACD)-After	31.25	37.5	25	31.25	31.25	37.5	31.25	37.5
Endpoint	**Algorithm accuracy tested by 10-fold Cross Validation (%)**							
Bee	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
(Pallas)-Before	41.7722	34.1772	32.9114	40.5063	31.6456	41.7722	41.7722	36.7089
(Pallas)-After	41.7722	31.6456	31.6456	41.7722	30.3797	40.5063	41.7722	40.5063
(ACD)-Before	35.443	36.7089	39.2405	39.2405	35.443	44.3038	41.7722	39.2405
(ACD)-After	35.443	21.519	35.443	39.2405	31.6456	44.3038	41.7722	40.5063

For this endpoint since the value difference between descriptors is not too high just one experiment has been performed. The results are shown in table 27. LogP was swapped between two files to see how that affects the models. For first modelling using test sets the results have improved for IBK in Pallas files and for ACD two-algorithm J48 and JRip shows improvement. The other algorithm doesn't show improvement. For 10-fold Cross Validation for Pallas the accuracy increases using CVR and LMT and for ACD using LMT algorithm.

Table28: Algorithms accuracy after descriptor swap in datasets (Pallas) for Daphnia endpoint

Endpoint Daphnia	Algorithm accuracy against test set (%)							
Pallas	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
Original Model	42.5	47.5	40	40	42.5	45	40	42.5
LogP swap	40	52.5	40	45	40	37.5	40	40
LogP & LogDpH3 swap	42.5	55	40	42.5	40	45	40	47.5
LogP,LogDpH3 & LogDpH5 swap	45	55	45	52.5	40	42.5	40	37.5

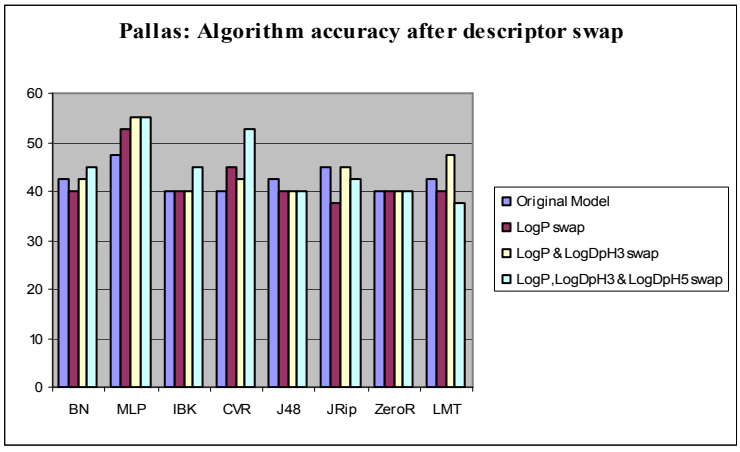


Figure46: Algorithms accuracy after descriptor swap in datasets (Pallas) for Daphnia endpoint

Table29: Algorithms accuracy after descriptor swap in datasets (ACD) for Daphnia endpoint

Endpoint Daphnia	Algorithm accuracy against test set (%)							
ACD	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
Original Model	47.5	50	35	45	40	37.5	40	45
LogP swap	47.5	47.5	40	42.5	40	42.5	40	45
LogP & LogDpH3 swap	47.5	45	45	40	40	40	40	45
LogP,LogDpH3 & LogDpH5 swap	47.5	45	45	40	45	40	40	45

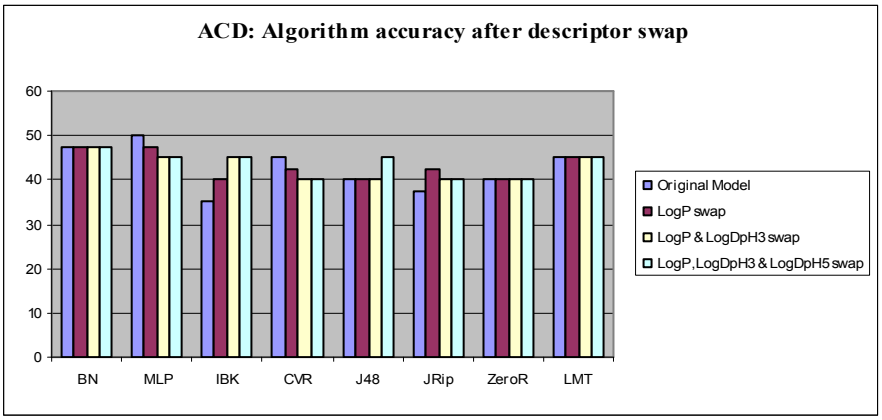


Figure47: Algorithms accuracy after descriptor swap in datasets (ACD) for Daphnia endpoint

Table30: Algorithms accuracy after descriptor swap in datasets (Pallas) for Daphnia endpoint using 10-fold Cross Validation

Endpoint Daphnia	Algorithm accuracy tested by 10-fold Cross Validation (%)							
Pallas	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
Original Model	44.1176	47.549	38.2353	43.1373	48.5294	43.1373	44.6078	45.5882
LogP swap	46.0784	47.0588	38.7255	44.6078	44.1176	46.0784	44.6078	45.5882
LogP & LogDpH3 swap	48.5294	48.5294	39.2157	46.0784	47.0588	50.4902	44.6078	47.0588
LogP,LogDpH3 & LogDpH5 swap	47.0588	48.0392	36.7647	48.0392	49.5098	48.0392	44.6078	48.5294

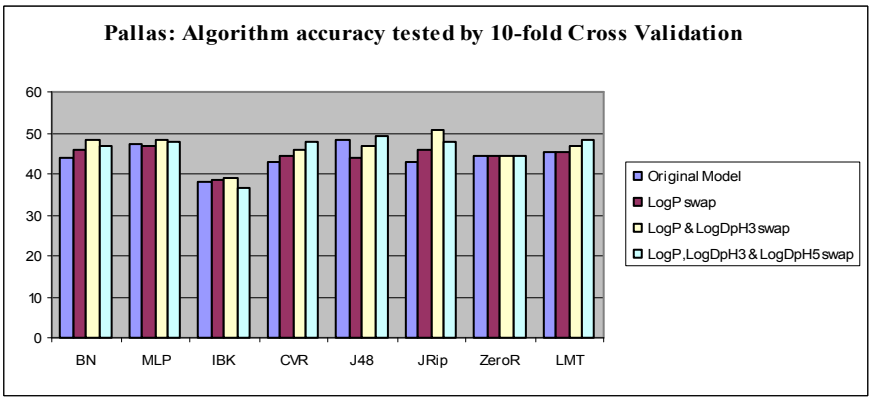


Figure48: Algorithms accuracy after descriptor swap in datasets (Pallas) for Daphnia endpoint using 10-fold Cross Validation

Table31: Algorithms accuracy after descriptor swap in datasets (ACD) for Daphnia endpoint using 10-fold Cross Validation

Endpoint Daphnia	Algorithm accuracy tested by 10-fold Cross Validation (%)							
ACD	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
Original Model	42.6471	47.0588	42.1569	50.4902	48.0392	48.5294	44.6078	46.0784
LogP swap	43.1373	44.1176	39.7059	44.1176	45.098	47.0588	44.6078	46.0784
LogP & LogDpH3 swap	45.5882	47.0588	37.2549	46.5686	45.5882	46.0784	44.6078	48.0392
LogP,LogDpH3 & LogDpH5 swap	45.098	45.5882	37.2549	49.0196	45.098	48.0392	44.6078	47.0588

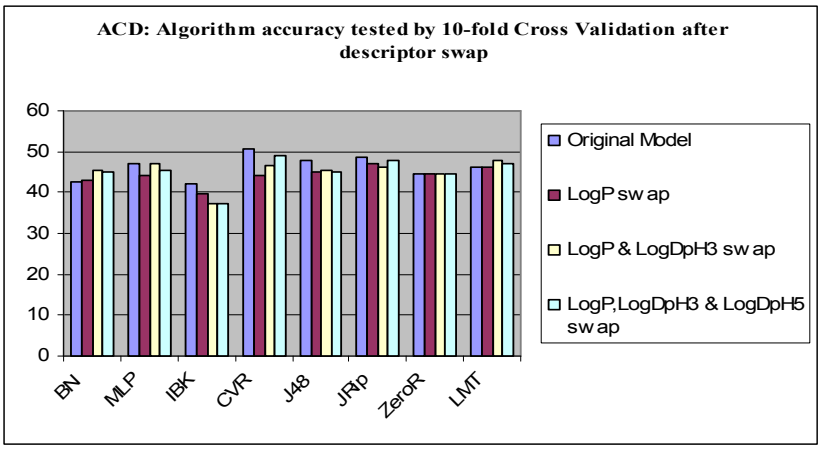


Figure49: Algorithms accuracy after descriptor swap in datasets (ACD) for Daphnia endpoint using 10-fold Cross Validation

Analysis: for this endpoint the descriptors which had highest value difference in both files were selected. As it shown in table 28, for modelling using test set the results are improved for Pallas files, especially for CVR, MLP and BN algorithm but for ACD the results in table 29 show improvement for IBK, J48 and JRip algorithm. The results are graphically represented in figures47 and 48. For modelling using 10-fold Cross Validation the results in tables 30 show massive improvements for Pallas for all algorithms especially the correlation between LogP and LogDpH3 with other descriptors values is considerable but results for ACD files in table 31 does not improve except for BN algorithm. The results are shown graphically in figure 49.

Table32: Algorithms accuracy after descriptor swap for LogP in ACD and Pallas datasets for OQ endpoint

Endpoint	Algorithm accuracy against test set (%)							
OralQuail	**BN**	**MLP**	**IBK**	**CVR**	**J48**	**JRip**	**ZeroR**	**LMT**
(Pallas)-Before	44.4444	44.4444	33.3333	44.4444	44.4444	44.4444	44.4444	44.4444
(Pallas)-After	44.4444	44.4444	38.8889	44.4444	44.4444	44.4444	44.4444	44.4444
(ACD)-Before	44.4444	44.4444	50	44.4444	44.4444	44.4444	44.4444	44.4444
(ACD)-After	44.4444	44.4444	44.4444	44.4444	44.4444	44.4444	44.4444	44.4444
Endpoint	**Algorithm accuracy tested by 10-fold Cross Validation (%)**							
OralQuail	**BN**	**MLP**	**IBK**	**CVR**	**J48**	**JRip**	**ZeroR**	**LMT**
(Pallas)-Before	58.1395	53.4884	44.186	54.6512	55.814	58.1395	58.1395	56.9767
(Pallas)-After	58.1395	53.4884	43.0233	58.1395	56.9767	58.1395	58.1395	56.9767
(ACD)-Before	58.1395	55.814	34.8837	58.1395	58.1395	58.1395	58.1395	58.1395
(ACD)-After	58.1395	54.6512	43.0233	54.6512	58.1395	58.1395	58.1395	56.9767

As the result show in table 32, since the value difference between descriptors in two files is not considerable just one swap for LogP has been done for this endpoint. The results have been improved in some algorithms (J48) for Pallas using 10-fold Cross Validation and also IBK for ACD. In modelling using test sets the results are same and in the case of IBK algorithm for ACD is even worse. Figures50 and 51 show the results in graphical form.

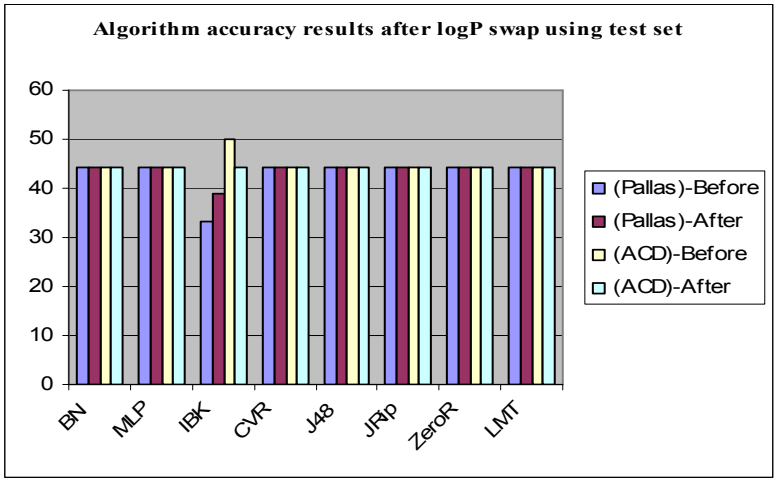


Figure50: Algorithms accuracy after descriptor swap for LogP in datasets (ACD and Pallas) for OQ endpoint

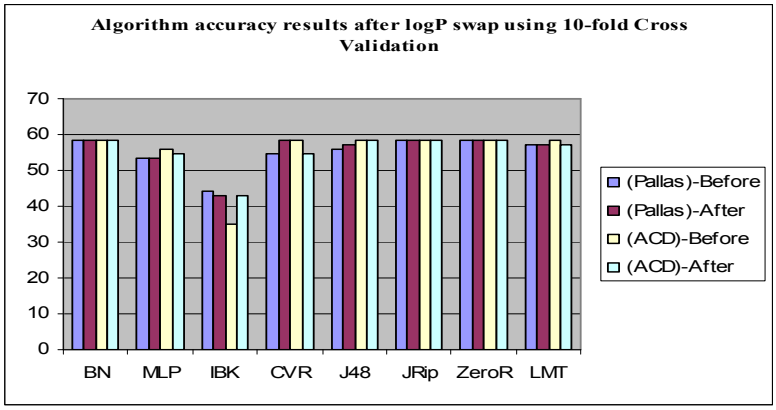


Figure51: Algorithms accuracy after descriptor swap for LogP in datasets (ACD and Pallas) for OQ endpoint using 10-fold Cross Validation

Table33: Algorithms accuracy after descriptor swap for LogP in datasets (ACD and Pallas) for DQ endpoint

Endpoint	**Algorithm accuracy against test set (%)**							
DietryQuail	**BN**	**MLP**	**IBK**	**CVR**	**J48**	**JRip**	**ZeroR**	**LMT**
(Pallas)-Before	33.3333	38.8889	38.8889	16.6667	27.7778	33.3333	33.3333	50
(Pallas)-After	33.3333	38.8889	38.8889	27.7778	22.2222	33.3333	33.3333	44.4444
(ACD)-Before	33.333	50	50	22.2222	33.3333	33.3333	33.3333	44.4444
(ACD)-After	33.3333	33.3333	33.3333	16.6667	38.8889	27.7778	33.3333	38.8889
Endpoint	**Algorithm accuracy tested by 10-fold Cross Validation (%)**							
DietryQuail	**BN**	**MLP**	**IBK**	**CVR**	**J48**	**JRip**	**ZeroR**	**LMT**
(Pallas)-Before	30.3371	25.8427	31.4607	34.8315	32.5843	29.2135	32.5843	32.5843
(Pallas)-After	30.3371	37.0787	32.5843	30.3371	32.5843	24.7191	32.5843	30.3371
(ACD)-Before	31.4607	33.7079	25.8427	29.2135	31.4607	29.2135	32.5843	28.0899
(ACD)-After	31.4607	29.2135	29.2135	29.2135	26.9663	31.4607	32.5843	26.9663

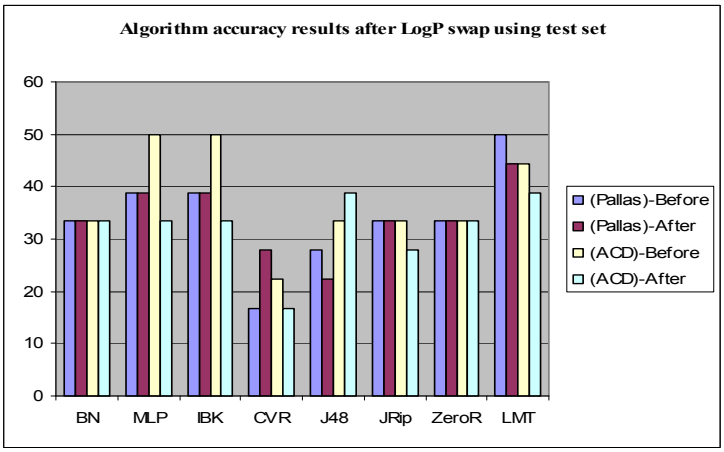


Figure52: Algorithms accuracy after descriptor swap for LogP in datasets (ACD and Pallas) for DQ endpoint

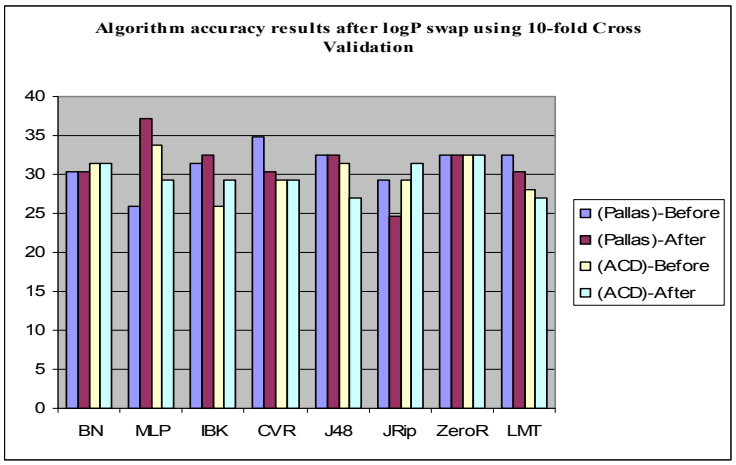


Figure53: Algorithms accuracy after descriptor swap for LogP in datasets (ACD and Pallas) for DQ endpoint using 10-fold Cross Validation

For this endpoint the results in table 33 don't show much improvement, although in number of cases for ACD files using Cross Validation (IBK, JRip) there is an increase. This is because the value difference produce by two programs is not considerable. Figures52 and 53 show the results graphically.

4.2.6 Adding Artificial Data (using average- first time)

For this task artificial data was added to each file. Between each row the average values of descriptors in that row was added to make a new row and new compound. Then new file was trained twice. For each file (ACD and PALLAS) we have two versions. One version is original file with no artificial data added and another version with artificial data added. For instance in table 34 they have been recorded as T_P original mean the file with no added artificial data for Trout endpoint produced by PALLAS. T_P artificial means the file with added artificial data for Trout endpoint produced by PALLAS.

Table34: Algorithms accuracy after adding artificial data to all the datasets (ACD and Pallas) to all endpoints

Endpoints	Algorithm accuracy against test set (%)							
	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
T_P original	56.5217	54.3478	36.9565	56.5217	54.3478	47.8261	43.4783	52.1739
T_P artificial	59.4203	56.5217	39.1304	47.8261	42.029	39.1304	43.4783	37.6812
T_A original	63.0435	65.2174	47.8261	63.0435	56.5217	58.6957	43.4783	60.8696
T_A artificial	60.8696	65.2174	44.9275	52.1739	42.029	36.2319	43.4783	40.5797
D_P original	42.5	47.5	40	40	42.5	45	40	42.5
D_P artificial	38.3333	45	38.3333	26.6667	38.3333	31.6667	40	30
D_A original	47.5	50	35	45	40	37.5	40	45
D_A artificial	53.3333	53.3333	45	46.6667	40	30	40	38.3333
B_P original	31.25	37.5	18.75	37.5	37.5	37.5	31.25	37.5
B_P artificial	25	25	41.6667	33.3333	29.1667	29.1667	29.1667	25
B_A original	31.25	37.5	31.25	31.25	25	31.25	31.25	37.5
B_A artificial	25	29.1667	25	29.1667	16.6667	25	29.1667	12.5
OQ_P original	44.4444	44.4444	33.3333	44.4444	44.4444	44.4444	44.4444	44.4444
OQ_P artificial	44.4444	44.4444	44.4444	44.4444	44.4444	44.4444	44.4444	48.1481
OQ_A original	44.4444	44.4444	50	44.4444	44.4444	44.4444	44.4444	44.4444
OQ_A artificial	44.4444	44.4444	51.8519	44.4444	44.4444	44.4444	44.4444	44.4444
DQ_P original	33.3333	38.8889	38.8889	16.6667	27.7778	33.3333	33.3333	50
DQ_P artificial	25.9259	37.037	37.037	25.9259	25.9259	25.9259	33.3333	29.6296
DQ_ A original	33.333	50	50	22.2222	33.3333	33.3333	33.3333	44.4444
DQ_A artificial	22.2222	48.1481	33.3333	25.9259	29.6296	25.9259	33.3333	29.6296

Table35: Algorithms accuracy after adding artificial data to all the datasets (ACD and Pallas) to all endpoints using 10-fold Cross Validation

Endpoints	Algorithm accuracy tested by 10-fold Cross Validation (%)							
	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
Trout (Pallas)	57.4074	50.463	43.0556	54.6296	52.3148	50.9259	44.9074	50
Trout (Pallas) artificial	62.6543	54.9383	48.4568	63.8889	66.358	58.642	45.0617	64.5062
Trout (ACD)	56.4815	49.0741	51.3889	54.6296	51.3889	56.0185	44.9074	51.8519
Trout (ACD) artificial	63.8889	49.3827	55.8642	59.8765	65.7407	63.8889	45.0617	66.0494
Daphnia (Pallas)	44.1176	47.549	38.2353	43.1373	48.5294	43.1373	44.6078	45.5882
Daphnia (Pallas)artificial	61.7647	51.634	45.4248	55.8824	61.1111	59.8039	44.4444	58.8235
Daphnia(ACD)	42.6471	47.0588	42.1569	50.4902	48.0392	48.5294	44.6078	46.0784
Daphnia(ACD)artificial	62.0915	50.9804	55.2288	60.7843	60.7843	63.3987	44.7712	63.3987
Bee(Pallas)	41.7722	34.1772	32.9114	40.5063	31.6456	41.7722	41.7722	36.7089
Bee(Pallas)artificial	47.4576	28.8136	36.4407	47.4576	42.3729	50.8475	41.5254	51.6949
Bee(ACD)	35.443	36.7089	39.2405	39.2405	35.443	44.3038	41.7722	39.2405
Bee(ACD)artificial	47.4576	35.5932	38.1356	51.6949	47.4576	44.0678	42.3729	41.5254
OralQuail(Pallas)	58.1395	53.4884	44.186	54.6512	55.814	58.1395	58.1395	56.9767
OralQuail(Pallas)artificial	64.3411	52.7132	41.8605	66.6667	68.9922	62.0155	58.1395	65.1163
OralQuail(ACD)	58.1395	55.814	34.8837	58.1395	58.1395	58.1395	58.1395	58.1395
OralQuail(ACD)artificial	64.3411	56.5891	38.7597	62.7907	68.2171	67.4419	58.1395	63.5659
DietryQuail(Pallas)	30.3371	25.8427	31.4607	34.8315	32.5843	29.2135	32.5843	32.5843
DietryQuail(Pallas)artificial	43.609	41.3534	30.0752	44.3609	50.3759	43.609	32.3308	40.6015
DietryQuail(ACD)	31.4607	33.7079	25.8427	29.2135	31.4607	29.2135	32.5843	28.0899
DietryQuail(ACD)artificial	45.1128	31.5789	29.3233	39.0977	42.1053	48.8722	32.3308	42.8571

Statistical Representation of the Results in This Chapter

Findings from all the experiments show following results:

Table36: Proportion of missing values in all the datasets (ACD and Pallas) in all endpoints

	ACD (%)					Pallas (%)				
	T	D	B	OQ	DQ	T	D	B	OQ	DQ
Rows with missing values	5.3	6.0	4.3	8.6	8.9	2.1	1.8	1.8	1.7	4.3
Rows with disguised missing value	30.8	22.7	23.9	25.8	42.2	51	45.4	31.6	48.2	60.1

Table 37 shows the increase of classification accuracy (%) after doing tasks compare to the original modelling when files where in their original format training data against test set. Table 39 shows the results using 10-fold Cross Validation. As it shown in some cases the classification accuracy has increased as a result of the performed tasks. For instance in table 37 the accuracy after modelling Pallas_LogP dataset (which is a dataset with LogP swapped) for BN algorithm has increased by 15.5% compare to the result from the original dataset modelled using the same algorithm. When there is a decrease in accuracy it is shown with minus sign in the table. In some cases there has been no difference which is shown by zero.

Table 38 shows the results for the same parameters in ACD files training data against test set. Table 39 shows the results when the files have been modelled using 10-fold Cross Validation.

Table37: Proportion of the classification accuracy results (%) for all endpoints after descriptor swap and adding artificial data (Pallas)

Swapping Descriptors & Adding Artificial data	Increased accuracy using training set against test set, Pallas (%)							
Trout	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
Pallas_LogP	15.5	4.34	12.8	13.69	7.8	5.9	3.4	15
Pallas_LogP & LogDpH5	9.02	0	23.69	-5.86	14.3	23.3	3.4	-4.3
Pallas_LogP, LogDpH5 & LogDpH7	9.02	6.5	12.8	5	10	16.8	3.4	2.8
Artificial data	-2.18	0	-2.9	-10.87	-14.5	-22.46	0	-20.29
Daphnia								
Pallas_LogP	0	-3.5	5	-2.5	0	5	0	0
Pallas_LogP & LogDpH5	0	-5	10	-5	0	2.5	0	0
Pallas_LogP, LogDpH5 & LogDpH7	0	-5	10	-5	5	2.5	0	0
Artificial data	5.83	3.33	10	1.66	0	-7.5	0	-6.67
Bee								
Pallas_LogP	0	0	-6.25	0	6.25	6.25	0	0
Artificial data	-6.25	-8.33	-6.25	-2.08	-8.34	-6.25	-2.08	-25
OralQuail								
Pallas_LogP	0	0	-5.56	0	0	0	0	0
Artificial data	0	0	1.85	0	0	0	0	0
DietaryQuail								
Pallas_LogP	0	-16.67	-16.67	-5.56	5.55	-5.55	0	-5.55
Artificial data	-11.11	-1.86	-16.67	3.7	-3.7	-7.4	0	-14.82

Table38: Proportion of the classification accuracy results (%) for all endpoints after descriptor swap and adding artificial data (ACD)

Swapping Descriptors & Adding Artificial data	Increased accuracy using training set against test set, ACD (%)							
Trout	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
ACD_LogP	16.1	6.8	-0.8	16.5	9.6	15	3.47	14
ACD_LogP & LogDpH5	16.1	-1.8	18.6	3.4	9.6	11.5	3.47	5.3
ACD_LogP,LogDpH5 & LogDpH7	14	9.0	3.4	16.5	16.1	2.8	3.47	14
Artificial data	2.9	2.18	2.18	-8.7	-12.32	-8.69	0	-14.49
Daphnia								
ACD_LogP	-2.5	5	0	5	-2.5	-7.5	0	-2.5
ACD_LogP & LogDpH5	0	7.5	0	2.5	-2.5	0	0	0
ACD_LogP, LogDpH5 & LogDpH7	2.5	7.5	5	12.5	-2.5	-2.5	0	2.5
Artificial data	-4.17	-2.5	-2.17	-13.34	-4.17	-13.34	0	-7.5
Bee								
ACD_LogP	0	-6.25	18.75	0	-6.25	-12.5	0	0
Artificial data	-6.25	-12.5	22.91	-4.17	-8.33	-8.33	-2.08	-12.5
OralQuail								
ACD_LogP	0	0	5.55	0	0	0	0	0
Artificial data	0	0	11.11	0	0	0	0	4.3
DietaryQuail								
ACD_LogP	0	0	0	11.11	-5.55	0	0	-5.55
Artificial data	-7.4	-1.85	-1.85	9.26	-1.85	-7.4	0	-20.38

Table39: Proportion of the classification accuracy results (%) for all endpoints after descriptor swap and adding artificial data (Pallas) using 10-fold Cross Validation

Swapping Descriptors & Adding Artificial data	Increased accuracy using 10-fold Cross Validation, Pallas (%)							
Trout	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
Pallas_LogP	13.37	6.6	5.5	5.4	6.59	6.57	0.3	7.17
Pallas_LogP & LogDpH5	2.94	-0.49	-3.43	-2.95	-2.4	-3.4	0	-2.45
Pallas_LogP, LogDpH5 & LogDpH7	15.69	8.5	4.6	5.99	5.67	3.3	0.30	7.17
Artificial data	7.4	0.3	4.48	5.2	14.3	7.8	0.16	14.1
Daphnia								
Pallas_LogP	0.49	-2.9	-2.4	-6.3	-2.9	-1.4	0	0
Pallas_LogP & LogDpH5	2.9	0	-4.9	-3.9	-2.4	-2.4	0	1.96
Pallas_LogP, LogDpH5 & LogDpH7	2.4	-1.47	-4.9	-1.4	-2.9	-0.4	0	0.98
Artificial data	19.44	3.9	13.07	10.29	12.7	14.8	0.16	17.3
Bee								
Pallas_LogP	0	-15.1	-3.7	0	-3.7	0	0	1.2
Artificial data	12.0	-1.1	-1.1	12.4	12.0	-0.2	0.6	2.2
OralQuail								
Pallas_LogP	0	-1.1	8.1	-3.4	0	0	0	-1.1
Artificial data	6.2	0.7	3.8	4.6	10.0	9.3	0	5.4
DietaryQuail								
Pallas_LogP	0	-4.4	3.3	0	-4.4	2.2	0	-1.1
Artificial data	13.6	-2.1	3.4	9.8	10.6	19.6	-0.2	14.7

Table40: Proportion of the classification accuracy results (%) for all endpoints after descriptor swap and adding artificial data (ACD) using 10-fold Cross Validation

Swapping Descriptors & Adding Artificial data	Increased accuracy using 10-fold Cross Validation, ACD (%)							
Trout	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
ACD_LogP	13.29	4.77	6.21	11.49	2.4	7.79	0.3	4.05
ACD_LogP & LogDpH5	2.45	1.96	-6.21	5.39	-3.43	3.92	0	4.42
ACD_LogP,LogDpH5 & LogDpH7	12.83	4.77	9.9	8.25	4.72	14.27	0.3	0.7
Artificial data	5.2	4.4	5.4	9.2	14	7.7	0.16	14.5
Daphnia								
ACD_LogP	1.9	-0.49	0.49	1.47	-4.4	2.9	0	0
ACD_LogP & LogDpH5	4.4	0.98	0.98	2.9	-1.4	7.3	0	1.4
ACD_LogP, LogDpH5 & LogDpH7	2.9	0.49	-1.47	4.9	0.98	4.9	0	2.9
Artificial data	17.6	4.0	7.1	12.7	12.5	16.6	-0.16	13.2
Bee								
ACD_LogP	0	-2.5	-1.2	1.2	-1.2	-1.2	0	3.79
Artificial data	5.6	-5.3	3.5	6.9	10.7	9.0	-0.2	14.9
OralQuail								
ACD_LogP	0	0	-1.1	3.4	1.16	0	0	0
Artificial data	6.2	-0.7	-2.3	12.0	13.1	3.8	0	8.1
DietaryQuail								
ACD_LogP	0	11.2	1.1	-4.4	0	-4.4	0	-2.2
Artificial data	13.2	15.5	-1.3	9.5	17.7	14.3	-0.2	8.0

Table41: Proportion of classes in each training dataset after adding artificial data first time

%	Class1	Class2	Class3	Class4	Class5
Trout	32.7	31.7	13.5	9.5	0
Daphnia	44.7	25.4	19.9	9.8	0
Bee	15.2	17.8	12.7	41.5	12.7
OralQuail	1.6	19.3	18.6	58.1	0
DietaryQuail	8.2	32.3	30.8	19.5	9.0

Analysis: as the results show (table 37, 38, 39 and 40)10-fold Cross Validation method increases the accuracy results more than using the training set against test set. Also three algorithms LMT, ZeroR and IBK have less increase of classification accuracy in all the models. As the result for following modelling these three algorithms are not going to be used. The proportion of the classes is different in each data set so it is important to see how changing these are going to affect the model. Next experiment will concentrate on this with adding artificial data on classes that have big proportion in the datasets. With adding more rows to the class that is big subset of the data we experience the increase in the classification accuracy. Trout and Daphnia respond with artificial data very well. There is an increase in classification accuracy for most algorithms used for training ACD and Pallas data sets. In ACD files for Bee endpoint there are three negative values for MLP, IBK and JRip and in Pallas files this applies to MLP and ZeroR. For OralQuail in ACD files are also one zero value for ZeroR algorithm and in Pallas files the values are negative for MLP and IBK and Zero for ZeroR. For DietaryQuail in ACD files there are two negative values

for MLP and ZeroR and in Pallas files this applies to IBK and ZeroR. So considering all the results if still the results for IBK, MLP and ZeroR algorithms are not satisfactory we will discard them in further modelling.

4.2.7 Adding Artificial Data (using average-second time)

For this task we need to add artificial data to all datasets the way we did before (calculation of the average value of top and bottom row and inserted empty row between and fill with average value). The models from previous task need to be looked at to see the proportion of the classification accuracy for negative and positive values applies to which classes then we can change the proportion of the classes in training data set accordingly to see the effect in the models. As it shows in the table 41(class proportion), for Trout and Daphnia the proportion of Class1 and Class2 compounds are much higher than Class3 and Class4. For Bee endpoint the proportion of Class 4 is very high and other classes are almost same. For OralQuail the proportion of class1 is very low and Class4 is very high. For DietaryQuail Class1 and Class5 proportions are very low and Class2 and Class3 have almost same proportion as Class1 and Class2 in Trout. The results of modelling after adding artificial data for the second time is shown in table 42.

Table42: Classification accuracy result after adding artificial data first and second time for all the endpoints

	Increased accuracy using 10-fold Cross Validation, (%)							
	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
T_P_artificial_1	5.2	4.4	5.4	9.2	14	7.7	0.16	14.5
T_P_artificial_2	8.02	4.52	12.39	10.56	17.29	17.05	0.10	17.51
T_A_artificial_1	7.4	0.3	4.48	5.2	14.3	7.8	0.16	14.1
T_A_artificial_2	14.05	1.97	2.43	10.56	19.37	7.32	0.10	18.21
D_P_artificial_1	17.6	4.0	7.1	12.7	12.5	16.6	-0.16	13.2
D_P_artificial_2	21.97	4.04	6.97	17.79	21.24	23.69	-0.13	19.52
D_A_artificial_1	19.44	3.9	13.07	10.29	12.7	14.8	0.16	17.3
D_A_artificial_2	23.44	3.55	12.14	17.07	17.31	17.31	0.10	19.76
B_P_artificial_1	5.6	-5.3	3.5	6.9	10.7	9.0	-0.2	14.9
B_P_artificial_2	6.63	-0.41	0.84	6.62	19.30	7.27	-0.37	9.1
B_A_artificial_1	12.0	-1.1	-1.1	12.4	12.0	-0.2	0.6	2.2
B_A_artificial_2	7.86	-1.04	-3.57	9.16	11.69	1.55	0.26	5.98
OQ_P_artificial_1	6.2	-0.7	-2.3	12.0	13.1	3.8	0	8.1
OQ_P_artificial_2	14.37	2.06	-9.09	14.35	16.70	9.11	-0.24	14.36
OQ_A_artificial_1	6.2	0.7	3.8	4.6	10.0	9.3	0	5.4
OQ_A_artificial_2	14.37	0.91	1.37	15.54	15.54	7.94	-0.24	14.95
DQ_P_artificial_1	13.2	15.5	-1.3	9.5	17.7	14.3	-0.2	8.0
DQ_P_artificial_2	19.38	9.18	-2.08	13.19	18.26	22.19	0.18	12.61
DQ_A_artificial_1	13.6	-2.1	3.4	9.8	10.6	19.6	-0.2	14.7
DQ_A_artificial_2	17.69	-2.06	-0.41	12.02	20.51	22.19	0.18	18.23

As it shown in table 42 the results for ZeroR and IBK algorithm are generally bad as before. For other algorithms Trout, Daphnia shows increase in accuracy after adding artificial data for second time but for Bee, results are worse in general so this dataset does not respond to artificial data. In this dataset the proportion of class2 and class4 is

very high probably that is the reason for decrease in prediction accuracy, which should be considered for further modelling. OralQuail dataset with very high Class4 proportion and very low Class1 percentage still responds to artificial data (except: IBK and ZeroR) and shows increase in prediction. DietaryQuail also shows better results (except for IBK) algorithm. So what we need to do is to adjust the proportions of the Classes for Bee endpoint (add artificial data just to Class1 and Class2) to see the effect on the prediction. The first logic is to reduce the Class4 proportion and increase Class1 and 2 proportions in the dataset (as it appears in Trout and Daphnia datasets and shows very high prediction accuracy). This time IBK and ZeroR algorithm wouldn't be used for modelling. Table 43 compare the proportion of the classes in files with artificial data added first time (first row for each dataset) and for second time (second row).

Table43: Proportion of the classes in datasets after adding artificial data first and second time for all the endpoints

%	Class1	Class2	Class3	Class4	Class5
Trout_1	32.7	31.7	13.5	9.5	0
Trout_2	45	32	13.5	9.5	0
Daphnia_1	44.5	25.7	20	9.58	0
Daphnia_2	44.7	25.4	19.9	9.8	0
Bee_1	15.2	17.8	12.7	41.5	12.7
Bee_2	15.2	17.8	12.7	41.4	12.7
OralQuail_1	1.6	19.3	18.6	58.1	0
OralQuail_2	3.5	19.8	18.7	57.8	0
DietaryQuail_1	8.2	32.3	30.8	19.5	9.0
DietaryQuail_2	7.9	32.7	30.5	19.7	9.0

Table44: Proportion of artificial data in datasets first and second time for all the endpoints

Proportion of artificial data %	Trout	Daphnia	Bee	OralQuail	DietaryQuail
Artificial_1	33	33	33	33	33
Artificial_2	49.8	49.8	49.6	49.7	49.7

Table45: The classification accuracy algorithms with highest performance (time) for all the experiments on datasets

BN	J48	JRip	CVR	LMT	MLP
6	4	3	3	2	1

Based on what is represented in table 45 the conclusion would be that the best models are BN and J48.Table 46 shows the results of global parameter comparison for all the endpoints. These results have been shown in previous sections in this chapter separately for each endpoint. This table also displays the ID of the chemical compound with the maximum value and minimum value for all the descriptors. It shows the minimum value difference between two calculated values by ACD and Pallas and also the maximum value difference for the same descriptor presented by two programs.

Pallas (Daphnia endpoint)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	-2.701	-6.54063	-6.54052	-7.85046	-7.89228	-9.10505
Max	11.6915	11.6915	11.6915	11.6915	11.6915	11.6915
Min Value Difference	-7.40214	-8.11613	-8.11602	-8.50406	-8.10928	-7.40214
Max Value Difference	2.21882	4.00517	3.19537	3.58602	3.80667	3.81636
ID of Min	*346*	*51*	*51*	*417*	*417*	*143*
ID of Max	*418*	*418*	*418*	*418*	*418*	*418*
ACD(Daphnia endpoint)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	-2.3559	-5.4966	-5.8715	-6.499	-6.6644	-6.8685
Max	13.676	13.676	13.676	13.676	13.676	13.676
Min Value Difference	-7.40214	-8.11613	-8.11602	-8.50406	-8.10928	-7.40214
Max Value Difference	2.21882	4.00517	3.19537	3.58602	3.80667	3.81636
ID of Min	*143*	*143*	*143*	*143*	*143*	*143*
ID of Max	*90*	*90*	*90*	*90*	*90*	*90*
Pallas(Bee endpoint)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	-0.952306	-3.78509	-4.75384	-5.6052	-5.99314	-7.5105
Max	8.16996	8.16996	8.16996	8.16996	8.16996	8.16996
Min Value Difference	-3.00158	-3.00158	-3.00158	-3.00158	-3.00158	-3.27919
Max Value Difference	2.21882	2.22028	2.21881	3.58602	3.80667	3.81636
ID of Min	*192*	*382*	*457*	*373*	*373*	*373*
ID of Max	*146*	*146*	*146*	*146*	*146*	*146*
ACD(Bee endpoint)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	-1.4202	-3.4969	-4.2068	-5.0751	-5.2504	-5.8599
Max	8.2665	8.1404	8.1404	8.1404	8.1404	8.1678
Min Value Difference	-3.00158	-3.00158	-3.00158	-3.00158	-3.00158	-3.27919
Max Value Difference	2.21882	2.22028	2.21881	3.58602	3.80667	3.81636
ID of Min	*373*	*382*	*382*	*373*	*373*	*373*
ID of Max	*146*	*248*	*248*	*248*	*248*	*146*
Pallas(Trout endpoint)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	-2.701	-6.54063	-6.54052	-6.5299	-6.51344	-9.10505
Max	8.68196	8.68196	8.68196	8.68196	8.68196	8.68196
Min Value Difference	-7.40214	-8.11613	-8.11602	-8.1055	-8.08934	-7.40214
Max Value Difference	2.69049	4.00517	3.19537	3.58602	3.80667	3.81636
ID of Min	*346*	*51*	*51*	*51*	*51*	*143*
ID of Max	*93*	*93*	*93*	*93*	*93*	*93*
ACD(Trout endpoint)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	-2.3559	-5.4966	-5.8715	-6.499	-6.6644	-6.8685
Max	13.676	13.676	13.676	13.676	13.676	13.676
Min Value Difference	-7.40214	-8.11613	-8.11602	-8.1055	-8.08934	-7.40214
Max Value Difference	2.69049	4.00517	3.19537	3.58602	3.80667	3.81636
ID of Min	*143*	*143*	*143*	*143*	*143*	*143*
ID of Max	*90*	*90*	*90*	*90*	*90*	*90*
Pallas(DietryQuail endpoint)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	-2.244	-2.36733	-2.51701	-3.92561	-3.99634	-4.04894
Max	8.16996	8.16996	8.16996	8.16996	8.16996	8.16996
Min Value Difference	-2.38923	-5.34828	-6.73168	-5.77022	-5.07963	-4.56813
Max Value Difference	2.47788	2.47788	2.47788	5.460421	5.677121	5.820521
ID of Min	*442*	*337*	*230*	*230*	*230*	*230*
ID of Max	*146*	*146*	*146*	*146*	*146*	*146*

ACD(DietryQuail endpoint)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	-1.4136	-1.4184	-2.5887	-5.6348	-5.8515	-5.9949
Max	8.5027	8.5012	8.5008	8.4633	8.4118	8.1678
Min Value Difference	-2.38923	-5.34828	-6.73168	-5.77022	-5.07963	-4.56813
Max Value Difference	2.47788	2.47788	2.47788	5.460421	5.677121	5.820521
ID of Min	*447*	*447*	*230*	*447*	*447*	*447*
ID of Max	*411*	*411*	*411*	*411*	*411*	*146*
Pallas(OralQuail)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	-2.82609	-6.54063	-6.54052	-6.5299	-7.05401	-7.78253
Max	8.16996	8.16996	8.16996	8.16996	8.16996	8.16996
Min Value Difference	-7.40214	-8.11613	-8.11602	-8.1055	-8.08934	-7.40214
Max Value Difference	2.47788	4.00517	3.19537	2.47788	2.47788	2.47788
ID of Min	*433*	*51*	*51*	*51*	*372*	*372*
ID of Max	*146*	*146*	*146*	*146*	*146*	*146*
ACD(OralQuail)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	-1.6559	-4.7181	-3.763	-4.6224	-5.0095	-5.2056
Max	13.676	13.676	13.676	13.676	13.676	13.676
Min Value Difference	-7.40214	-8.11613	-8.11602	-8.1055	-8.08934	-7.40214
Max Value Difference	2.47788	4.00517	3.19537	2.47788	2.47788	2.47788
ID of Min	*347*	*347*	*347*	*372*	*372*	*372*
ID of Max	*90*	*90*	*90*	*90*	*90*	*90*

Table46: Summary result of all statistical parameters from all the experiments

	Classification accuracy for training set against test set							
	BN	**MLP**	**IBK**	**CVR**	**J48**	**JRip**	**ZeroR**	**LMT**
Trout (Pallas)	56.52	54.35	36.96	56.52	54.35	47.83	43.48	52.17
Trout (ACD)	63.04	65.22	47.83	63.04	56.52	58.70	43.48	60.87
Daphnia (Pallas)	42.50	47.50	40.00	40.00	42.50	45.00	40.00	42.50
Daphnia(ACD)	47.50	50.00	35.00	45.00	40.00	37.50	40.00	45.00
Bee(Pallas)	31.25	37.50	18.75	37.50	37.50	37.50	31.25	37.50
Bee(ACD)	31.25	37.50	31.25	31.25	25.00	31.25	31.25	37.50
OralQuail(Pallas)	44.44	44.44	33.33	44.44	44.44	44.44	44.44	44.44
OralQuail(ACD)	44.44	44.44	50.00	44.44	44.44	44.44	44.44	44.44
DietryQuail(Pallas)	33.33	38.89	38.89	16.67	27.78	33.33	33.33	50.00
DietryQuail(ACD)	33.33	50.00	50.00	22.22	33.33	33.33	33.33	44.44
Endpoints	**Classification accuracy using 10-fold Cross Validation**							
	BN	**MLP**	**IBK**	**CVR**	**J48**	**Jrip**	**ZeroR**	**LMT**
Trout (Pallas)	53.05	51.91	42.37	54.20	48.86	50.76	44.66	50.38
Trout (ACD)	53.82	51.15	52.29	57.25	58.40	54.58	44.66	55.73
Daphnia (Pallas)	42.62	41.39	37.70	37.70	51.23	45.90	43.85	47.95
Daphnia (ACD)	45.90	44.67	41.39	51.64	43.85	46.72	43.85	48.36
Bee (Pallas)	40.00	32.63	30.53	44.21	32.63	40.00	40.00	40.00
Bee (ACD)	37.89	27.36	38.95	33.68	33.68	37.89	40.00	36.84
OralQuail (Pallas)	55.77	49.04	31.73	52.88	51.92	54.81	55.77	52.88
OralQuail (ACD)	55.77	52.88	34.62	50.00	53.85	55.77	55.77	54.81
DietryQuail (Pallas)	32.71	31.78	28.04	31.78	20.56	27.10	32.71	25.23
DietryQuail (ACD)	31.77	29.90	30.84	25.23	29.91	28.97	32.71	35.51

Table47: The summary result of classification accuracy on datasets from all previous experiments

Table 47 shows the result of the classification accuracy for all the endpoints in one table. The results have been shown before for each endpoint in separate tables.

4.2.8 Collective Summary Results

We found the following deficiencies in data files:

Check of Input Values: there were number of rows in which the values for all columns (descriptors) were identical for specific chemical compounds.

This might have happened as a result of a mistake in value generation by the software used due to the complexity of the calculation of the chemical compounds properties (ex: Trout data set). These values might be the default values for descriptors, which are generated when the exact measures for compounds attributes cannot be produced. For whatever reasons these values appear in the dataset, they need further consideration and study and they cannot be relied on.

We also found a contradiction between ID number and matching chemical specified by one program to another in the sense that the ID for the specific chemical was the same in both files but the matching name and CAS number were different. For example for endpoint Bee LD50, in the file with ACD descriptors, chemical compound with ID=450=Allethrin has been given CAS no: 584-79-2 but in the file produced by Pallas, ID=450=28434-00-6=s-bioallethrin, which in Toxnet comes with a different name for the same chemical having the same CAS: 284-79-2.

Moreover, a breach of the homogeneity rules was found: in the dataset for endpoint Trout, legend (descriptors definition) for Pallas is different from the other endpoints although for this work the descriptors were selected accordingly (ex: Pallas04=LogDpH7 but for other endpoints Pallas05=LogDpH7). Also the number of significant places that represent values in each column and for every row is different, which shows inconsistencies of data representation. We have presented the results of calculation for Min, Max values and their difference of the same descriptor for the same compound available in two data files related to the software used to calculate chemical descriptors and also showed the ID number of the chemical compound with the Min or Max value for the specific descriptor. What we found are significant differences between calculated values for the same descriptor presented by ACD and Pallas. In some cases, for example for endpoint Trout LC50, the maximum values for LogP are 8.6 (Pallas) and 13.6 (ACD) and for OralQuail LD50 are and 8.1 (Pallas)

and 13.6 (ACD). This is almost double from one to another and flags out a significant warning, since information provided for this descriptor identifies compound solubility in water and ability to cross cell membranes and is therefore of high importance for toxicity prediction models.

Model Performances: descriptor value differences also create doubts of reliability. This problem applies to all descriptors and for all endpoints in DEMETRA datasets. In this chapter the accuracy of models using various algorithms for classification is compared: values for the first experiment, which was model development based on training set using eight algorithms (see above) and validation against original testing set. The performance in general presents better results for data values generated by program ACD. Mean Square Error and Root Mean Absolute Error have been used to measure the errors of classification accuracy (not displayed here) are lower for models related to ACD data. This shows better correlations between ACD descriptors values and the toxicity output. Performance of the models has also improved with swapping descriptor between two dataset from ACD and Pallas.

Range Margins' IDs: ID numbers for chemical compounds defining Min and Max value for same descriptor were also considered. If a chemical compound with specific ID number has the Min value for a specific descriptor in one data file, the same chemical compound should possess the same parameter property for all source files. For instance for endpoint Trout, the ID of compound, which has the minimum value for LogP (Pallas) is 346 but generated by ACD is 143.

Min-Max value difference between two columns (value for the same descriptor, one generated by ACD and one by Pallas) in the same row considerably vary (ex: for Trout LC50 endpoint vary by up to 8.1 unsigned numerical value) [60].

4.3 A new algorithm for data quality assessment process

Based on the findings presented in previous chapter we have defined number of criteria and also a procedural framework in order to assess data quality in predictive toxicology.

4.3.1 Proposed Criteria for Data Quality in Predictive Toxicology

Figure54 shows values variation for LogP between the two programs (data for OralQuail LD50 endpoint). There are number of big peaks in the graph for values calculated by both programs which clearly identify the presence of outliers. As it

shown the values follow same pattern but in different proportion. This again depends on the computer program calculation default values setup, which is not the same in two programs.

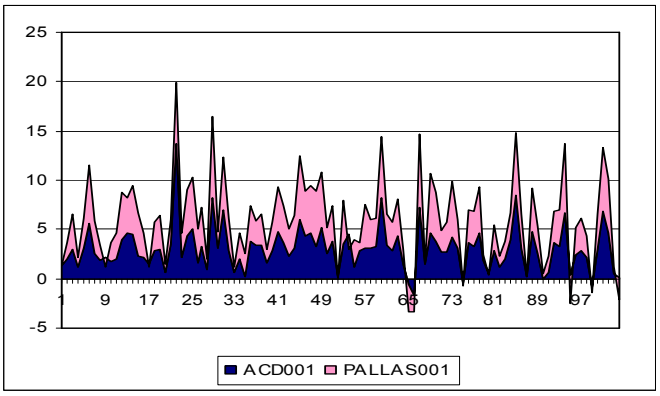


Figure54: Comparison of LogP variation values presented by ACD and Pallas for OralQuail.

Table48: Calculated variance for OralQuail

OralQuail	LogP	LogDpH3	LogDpH5
VARP (ACD)	4.83210	5.52960	5.62856
VARP(Pallas)	4.87138	7.21954	7.29682
	LogDpH7	LogDpH7.4	LogDpH9
VARP (ACD)	6.18880	6.26632	6.23918
VARP(Pallas)	7.51035	7.51433	7.37912

From this experiment we propose as property for data quality the definition domain for each variable (a value range for each descriptor) and decide that we just accept the values in this range and categorized the peaks outside the range as outliers so they could be studied separately. This bias could be proposed as metric for every descriptor considering the measurements of every descriptors confidence interval for each endpoint and acceptance of the values within this range. Later we need to define a method to describe how the outliers could be modelled separately and how we can combine these models with the results of the training the rest of the data.

Table 48 shows the variance VARP calculated for descriptor values obtained by using ACD and Pallas for OralQuail LD50 endpoint according to formula: $\frac{\sum (x - \bar{x})^2}{n}$ where x is a sample Mean and n is a sample size. The variance values are greater for values produced by Pallas, which shows bigger distribution with a negative impact on the model development. The descriptor variance qualifies as a meaningful property of the source values.

Noise in data identified by rows with the same values in each column could be another measurement for signalling wrong data inputs. These rows should be eliminated or recalculated. If they are produced by program, there should be further confidence and reliability issues in using that program.

A correlation of the margins (Min and Max values) for each descriptor as calculated by different software represents a quality flag variable as well. If these extreme values (generated by various sources) for each endpoint do not belong to the same compound, then that particular descriptor needs further study. This is especially requires further consideration for descriptors (i.e. LogP) that are likely to be included as inputs for models based on feature extraction algorithms.

Descriptor swap (LogP) increased the classification accuracy. This showed the change of input balanced the model, which also can be used in defining bias for descriptors min and max values.

Apart from these quality criteria proposed based on the DEMETRA data other quality issues that have been discussed in previous chapters in other domains such as: data source reputation, consistency and integrity could also be added to our framework which all depends on the users of the system and their preferences.

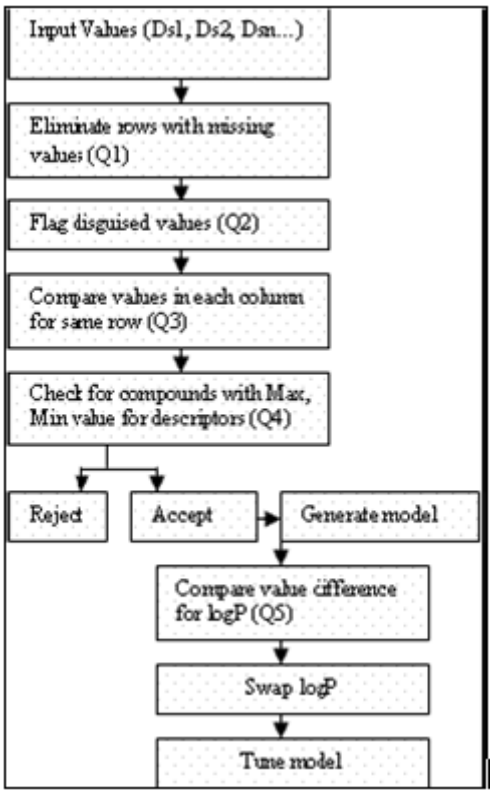


Figure55: Data quality assessment procedure

4.3.2 Quality Processing Flow Chart for Proposed Metrics

Based on empirical results obtained from studying the five toxicity datasets, we propose a data quality assessment process. Figure 68 shows this necessary process to prepare data for further modelling based on highlighted defects in out experimental work at this stage. Note that investigation was carried out on internal data and the proposed process has been based on discovered results.

4.3.3 A New Quality Assessment Algorithm for Data Quality

In Figure 69, we propose a quality check and assessment algorithm for the above procedure. The proposed algorithm could be improved and extended to provide further quality checks. At this stage the main aim was to direct our attention to first stage, error identification defects and propose possible ways of discovering and overcoming these in toxicology data.

Considering data quality parameters and criteria identified by our study and the experimental work presented above, some issues related to data quality have been highlighted, which indicate the need for a framework for quality assessment and measurements. The experimental work has identified some deficiencies related to data values and presentation. All highlighted data defects have direct effect on QSAR model performances, which are used for toxicity prediction of untested chemicals.

The importance of models requires use of high quality data [66] [67] [68].

Double data source

Input: Ds: Data Source, Ro: Result Output (data processed, ready for modelling), Rw: Instance(compound, row), Dc: Descriptor(column), result: Rs (final model), Quality Metrics : Q1=missing values in rows, Q2=column values are same in one row, Q3=values for the descriptor in each row is out of range Minv→Maxv, Q4=flag if Min and Max value for same descriptor in two files do not belong to the same compound, Q5=bias for value difference between same descriptor value for same compound in Ds1, Ds2, Dsn…Please note in our example data sources are ACD and Pallas.

Clean the data

```
//check for rows with missing values and eliminate
//check for rows with same value in each column and eliminate
//compare if value for each descriptor (column) and every row falls

Within bias (value range)
Start: SearchSheet (ACD & Pallas)
Foreach (Result as sheet→Ro)
For (i=0; i<count (Rw); i++)
If (Rw = (Q1))
Delete Rw else
For (j=0; j<count (Dc); j++)
If (Dc= (Q2))
Flag (error): "disguised data" else
If not Rw= (Q3) & Dc= (Q3)
Flag (error): "suspicious values" else
If not Rw= (Q4) & Dc= (Q4)
Flag (error): "Min, Max do not belong to same compound" else
Display Ro
End
```

Generate model

```
//check similar fields; if value difference for same descriptor
(logP) and for same compound is high, then train model, produce
result, swap logP, train again.
 Input: Ro + added new column which shows the difference between two
values (logPPallas-LogPACD=Dsw), LogP descriptor=DLogP
//generate model with cleaned data
Start: generate model (ACD, Pallas)
//swap logP and generate again
Foreach (Result as sheet→Rsw)
For (j=0; j<count (Dc); j++)
If (Rsw = Q5)
Swap (DLogP)
Display Rsw
Generate model
```

End

Figure56: Data quality assessment algorithm

4.4 Summary and Conclusions

In this chapter we have shown the results of our investigations on online toxicity databases and highlighted the inconsistencies in values presentation and also the structural differences from source to source. We have provided the results of our detailed investigation and experimental work on Demetra data. We studied two different file presented by two programs ACD and Pallas for same chemical compound and same species. The global values (min, max, average, standard deviation) have been measured and presented and also difference between values for same compounds generated by two programs highlighted. Based on these values we have selected number of descriptors to swap between two files for same species in order to see how variation in values could affect the model performance. The idea was to show that the good model is purely depend on how the descriptors have been generated and by what tool which directly affect the data quality. We have also tried to understand the data characteristics and insight view of the relations between descriptors. The results of this investigation have led us to identifying a general framework for data quality. We have provided the quality flow chart which shows the quality check of the data in steps with five identified criteria. An algorithm has also been proposed in details which explain how data is assessed and processed before modeling. These criteria are related purely to data values and can be added to other quality criteria which have been proposed in previous chapters to form a complete quality framework. Identifying these issues is great help in knowing our data before further modeling. One can assess data considering these criteria before training. Since reliability of the models purely depends on quality of the data, our algorithm shows how data can be validated before any models are constructed.

5. A NEW ALGORITHM FOR MISSING VALUE GENERATION IN TOXICITY DATASETS

For chemical compounds in toxicity databases, missing values problem is also a big issue. It might appear in two forms, absence of affect of chemical compounds toxicity on specific species and environment or missing values of chemical compounds properties and attributes. The reason behind the first issue apart from the experimental procedure has not been taken place in some environment nor on some species, is the idea of the effect not being detected even after the testing. To resolve this problem there are some guidelines that have been provided by the Environmental Protection Agency (EPA) [69]. In some datasets the toxicity values have been calculated but there may still be a large number of empty rows for chemical compounds attributes.
Since these datasets are used for data analysis and modelling using data mining and machine learning tools for prediction of toxicity of untested chemicals generating QSAR (Quantity Structure Activity Relationship)[70] models, it is very important to find an efficient way to overcome the problem of missing values. For the first problem (missing toxicity values) we present below the approach suggested by EPA. For the second type of missing values we propose a framework in the coming sections.

5.1 Toxicology Approach for Missing Values at the Collection Stage (EPA)

The Environmental Protection Agency has developed the data quality assessment process as an important tool for environmental scientist in order to assure the quality, quantity, collection and analysis of the environmental data has been satisfied. It provides solutions to overcome the problem of missing values relate to values below detection limits for chemical analysis. These are the cases where measurement data are described as not detected; the concentration of the chemical is unknown although it is between zero and the detection limit (DL). For this problem the following guidance table has been produced [69].

Table49: Missing values percentage categories

Percentage of non- detects	Statistical analysis method
<15%	Replace non-detects with DL/2 or a very small number
15%-50%	Cohen's adjustment, Trimmed mean, Winsorized mean and standard deviation
>50%-90%	Use test for proportions

A) Less than 15% non-detects-substitution methods: if there are small portion of the data is missing, we can replace it with a small number, usually the detection limit divided by 2.

B.1) Between 15-50% non-detects: Cohen's method provides adjusted estimates of the sample mean and standard deviation that is used for data below the detection level. This method is based on statistical techniques of maximum possibility estimation of the mean and variance so the prediction for values for non-detects may not be zero.

B.2) Trimmed mean: this method discards the data in the tails of a dataset to develop an unbiased estimate of the population mean. For environmental data, missing values normally occur in the left tail of the data, so trimming the data can account for estimation of the mean value.

B.3) Winsorized mean and standard deviation: this method replaces data in the tails of a dataset with the next most extreme value. This also can adjust the dataset for non-detect value, which would help for calculation of mean and standard deviation parameters.

C) Greater than 50% non-detects-test of proportion: if more than 50% of the data are below the detection limit but at least 10% of the observations are quantified, tests of proportion may be used to test hypotheses using the data.

All of these methods help to find estimation for mean and standard deviation values when part of the data is missing and the exact parameters cannot be calculated. In the environmental testing these parameters are the most important statistical elements that need to be identified.

But even with these methods, still the second issue of recovering missing values remains unresolved. In the followings sections we propose a framework to recover missing values and the implementation of the defined methods in this framework on a number of toxicity datasets produced by two different applications with the help of statistical methods and measurements. Also we show how the generation of artificial data with the use of this framework can affect model performance.

5.2 Data Recovery: Proposed Framework

This framework uses objective measures based on statistical strengths or properties of data [71] to recover missing values in two circumstances. First when there are two versions of same datasets provided for the data with the same parameters and second when there is just one single dataset to be considered with missing values appearing for number of attributes. In the case of existence of multiple (historical or prototype) versions of the dataset, the missing values appear for all attributes for a specific field but in the case of considering just one version of the dataset, missing values might appear for just a number of attributes but not all. In worst case scenario we produce a method to recover data when for a field just one attribute value has been presented and the rest are missing.

If we consider the data in the form of M*N table with Mk columns and Nk rows then the missing values in datasets appear in the following form:

Table50. The structure of missing values in a dataset

		M_1	M_2	…	M_k
	N_1	m_1n_1	m_2n_1	…	m_kn_1
First case →	N_2	?	?	?	?
Second case →	N_3	m_1n_3	?	?	?
	:	:	:	:	:
	N_k	m_1n_k	m_2n_k	…	m_kn_k

5.2.1 Paired Datasets (Least Square Method)

Given two datasets (MN, AB) for the same data, which contain the same attributes, but with two sets of values (i.e.: depend on data generators applications or historical sources), we can find the relationship between column M from the file MN and column A from the file AB (provided they represent the same attribute).

$M_{1:\text{ correlated \_with: }}A_1$, $M_{2:\text{ correlated \_with: }}A_2$ …… $M_{k:\text{ correlated with : }}A_k$

Table51: The structure of missing values in multiple versions of the same dataset

	M_1	M_2	…	M_k
N_1	m_1n_1	m_2n_1	…	m_kn_1
N_2	m_1n_2	m_2n_2	…	m_kn_2
N_3	?	?	?	?

	A_1	A_2	…	A_k
B_1	a_1b_1	a_2b_1	…	a_kb_1
B_2	?	?	?	?
B_3	a_1b_3	a_2b_3	…	a_kb_3

In this case we have a number of rows with missing values in each file. Considering our dataset with its special characteristics, if there exists strong relationships globally and locally between attributes, with the use of Least Square Method for regression line [72] we can calculate the missing values based on the following formula (considering the straight-line model):

5.1	$y = \beta_0 + \beta_1 x + \varepsilon$

The least square method involves the determination of β_0, β_1 to minimize ***Q*** and they are treated as the variables in the optimization and the predictor variable values, x_1, x_2, .. , x_n are treated as coefficients.

For this model the least squares estimations of the parameters are computed by:

5.2	$Q = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$
5.3	$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

5.2.2 Single Dataset (First Serial Correlation)

This scenario is for when there is just one version of the dataset and the value for at least one attribute has been provided so we can calculate the value for the whole row of data based on the first value. In this situation we need to examine the correlation between descriptors themselves with the use of First Serial Correlation [72], which investigates the dependencies of the variables.

We consider the relationship between attributes as follows:

$$(M_1, M_2), (M_2, M_3) \ldots (M_{k-1}, M_k)$$

$$M_{1:\text{ correlated \_with:}} M_2, \quad M_{2:\text{ correlated with:}} M_3 \ \ldots\ldots \ M_{k-1:\text{ correlated \_with :}} M_k$$

Table52: The structure of the missing values in a single version of the dataset

	M_1	M_2	…	M_k
N_1	m_1n_1	m_2n_1	…	m_kn_1
N_2	m_1n_2	?	?	?
:	:	:	:	:

5.3 Experimental Work

Extensive experimental work has been carried out in order to examine the framework for a number of toxicity datasets to recover missing values. With the consideration of the datasets possessing the special characteristics requirements, these methods have been tested and the results are as follows.

5.3.1 Background

Given the current facilities available for complex calculations, it seems that high confidence is implicitly awarded to data downloaded from online resources. The same applies to data generated by specialist software. We used the opportunity to study the DEMETRA data sets on some issues on data quality for large databases. We started with identification of descriptors sharing the same name and duplicated as generated by various software used by research laboratories involved in the project.

Data on five toxicity endpoints are provided by the DEMETRA project for four different species: Bee, Daphnia, Trout, OralQuail and DietaryQuail. For each dataset, values for six compound descriptors calculated by two specialist programs: ACD and Pallas, have been considered. Our aim was to highlight the variation of values for each descriptor produced from one program to another and also to compare any further quantitative differences between specific descriptors calculated by one program with the value for the same descriptor and chemical compound generated by the other one. Then we compared the accuracy of basic classification model using input data presented for each endpoint by descriptors calculated by ACD and Pallas.

5.3.2 Data Preparation

For each dataset the same number of compounds has been selected. Data cleaning has also been performed in the form of eliminating rows with missing values. Six common descriptors have been selected from both datasets. These are as follows: LogP, LogDpH3, LogDpH5, LogDpH7, LogDpH7.4 and LogDpH9. The data have been divided into training set and testing set based on predefined rules (85% training, 15% testing) by DEMETRA project. Weka data mining tool has been used to develop models. The data format has also been changed for modelling into Weka compatible format (arff). The conditions of experiments for each endpoint containing two datasets

(i.e. the same running parameters for algorithms, identical host machine etc.) were identical in order to assure an accurate comparison.

5.4 Methods Implementation for Toxicity Datasets

For all the experiments the data has been cleaned in the mean of omitting empty rows with missing values. Table 53 shows the proportion of missing values in each dataset: T for Trout, D for Daphnia, B for Bee, OQ for OralQuail and DQ for DietaryQuail.
First row of the table shows the number of chemical compounds for each endpoint in each dataset. Second row presents number of compounds after the data cleaning. The other four rows show the proportion of lost data after cleaning and also the proportion of empty rows before cleaning.
As it shown in some cases for accuracy and identically of the experiments (all our experiments were based on value comparisons between ACD and Pallas files) for specific endpoint in both datasets (ACD, Pallas) the exact same chemicals had to be selected so the proportion of the lost data after cleaning is even more than the proportion before cleaning. For example for Trout endpoint in ACD file empty rows are 5.3% of the whole dataset originally but after cleaning this increases to 7.09%. As it shown in the table in some cases such as DietaryQuail, 8.9% of the data is missing in files produced by ACD application.

Table53: The proportion of missing values in each dataset after and before cleaning

	T	D	B	OQ	DQ
Number of original compounds	282	264	105	116	123
Number of compounds after cleaning	262	244	95	104	107
Lost data after cleaning (%)	7.09	7.5	10.5	10.3	13.0

Empty rows before cleaning (%)

ACD					Pallas				
T	D	B	OQ	DQ	T	D	B	OQ	DQ
5.3	6.0	6.6	8.6	8.9	2	2	3	1.7	4.3

5.4.1 Test of the Methods Requirements (Existence of the Relationship)

In toxicity datasets the missing values are in the form of whole row (first case) or in the case of ACD data sets only the value for LogP exists and other values are missing (second case). To test the relations between attributes, a number of tasks were performed using statistical tools. Firstly, correlation between one descriptor from one dataset with exact descriptor from the second dataset has been measured.
Figure57 shows the LogP value variation for DietaryQuail endpoint presented by ACD and Pallas. The yellow line on the graph shows the difference between the two

values. As the graph shows although the values are different but they follow the same pattern so there exists a correlation between them, which can be measured.

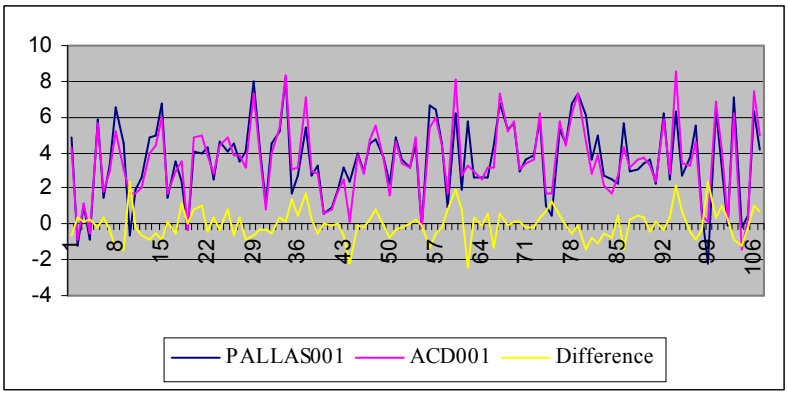


Figure57: LogP variation for DietaryQuail endpoint presented by ACD and Pallas

Figure58 shows the regression line for LogP value in two files (ACD, PALLAS) for four endpoints. Since there is strong relationship between two attributes we can calculate missing values if the value for same compound exists in either file, based on regression line equation. Table54 shows the statistical parameters measured for the relationship between two LogP calculated by ACD and Pallas for DietaryQuail endpoint. On the left side of table we produced the parameters based on X variable considered as LogP Pallas or PALLAS001 and Y variable as LogP ACD or ACD001.

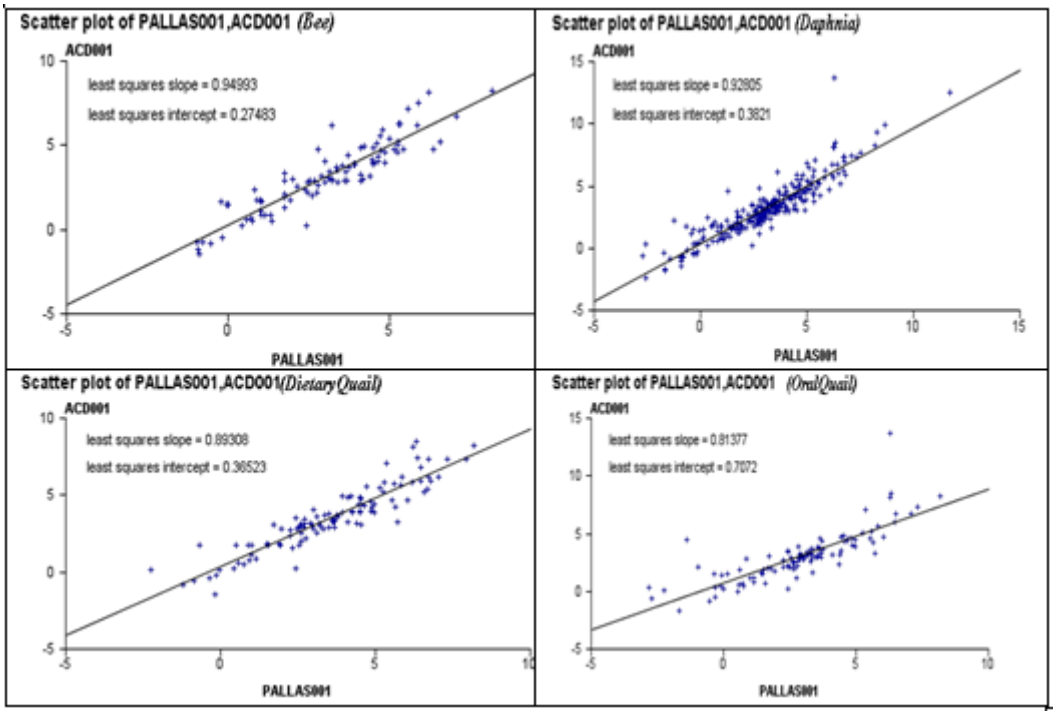


Figure58: Correlation between two LogP values for four endpoints

The regression line is also based on calculating dependent variable Y based on independent value X. On the right hand side of the table we calculated the regression based on ACD001 considered as X variable (in the cases when the value for an attribute LogP for a chemical compound is missing in dataset produced by PALLAS) and on the left the regression function is calculated based on PALLAS001 as X variable (in the cases when the value for same attribute is missing in the dataset produced by ACD). These functions have been produced by our statistical tool.

We should also note that in each case the value for X exist so the value for Y need to be calculated based on the function. Also the values for confidence intervals and standard deviation are very close.

Table54: The statistical parameters show the relationship between two LogP for DietaryQuail

Regression line x variable: PALLAS001 y variable: ACD001 The equation of the regression line of y on x is: y = 0.365234 + 0.893078x	Regression line x variable: ACD001 y variable: PALLAS001 The equation of the regression line of y on x is: Y = 0.190893 + 0.951051x
Confidence interval for PALLAS001 Mean 3.57315 Standard deviation 2.14349 95% confidence interval: 3.167 to 3.9793	Confidence interval for ACD001 Mean 3.55633 Standard deviation 2.07714 95% confidence interval: 3.16276 3.94991

Since the missing values in datasets appear in the whole row for a number of compounds, the same procedure can be repeated for all pair variables to fill the empty cells. As it has been specified in earlier sections the assumption is for every row of the missing values (Y parameter), related to a specific chemical compound there are existing values (X parameter) for the corresponding compound in the paired file (either ACD or PALLAS).

5.4.2 Recovering Missing Values (Multiple Datasets)

For this task based on regression function measured for every pair attribute (i.e.: LogP from ACD and LogP from PALLAS) and statistical analysis, explained in previous section, the missing values in each file for DietaryQuail endpoint have been replaced with calculated values (the same procedure has been repeated for other five pair attribute in both files). Recovered data then was trained using the same algorithms as in previous experiments in Weka. Table55 shows the results for this experiment for DietaryQuail.

Table55: The results for modelling original dataset (with omitted rows) and with recovered data using regression for DietaryQuail endpoint

	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT	Average	Increase (%)
DQ_P	30.34	25.84	31.46	34.83	32.58	29.21	32.58	32.58	31.18	2.55
DQ_P_R	26.42	33.96	32.08	41.51	35.85	35.85	28.30	35.85	33.73	
%diff	-3.92	8.12	0.61	6.68	3.26	6.64	-4.28	3.26		
DQ_A	31.46	33.71	25.84	29.21	31.46	29.21	32.58	28.09	30.20	1.53
DQ_A_R	29.25	34.91	35.85	32.08	29.25	29.25	29.25	33.96	31.72	
%diff	-2.22	1.20	10.01	2.86	-2.22	0.03	-3.34	5.87		

DQ_P shows the classification accuracy for dataset produced by Pallas program when the rows with missing values have been omitted. DQ_P_R: shows the results for the same experiment after data has been recovered. The value for the row "diff" shows the difference between the two results. Last three rows of the table show the results for files produced by ACD program. As it shown in the table, recovered data had dramatic affect on the results especially in the case of MLP and CVR algorithm for Pallas dataset and LMT and IBK algorithms for ACD dataset.

5.4.3 Recovering Missing Values for DietaryQuail Endpoint (Single Dataset)

In DietaryQuail dataset presented by ACD there were values for LogP descriptors and then the values for other descriptors were missing. These missing values have been estimated based on First Serial Correlation. The data has been trained afterwards using the same algorithms. The results of this experiment are listed in Table56.

Table56: The results for modelling original dataset (with omitted rows DQ_ACD) and with recovered data (DQ_LogP_corr_ACD) for DietaryQuail endpoint

Endpoint	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT	Average	Increase (%)
DQ _ACD	31.77	29.90	30.84	25.23	29.91	28.97	32.71	35.51	30.61	
DQ_LogP_corr_ACD	29.52	32.38	31.43	26.67	22.86	21.90	29.52	31.43	28.21	-2.39

5.5 Increasing the Model Performance with Generation of Artificial Data Using LSM Method

As shown in Table57, the performance of models (in the case of single dataset) has improved with some algorithm and decrease with others. But in general models are not better with recovered data. In order to have better models we propose the generation of artificial data. Since the correlation of descriptors is depending on the values variation produced by two source programs, the existence of outliers has a direct effect on the models. The results of second experiment on DietaryQuail in

Table56 shows that the methods have caused the effect on reducing the classification accuracy since the recovered values in some cases may belong to an outlier.

For this reason to balance the data we used the same method and procedure described in section 6.2 but this time we performed the task on the outliers which have been separated and used as training set. With Least Square Methods (1, 2, 3) we calculated the correlation between the descriptors (as we did for paired datasets). Then we generated artificial row (or compound) into the dataset.

With this method we balanced the data around outliers in both datasets (ACD and Pallas) in order to include the outliers into the regression. The experiment has been tested for three endpoints: Bee, OralQuail and Daphnia (see Table57 for the results).

Table57:The results for modelling original dataset and data with generated artificial values

Endpoint	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT	Average	Increase (%)
OralQuail (ACD)	55.77	52.88	34.62	50.00	53.85	55.77	55.77	54.81	51.68	
OQ_ACD_outlier+arti	57.76	51.72	47.41	56.03	56.90	57.76	57.76	56.90	55.28	3.60
OralQuail (Pallas)	55.77	49.04	31.73	52.88	51.92	54.81	55.77	52.88	50.60	
OQ_P_outlier+arti	57.76	52.59	37.07	56.03	55.17	57.76	57.76	56.90	53.88	3.28
Daphnia (ACD)	45.90	44.67	41.39	51.64	43.85	46.72	43.85	48.36	45.80	
D_ACD_outlier2+arti	44.15	52.45	43.02	49.06	42.26	49.43	44.91	46.79	46.51	0.71
Daphnia (Pallas)	42.62	41.39	37.70	37.70	51.23	45.90	43.85	47.95	43.55	
D_P_outlier+arti	44.53	48.30	39.62	45.66	46.79	46.42	44.91	43.77	45.00	1.45
Bee (ACD)	37.89	27.36	38.95	33.68	33.68	37.89	40.00	36.84	35.79	
B_ACD_outlier+arti	37.50	34.62	50.00	35.58	34.62	40.38	42.31	38.46	39.18	3.39
Bee (Pallas)	40.00	32.63	30.53	44.21	32.63	40.00	40.00	40.00	37.50	
B_P_outlier+arti	40.38	34.62	30.77	41.35	29.81	41.35	42.31	44.23	38.10	0.60

5.6 Algorithm for Generation of Missing Values

As a result of our study and special characteristics of our data, the following procedure can be proposed for recovery of the missing values:

Single dataset MN : M*N(if M_1: exist)

```
1.Read file (MN)
2.Calculate correlation using (formula: 1,2,3)
3.If correlation false exit, else
4.Sort data, based on toxicity (high →low)
5.Locate missing value
6.Locate first value in first cell
7.Consider value as X
8.Using the function calculated in step 2, calculate Y
9.Fill the value for next descriptor by calculated value (Y)
10.Repeat step 6, 7, 8, 9 until row is recovered
11.Repeat step 5, 6, 7, 8, 9
12.Repeat until 5 is empty
13.End
```

Paired dataset (MN: M*N & AB: A*B)

```
1.Read files (MN & AB)
2.Sort data based on toxicity (high →low) in both files
```

3.Calculate correlation using (formula: 1,2 & 3): If M1: correlated _with: A1, M2: correlated _with: A2 …… Mk: correlated_with: Ak True then
4.Locate missing value (cell) in MN, When $Y=Y_{MN}$ & $X=X_{AB}$
5.Calculate mn when $Y=Y_{MN}$
6.Calculate $Mn+1 \rightarrow Mk+1$ when $Y=Y_{MN}$
7.Do until data is estimated in entire row
8.Locate the next row with missing values
9.Repeat steps 4-8 until step 8 is empty
10.Calculate correlation using (formula: 1,2 & 3): If A1: correlated _with: M1, A2: correlated _with: M2 … Ak: correlated_with: Mk True then
11.Locate missing value (cell) in AB, When $Y=Y_{AB}$& $X=X_{MN}$
12.Calculate ab when $Y=Y_{AB}$
13.Calculate $Ab+1 \rightarrow Ak+1$ when $Y=Y_{AB}$
14.Do until data is estimated in entire row
15.Locate the next row with missing values
16.Repeat steps 11-15 until step 15 is empty
17.End

Figure59: Missing Values Recovery algorithm

5.7 Summary and Conclusions

In this chapter, the problem of missing values in datasets has been addressed specially in toxicology domain. The toxicology approach have been discussed that deal with the problem at the collection stage. Least Square Method for paired dataset and Serial Correlation for single dataset provided the solution for the problem in two different situations. An algorithm using these two methods has been proposed in order to overcome the problem of missing values. The proposed algorithm has been tested on number of DEMETR datasets to test the effectiveness on the outcome model after recovery of missing values. Also the Least Square Method has been used to generate artificial data around outliers to improve model performance. The implementation of the proposed algorithm requires the existence of the high correlation between descriptors (attributes) in the dataset.

6. ARTIFICIAL DATA GENERATION, DATA CHARACTERISTICS AND MODEL PERFORMANCE

Improving the learner performance over imbalanced and multidimensional datasets raises a challenging task for the machine learning community as well as for the data user. Although a salient characteristic in data modelling is the amount of data provided for the learner, the proportional distribution of that data in each class has also direct relationship with the classifier performance. In imbalanced datasets when data is distributed into different classes, various in size, understanding of data structure and characteristics plays an important role in improving the learner accuracy.

In this chapter we introduce a new approach that combines the information gained from traditional classification algorithms, confusion matrix parameters and density-based clustering to generate artificial data in order to increase the learner performance. First a classification algorithm is run on training data. Then the confusion matrix is studied and the True Positive (TP) rate of each class is measured. The class with the lowest TP rate is selected. Using density-based clustering we identify the centroid of the class and measure the samples distribution in multidimensional space in the next step. With the values gained from Probability Density Function estimations for clusters, extra samples are generated and added to the original dataset to rebalance the class proportion and the weight of different classes in the whole training set. Our method has been evaluated in terms of TP, F-Measure and also overall accuracy against a number of Demetra and UCI datasets. We also report evaluation of the performance of other classifiers, trained on the expanded datasets (datasets with added artificial data using our method) at the later stage. Our method provides an insight view of the data structure and characteristics in order to identify how much and where the data need to be added for increasing the classification accuracy of the learner.

6.1 Introduction

In data mining, classification learning is a supervised learning scheme that uses knowledge gained through the training process of classified instances for classification of unseen examples. One of the main issues for classifier during this

process is the samples distribution of classes or class balance. Imbalanced or skewed [51] dataset, affect the performance of classification algorithms. The over represented classes provide enough information for training the classifier because of their sufficient number of samples against the under represented class. Real world scientific applications often face this problem for a number of reasons [52].

For instance, in toxicology domain this problem is severe. When the chemical compounds need to be tested on different species, high toxicity chemicals cannot be sampled as many as low toxicity compounds. In these datasets the important task of classification has to focus on high toxic chemical compounds since misclassification of high toxic chemicals may lead to disastrous consequences.

We propose a new approach, which combines the supervised classification task with unsupervised clustering in order to maximize the knowledge gained from the data characteristics. Firstly selected datasets from Demetra project and UCI repository are trained using a classification algorithm. At the second stage the poorly classified samples are identified by studying the produced confusion matrix of classification task. Then TP rate for these samples is measured and compared with other samples belonging to classes with higher classification accuracy or TP. The class with lowest prediction accuracy produced on its samples is separated and used for the density-based clustering task study. This task is performed on the selected class in order to identify the samples distribution density inside its clusters. The cluster, which contains more samples or with higher prior probability would be identified as the representative set.

Based on the class population and also cluster density, artificial data are generated. The generated data are added to the original dataset and a new training dataset is constructed. With this method we increase the classification accuracy of the less represented class and in most cases with effect on learner accuracy on other classes and also the overall prediction accuracy.

6.2 Related Work

Various approaches and methods have been proposed to tackle imbalanced data problem. One of these methods is one-sided selection [73] in which the borderline/negative examples or the ones overlapping in two class dimensional space are removed.

Another method is DataBoost-IM approach [74]. According to this method the hard examples from minority and majority class are identified. Then the synthetic samples are generated using the hard samples and added to the original dataset. The class distribution and the total weights of the different classes in the new training set are re-balanced at the last stage.

Guided re-sampling technique [75] is another solution which first determines the subcomponents within each class. The element in each subcomponent is re-sampled until each subcomponent has the same number of examples as biggest subcomponent. Then the between-class imbalance is eliminated by randomly selecting and duplicating members of the minority class.

SMOTEBoost [76] is another method which increases the learner performance in classification of minority class with creating synthetic instances by operating in the feature space rather than data space. Using this method a new minority class sample is created in the neighbourhood of the minority class target.

There are also some methods which down-size the majority class in order to equalize the distribution of two classes [77][78]. All these methods concentrate on the two-class problem with minority and majority class: either over-sampling or under-sampling presentation by overlooking the distribution of the class subcomponents [75]. The statistical relationship between these elements is not addressed in detail. This could be very important in terms of how the new samples are generated in order to improve this relationship and help the learner in the classification process.

6.3 Density-based Class-Boost Algorithm (DCBA)

The Density-based Class-Boost Algorithm applies to multi-class domain problem and is based on insight view of class characteristics in order to determine the distribution density of class samples. The idea is based on boosting the core of the hard recognizable class in order to highlight class influence zones [79] or boundaries. The algorithm is presented in Figure61.

6.3.1 Probability Density Clustering

Clustering is based on a statistical model called finite mixture. A mixture is a set of k probability distributions of k clusters. The distribution gives the probability that an instance has a certain set of attribute values if it was identified to be a member of that cluster [80].

With Probability Density Clustering there are few parameters measured for each attribute in the data set and also each cluster within a class. For each attribute, mean, standard deviation and sampling probability are produced. For each cluster S with mean (μ_S) and a standard deviation (σ_S), if the classification is already determined for each sample then:

Mean (average):

6.1	$$\mu = \frac{1}{n}\sum_{1}^{n} X_i$$

Standard Deviation:

6.2	$$\sigma^2 = \frac{1}{n-1}\sum_{1}^{n} (X_i - \mu)^2$$

Sampling Probability for the class (S):

P(S) = the estimation of the number of instances belonging to the class.

With these parameters already identified: The probabilities that instance X belonging to cluster S is:

6.3	$$P_{(s\mid x)} = \frac{p(x \mid s)p(s)}{p(x)}$$

Where P (s|x) is the density function for:

6.4	$$S, f(x; \mu_S, \sigma_S) = \frac{1}{\sqrt{2\pi}\sigma_S} e^{\frac{-(x-\mu_S)^2}{2\sigma_S^2}}$$

Finally the joint probability of an instance is calculated as a sum of the probabilities of all its attributes which is produced as prior probability of instances distribution for each cluster [80] [81].

Figure60 shows the Density-based clustering for class3 in Demetra Trout dataset. The class is divided into two clusters: 0 and 1 with cluster 0 with more members and higher prior probability.

6.3.2 ROC analysis/Evaluation Measures

In this work we used a number of measures which are produced by confusion matrix during classification process. A brief description is provided below. Although in the results table of the experiments we have produced values for ROC curves as they have been produced by confusion matrix but our emphasis is on values produced for True Positives. The reason is that ROC diagram presents the binary data information.

- Confusion Matrix: in a dataset when the classification is performed the prediction for each sample has four possible outcomes: True Positive, False Positive, True Negative and False Negative. They are produced in the form of Confusion Matrix.

6.5	Overall Accuracy = (TP+TN)/(TP+TN+FP+FN)

- True Positives are the members of the class that have been predicted correctly for which the predicted and actual value for class membership are equal.

6.6	True Positive Rate = (TP)/(TP+FN) = Recall

- Recall: shows the proportional relationship between TP and FN rate.

- Precision: shows the proportional relationship between TP and FP.

6.7	Precision =(TP)/(TP+FP)

- F-Measure: this statistical figure simply produces the relationship between Precision and Recall as follows:

6.8	F = (2PR)/(R+P)

6.3.3 Artificial Data Generation

As the weak class is identified after first training with the classifier (the class with lowest TP rate), unsupervised Density-based Clustering is performed.

For every single attribute, mean, standard deviation and sampling prior probability are calculated. The determining facts for the size of additional artificial data are:

-The whole class proportion (the number of samples in target class).

In some cases the target class members are as little as four in OralQuail data set. In order to affect the learner performance, enough artificial data need to be generated. Table58 shows the number of classes and also samples in each class in all data sets.

-The cluster size inside the class (the number of samples in the cluster).

After performing Density-based Clustering task, the class is divided into two clusters with different proportion. When the original class member's size is small, consequently the constructed clusters would be smaller. For every cluster the following formula is applied to determine the cluster size: $S(x) = S_1(x) + S_2(x)$ when cluster S(x) consists of cluster $S_1(x)$ and cluster $S_2(x)$ and x means clusters' member set.

-The effect that the additional data is caused (increase in the classification accuracy).

The data (numerical values) is generated based on the normal distribution/mean values of each attribute based on following: if the frequency distribution has k attributes/features intervals with midpoints: m_1, m_2,..., m_k and corresponding frequencies f_1,f_2,...,f_k, then:

6.9	Grouped mean: $\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{n} = \frac{\sum_{Cells} (Midpo\,\text{int}\times Frequency)}{Totalfrequency}$

In general in our work the added artificial data size is between 10 to 100 percent.

Table58: Datasets class distribution;

Datasets	No. of Classes	Class1	Class2	Class3	Class4	Class5	Class6
Wine	3	51	79	48			
Iris	3	50	50	50			
Vehicle	4	199	217	212	218		
Ecolio	5	143	77	52	35	20	
Glass	6	70	76	17	13	9	29
Trout	4	117	84	34	27		
Daphnia	4	107	64	50	23		
OralQuail	4	4	21	21	58		
Bee	5	14	19	12	38	12	
DietaryQuail	5	8	35	32	22	10	

Table 58 shows the distributions of the members in each class. In Glass dataset class4 had no samples and it has been deleted. The label for other classes has been shifted accordingly.

If the constructed training data set (data set with added artificial data) is identified as T_{nm} and artificial training data as T_m then:

6.10	$\mathbf{T}_{nm} = \mathbf{T}_n \cup \mathbf{T}_m$

Then the artificial data are added to the class. This resized class would replace the original class in the training data set and the new training set is constructed and retrained. The stopping point for generation of more artificial data is when the classification accuracy start decreasing and also no more than the original set size.

Cluster 0: Prior probability: 0.7353
Cluster1: Prior probability: 0.2647

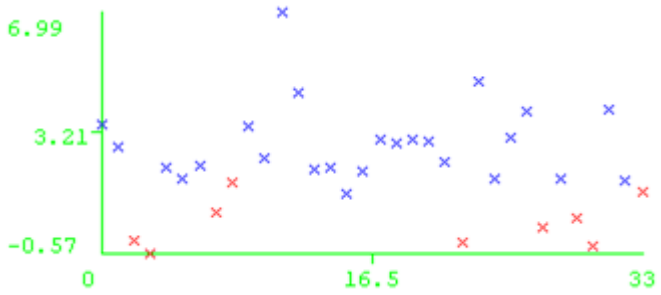


Figure60: Density-based clustering on class 3 in Trout dataset: x shows the number of instances in the class against y which is the value for an attribute ACD1 (-0.57 to 6.99)}

6.3.4 Algorithm Description

Firstly using a Meta learner classification algorithm [81][82] in the accuracy for each class is measured (Figure61 steps 1 and 2). The confusion matrix is presented after the process, the class with lowest TP rate is selected as target class (Figure61 step 3). Sometimes there are two classes with the same TP rate and both are targeted.

The targeted class or classes are then analyzed and with the help of unsupervised Density-based Clustering (Figure61 steps 4 and 5) the prior probability of each cluster is measured. Then within that class, the cluster with highest prior probability is selected (Figure61 step 6). At this stage the value of normal distribution mean of each attribute within the cluster and the frequency of the midpoint value for the samples (Figure61 step 7) are used for generating artificial data. Proportion or sample size in each cluster determines how much data need to be added. The numerical data is generated in a way that satisfies the cluster mean. For instance if the mean value for an attribute is about 0.7, with highest frequency distribution of 0.5-0.6 then the generated numerical values would fall in this range.

The generated data are added to the cluster in target class and finally to the original data set and the new training set is constructed (Figure61 step 8). The training data is balanced and weights are updated (step 9). The error of classification is calculated. This computed error is used to update the weights distribution of the samples (Figure61 steps 10, 11). The data is retrained using the same classification algorithm and result is presented (Output). The stopping criteria for adding more artificial data would be determined by overall classification accuracy of the whole training set in the consequent modelling.

TP rates of all classes are measured at every step of modelling since our experimental work proves that sometimes the increase in TP rate of one class affects the decrease of the same statistical measure in another class. Although initially the less representative class with lowest TP rate is targeted, the performance of the classification algorithm on all the other classes is also measured and watched in order to assure the effectiveness of the method.

The experimental results show that the implementation of the method (adding data to one class) highlights the boundaries and border lines of the other classes which causes the increase in the overall classification accuracy and also each individual class (in most cases).

6.4 Method Evaluation

As it has been mentioned in abstract for the purpose of the evaluation, we have chosen Trout, Bee, Daphnia, DietaryQuail and OralQuail data sets from real-world applications provided by Demetra project. We have also selected Glass, Iris, Wine, Ecolio and Vehicle from UCI Repository. All these datasets are multi-class and imbalanced (Table58) except Iris which is multi-class but balanced dataset. The results of experiment show that the method has been effective in all data sets in terms of increase in overall classification accuracy. In the case of Iris data set although the data set is not imbalanced but the original classification accuracy (on original data set with no artificial data) for class2 was much lower than other classes, so we tested the method to see if it is effective in order to increase the TP rate for this class which was successful.

As the results show for Demetra data sets (Table59) and for UCI data sets (Table60) the method not only increased the classification accuracy for the target class and the

overall accuracy of classification but also TP rate, F-Measure and ROC area of other classes as well (in most cases).

Density-based Class Boost Algorithm

Input: T $_n$ set of n examples $x_1 = (x_1^1, x_1^2, \ldots, x_1^m)$, x $_2 = (x_2^1, x_2^2, \ldots, x_2^m)$, x_2 = $(x_n^1, x_n^2, \ldots, x_n^m)$, with labels $c_i \in C^*$

- m_i , midpoint of distribution of class/cluster intervals (attribute values)

- f_i , corresponding frequencies

- L , number of iterations (1)

For $l = 1$ to L (2)

1. Initialize distribution weights on samples: $D_i(x_i) = \frac{1}{n}$ for all $x_i \in T_n$
2. Train data with meta learner
3. Identify target class S_i if $TP(S_i) = TP_{\min}$
4. Calculate $P(\mathrm{s}|x) = \frac{p(x \mid s)p(s)}{p(x)}$
5. Produce: $S, f(x; \mu s, \sigma s) = \frac{1}{\sqrt{2\pi\sigma s}} e^{\frac{-(x-\mu s)^2}{2\sigma_s^2}}$ in which
6. Calculate: $P(s \mid x) where P(s \mid x) = p(s \mid x)_{\min} + p(s \mid x)_{\max}$
7. For $p(s \mid x)_{\max}$, $generate\ \ T_m = \{x_i, i = (1...m)\}$ when $\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{n}$
8. Add artificial data T_m to original data set: T_n U $T_m = T_{nm}$
9. Balance training data and update weights
10. Train given the distribution $D_l, S_l = learner(T, D_l)$
11. Set $\beta_l = \varepsilon_l / (1 - \varepsilon_l)$ where ε_l error of S_L is: $S_l, \varepsilon_l = \sum_{x_i \in T, S_l(x_i) \neq y_i} D_l(x_i)$

Output: $S^*(x) = \arg\max_{y \in Y} \sum_{l: S_l(x) = y} \log \frac{1}{\beta_l}$

Figure61: DCBA Algorithm

Table59: Classification Accuracy for Demetra Datasets; target classes are in bold

Dataset	Class1	Class2	Class3	Class4	Class5
Bee					
TP	0.286	0.053	**0**	0.684	0.167
F-Measure	0.286	0.071	**0**	0.52	0.2
ROC area	0.765	0.511	**0.494**	0.605	0.702
Bee+artificial					
TP	0.571	0.105	**0.375**	0.658	0.25
F-Measure	0.516	0.143	**0.462**	0.543	0.286
ROC area	0.863	0.559	**0.571**	0.635	0.788
DietaryQuail					
TP	**0.1**	0.457	0.375	0.273	**0.1**
F-Measure	**0.2**	0.41	0.353	0.267	**0.154**
ROC area	**0.487**	0.543	0.581	0.569	**0.656**
DQ+artificial					
TP	**0.417**	0.457	0.406	0.409	**0.286**
F-Measure	**0.476**	0.421	0.366	0.439	**0.381**
ROC area	**0.668**	0.604	0.536	0.695	**0.719**
Trout					
TP	0.615	0.595	**0.147**	0.593	
F-Measure	0.643	0.549	**0.172**	0.533	
ROC area	0.71	0.72	**0.702**	0.847	
Trout+artificial data					
TP	0.667	0.583	**0.295**	0.556	
F-Measure	0.69	0.547	**0.329**	0.5	
ROC area	0.741	0.737	**0.751**	0.825	
Daphnia					
TP	0.664	0.313	0.38	**0.261**	
F-Measure	0.617	0.336	0.376	**0.316**	
ROC area	0.723	0.656	0.71	**0.852**	
Daphnia+artificial data					
TP	0.682	0.281	0.38	**0.407**	
F-Measure	0.635	0.324	0.365	**0.431**	
ROC area	0.7	0.659	0.699	**0.878**	
OralQuail					
TP	**0**	0.048	**0**	0.931	
F-Measure	**0**	0.074	**0**	0.701	
ROC area	**0.463**	0.456	**0.41**	0.482	
OQ+artificial					
TP	**0.143**	0.16	**0.276**	0.931	
F-Measure	**0.2**	0.235	**0.41**	0.697	
ROC area	**0.695**	0.62	**0.546**	0.669	

In the case of Glass, Bee, DietaryQuail (Figure 3:graphical representation) and Iris data sets after adding artificial data TP, F-Measure and ROC area increased for all the classes. In the Vehicle data set all the statistical measured for all the classes have

improved except there is a slight decrease in TP rate of class4 after addition of artificial data. For Daphnia data set although decrease in values of TP rate and F-Measure for class2 and class3 occurred all the other statistical measures have improved. In OralQuail except the decrease in F-Measure for class4 after addition of artificial data the other measures show good improvement.

Table60: Classification Accuracy for UCI Datasets; target classes are in bold

Dataset	Class1	Class2	Class3	Class4	Class5	Class6
Glass						
TP	0.786	0.737	**0.118**	0.769	0.667	0.828
F-Measure	0.738	0.723	**0.19**	0.769	0.6	0.842
ROC area	0.891	0.867	**0.832**	0.939	0.99	0.919
Glass+artificial data						
TP	0.814	0.737	**0.56**	0.769	0.889	0.828
F-Measure	0.792	0.737	**0.636**	0.741	0.727	0.873
ROC area	0.921	0.87	**0.929**	0.942	0.989	0.938
Ecolio						
TP	0.986	0.831	0.788	**0.514**	0.75	
F-Measure	0.956	0.8	0.837	**0.554**	0.833	
ROC area	0.979	0.95	0.942	**0.912**	0.985	
Ecolio+artificial data						
TP	0.972	0.87	0.788	**0.745**	0.7	
F-Measure	0.949	0.832	0.82	**0.792**	0.778	
ROC area	0.978	0.952	0.923	**0.953**	0.979	
Vehicle						
TP	0.94	0.507	**0.481**	0.963		
F-Measure	0.874	0.525	**0.523**	0.923		
ROC area	0.986	0.839	**0.864**	0.982		
Vehicle+artificial data						
TP	0.945	0.512	**0.575**	0.959		
F-Measure	0.87	0.534	**0.605**	0.937		
ROC area	0.988	0.845	**0.878**	0.992		
Wine						
TP	0.966	**0.915**	0.979			
F-Measure	0.958	**0.935**	0.959			
ROC area	0.991	**0.982**	0.997			
Wine+artificial data						
TP	0.949	**0.938**	0.979			
F-Measure	0.949	**0.943**	0.969			
ROC area	0.997	**0.991**	0.998			
Iris						
TP	0.917	**0.793**	0.923			
F-Measure	0.88	**0.821**	0.911			
ROC area	0.923	**0.908**	0.986			
Iris+artificial data						
TP	1	**0.96**	1			
F-Measure	1	**0.98**	0.96			
ROC area	1	**0.997**	0.995			

As it is shown in the result table (Table59) for this data set, the class1 and class3 in the first run classification had zero TP rate. None of the samples belonging to these two classes have been classified correctly. The method shows good implication for such data sets. In the case of Wine data set all the parameters have improved except the TP rate and F-Measure in class1. For Ecolio data set the effect is different. For class1 and class5 the result is not satisfactory also for class3 the F-Measure and ROC area shows decrease but all other parameters for the rest of the data set is good.

Table61: Classification Accuracy for all data sets after testing models

Datasets	**Cross-Validation: Overall Accuracy (%)**	
	Original	**Data added to class 3 (model1)**
Trout	54.6	57
Tested with model1	75.5	
		Data added to class 4 (model1)
Daphnia	47.54	49
Tested with model1	73	
		Data added to class1&5 (model1)
DietaryQuail	33.6	41
Tested with model1	63.5	
		Data added to class1&3 (model1)
OralQuail	52.8	56.3
Tested with model1	68.2	
		Data added to class3 (model1)
Bee	34.7	45.45
Tested with model1	61	
		Data added to class 2 (model1)
Iris	87.5	98.5
Tested with model1	96.2	
		Data added to class3 (model1)
Vehicle	71.9	74.04
Tested with model1	91.72	
		Data added to class4 (model1)
Ecolio	85.3	87.17
Tested with model 1	89.9	
		Data added to class3 (model1)
Glass	71.49	76
Tested with model 1	86.4	
		Data added to class2 (model1)
Wine	94	95.2
Tested with model 1	96.06	

The process of adding artificial data to datasets has been done in one iteration. The method can be applied and data can be added until the overall classification start decreasing. The confusion matrix has to be studied after every iteration, in order to target classes with lowest TP rate.

Although the application of the method show decrease in few parameters(TP, F-Measure or ROC area) in some cases, the overall result is promising and the method is very effective in severe imbalanced data sets such as Bee and OralQuail.

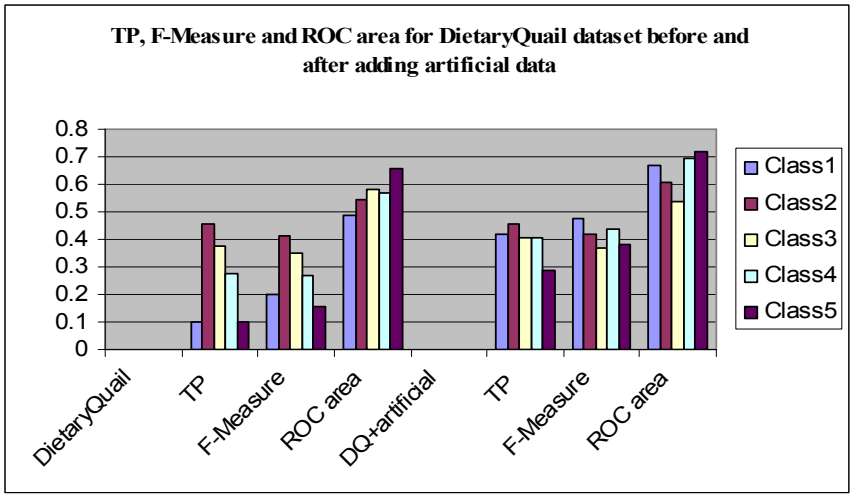


Figure62: TP, F-Measure and ROC area for DietaryQuail dataset before and after adding artificial data

Table61 shows the result of the testing gained models (from added artificial data) on the original data sets (using Cross-Validation). There is a good increase in classification accuracy after this process. There are three values for each dataset in this table. For examples in the case of Trout dataset the first value (54.6) is the overall classification accuracy for the original dataset with no artificial data. After adding artificial data to class3 the accuracy increased to (57). Then the dataset with artificial data was tested against the original dataset which caused the improvement in classification accuracy to (75.5). To ensure the effectiveness of our model, we trained a number of other classifiers to evaluate the performance. Table62 shows the results of this experiment. First the original dataset with no artificial data was trained. The first row of the table shows the classification accuracy for this procedure. Second time the dataset with added artificial data was trained. The second row of the table shows the result. There is an increase of the performance on the expanded datasets in majority of the cases. In very few cases with number of classifiers the performance is either the same or decreased in very low margin. The method proves to be effective with other classifiers as well.

6.5 Summary and Conclusions

In this chapter, we proposed a hybrid algorithm for generation of artificial data. We combined the supervised classification process with unsupervised clustering in order

to get insight view of the classes' internal components and characteristics. We focused our study and implementation of our method on imbalanced and multi-class data sets in which the severe class samples distribution exists.

We have shown that as long as we understand how class members are constructed in dimensional space in each cluster we can reform the distribution and provide more knowledge domain for classifier. Our results are promising and show the affect of the method even in special cases such as Demetra data sets where data is highly imbalanced with very low overall classification accuracy. Our process of data generation and the way the numerical values are produced proved to be effective [83].

Table62: Classification Accuracy for all data sets trained with other classifiers

Classification Accuracy % (Cross-Validation)									
Datasets	**BN**	**MLP**	**J48**	**JRip**	**Ridor**	**NB**	**SMO**	**NNge**	**IBK**
Trout	53.8	51.1	58.3	54.5	48.8	50.7	43.8	49.6	52
Trout+artificial data	55	54	58	55	50.3	54.4	45.5	52.2	52
Daphnia	45.9	44.6	43.8	46.7	45.9	40.5	45.9	45	41.3
Daphnia+artificial data	46	54	46	50	48	45.1	47.3	48	50
DietaryQuail	31.7	29.9	29.9	28.9	34.5	25.2	31.7	33.6	30.8
DQ+artificial	33.9	47.8	33.9	29.0	38	34	35	34	33.9
OralQuail	55.7	52.8	53.8	55	42	29	55.7	47	34
OQ+artificial	60	54	58.2	56.5	45.5	31	58	49	49.5
Bee	37.8	27.3	33.6	37.8	34.7	22	40	33.6	38.9
Bee+artificial	38	29.2	34	38	34.9	30.3	40	41.4	48.5
Iris	92.5	91.2	87.5	87.5	91.2	91.2	93.7	93.7	91
Iris+artificial data	97	98.5	95.5	95.5	97	97.1	94	98.5	95.5
Vehicle	61.5	81	73	68.6	71.1	45	74.5	64	69.6
Vehicle+artificial data	61.5	83	75	71	72	45.8	74.9	67.5	72
Ecolio	85.9	86.2	83.4	83.1	83.1	87	84	84.7	82.2
Ecolio+artificial data	85.9	86.2	84.2	84.2	84.8	87.4	87.7	87.7	82.2
Glass	74.7	67.2	65.8	69.6	68.6	49.5	57.4	66.8	70.5
Glass+artificial data	74.8	67.2	68.4	65.7	71.1	45.8	54.9	71.5	73
Wine	99.3	95.4	92.8	88.9	91.5	96.7	98.7	93.5	94.8
Wine+artificial data	98.9	97.2	93	90.7	92.3	97.2	98.8	96.1	95.6

7. CONCLUSIONS

The research work was based on various data collections, available publicly or for restricted users. We had detailed investigation and study on the toxicity data available on online sources and also on confidential Demetra data collections. We have also studied and used benchmark data for our experiments. We have shown that the toxicity data is unreliable and possess low quality for number of reasons. Its presentation is also not consistent throughout different sources and that makes the interpretability and accessibility of the information difficult. The toxicity data have different dimensions. The chemical compounds have a number of properties (descriptors), which relate to their biological activities. This relationship can be modelled by QSARs, which are used for toxicity prediction of untested chemicals. There are standard procedural steps to build QSARs. Data quality and descriptor selection play important roles in this process. Data quality and QSARs are two fundamental elements in toxicology studies.

Data quality is an important issue in scientific domains. We have studied number of proposed approaches for tackling the problem for predictive toxicology domain. Naumman introduces an information quality framework based on the user, information and the process of accessing this information. Another example is Fusionplex, which is a system that integrates information from multiple sources and also resolves data inconsistencies by use of fusion methods. COLUMBA is another system that performs the quality check by data cleansing procedures. The Information Quality Assessment Methodology introduced by Richard Y. Wang, overcome quality issue by defining number of criteria in its components. The methodology for establishing and maintaining quality in data context is another strategy, which assess the data at different levels. Helma also introduces some methods for measuring quality in predictive toxicology. Data cleaning is also used to enhance the quality of the data. There is also an issue of missing values in datasets, which reduces the quality and reliability of the data. There are some methods in use to overcome this problem. Some of these methods are such as: omit records, calculate average, single imputation, multiple imputations and expectation maximisation.

Another issue is modelling this data which raises the problem of how subcomponents of this data have been structured that identifis the data as balanced or imbalanced. The

imbalanced data affect the performance of classifiers during the classification process. Number of methods has been proposed to overcome imbalanced data problem such as; re-sampling, DataBoost-IM and SMOTEBoost.

During our study, we have also designed and implemented a prototype for data storage and representation for internal use. For this purpose three main stages have been processed: analysis, design and implementation. Through analysis stage, issues relate to user requirement have been considered such as system usability. Tasks chart also has been designed to show the graphical view of the whole system. There are some design issues from human computer interaction point of view, such as simple navigation, consistent layout and help feature which have been considered at this stage. The tools and application used for implementation were MySql database and PHP. These applications have been chosen for their simplicity of implementation and process and ease of run on any server and platform. The prototype is accessible internally and externally. The motivation behind this stage of the project was to implement an integration system which collects processes and evaluates data from different sources and stores internally. The whole system design chart has been produced in Appendix 1.

Another stage of the project was to investigate online toxicity databases in order to highlight the inconsistencies in values presentation and also the structural differences from source to source. We have also provided the results of our detailed investigation and experimental work on Demetra data. We studied two different file presented by two programs ACD and Pallas for same chemical compound and same species. The global values (min, max, average, standard deviation) have been measured and presented and also difference between values for same compounds generated by two programs highlighted. Based on these values we have selected number of descriptors to swap between two files for same species in order to see how variation in values could affect the model performance. The idea was to show that the good model is purely depend on how the descriptors have been generated and by what tool which directly affect the data quality. We have also tried to understand the data characteristics and insight view of the relations between descriptors. The results of this investigation have led us to identify a general framework for data quality and also clearer path for further study.

At the later stage based on our findings, experiments and the results obtained from previous work, number of data quality criteria has been identified in order to provide

a solution for valuation of the data in toxicology domain. We have also provided the quality flow chart which shows the quality check of the data in steps with five identified criteria. An algorithm has also been proposed in details which explain how data is assessed and processed before modelling. These criteria are related purely to data values and can be added to other quality criteria which have been proposed in previous chapters to form a complete quality framework.

The problem of missing values in datasets has been addressed specially in toxicology domain. The toxicology approach has been discussed that deals with the problem at the collection stage. Least Square Method for paired dataset and Serial Correlation for single dataset provided the solution for the problem in two different situations. An algorithm using these two methods has been proposed in order to overcome the problem of missing values. The proposed algorithm has been tested on number of Demetra datasets to test the effectiveness on the outcome model after recovery of missing values. Also the Least Square Method has been used to generate artificial data around outliers to improve model performance. The implementation of the proposed procedural algorithm requires the existence of the high correlation between descriptors (attributes) in the dataset.

Producing better modelling results based on generation of artificial data and affect of this process on data characteristics, we proposed a hybrid algorithm but from different point of view. We combined the supervised classification process with unsupervised clustering in order to get insight view of the class's internal components and characteristics. We focused our study and experiments on imbalanced and multi-class data sets in which the severe class samples distribution exists.

We have shown that as long as we understand how class members are constructed in dimensional space in each cluster we can reform the distribution and provide more knowledge domain for classifiers. Our results are promising and show the affect of the method even in special cases such as Demetra data sets where data is highly imbalanced with very low overall classification accuracy. Our process of data generation and the way the numerical values are produced proved to be effective.

7.1 Future Work

The project at this stage can take different routes. Three main elements: data and its components, classifiers and their performance and also statistical measures can be further studied from different aspects to highlight more relations.

One of the possible routes can focus on evaluation measures of our algorithm before and after addition of artificial data in order to see how the additional data could affect the whole data set characteristics not only the target classes. The number of artificial samples added to the set can also be considered [83]. Another suggestion could focus on classifiers performance. On our work we just used number of classifiers provided by Weka but using different data mining tools, more classifiers can be trained in order to highlight which classifier can perform better in artificial data domain.
Considering elements relate to data properties, nominal attributes can be studied as well as numerical values. We can concentrate the whole process of the artificial data generation with emphasis on nominal attributes. This can be implemented in the domains where these values affect the performance of the classifiers. We can also investigate the correlation between chemical compounds and generated data in order to assure how much the artificial data values could be close to the real values. All of these possible routes can start a new project in data mining and knowledge discovery.

7.2 Original Contributions

- Research on "online toxicity databases (ex: DssTox)" and their drawbacks. Chapter 2 discusses our detailed investigation on this issue. Some of these results have been produced in first and second paper presented to the University of Bradford workshop [7][27].
- Study Data Quality Assessment methods. The study on these methods is detailed in chapter 3.
- Detailed investigation on internal data: Demetra. The results of our experimental work on the issue are discussed in chapter 5.
- Detailed investigation on quality assessment methods and our data: Demetra dataset included in chapter 5.
- Identify data deficiencies and effects on QSAR models are also discussed in chapter 5. The results of this work have been summarized and presented at the UKCI international conference [66].
- Define a framework for model processing using the results of the investigation. The results of this work are discussed in chapter 6.
- Investigate on data quality criteria description based on mathematical concepts also included in chapter 6.

- Define data presentation and quality issues in predictive toxicology in chapter 6.
- Defining detailed “Data Quality Assessment Framework” in chapter 6. The results of the work are published and presented in AAIA international conference, also the extended version with the result of more experimental work has been published in TQ (Task Quarterly) computer science Polish national journal [67][68].
- Define method and algorithms for calculation of missing values which discussed and proposed in chapter 7.
- Define framework via generation of artificial data for improvement of model performance using defined method. It has also been proposed in chapter 7.
- Investigate the relationship between model performance and distribution of classes of compounds. This topic is discussed in chapter 8.
- Define algorithm to generate artificial data in imbalanced multidimensional datasets and improve model performance. The algorithm has been proposed in chapter 8. The results of this work have been published in ICDM international conference [83].

REFERENCES

[1] Seedcorn Proposal, Central Science Laboratory, York, 2004-2005.

[2] The Integrated Use of Alternative Approaches for Predicting Toxic Hazard, The report and recommendations of ECVAM Workshop 81,2, Reprinted with minor amendments from ATLA 23: 410-429, 1995. 1ECVAM - The European Centre for the Validation of Alternative Methods.

[3] IRIS: http://www.epa.gov/iris/

[4] Richard, A. M., C. Williams,R. Cariello, N., F., 2002. Improving Structure-Linked Access to Publicly Available Chemical Toxicity Information, *Current Opinion in Drug Discovery and Development* , 136-143

[5] Helma, C. Kramer, S., Pfahringer, B., Gottmann, 2000. E. Data Quality in Predictive Toxicology: Identification of Chemical Structures and Calculation of Chemical Properties, *Environmental Health Perspectives*, vol. 108, No. 11

[6] Warr, W. A. 2003. IUPAC project meeting: Extensible Markup Language (XML) Data Dictionaries and Chemical Identifier. National Institute of Standards and Technology. USA

[7] Malazizi, L. Neagu, D. Chaudhry Q., 2005. Knowledge Representation in Predictive Toxicology, *6th Informatics Workshop for Research Students*, University of Bradford, 115-119, ISBN 1-85143-220-5

[8] McCourt, D. Lopez, J. Benfenati, E., Mazzatorta, P., Romberg, M. Schuller, B., Dubitzky, W., 2003. Toward an Intelligent Data Type for Toxicity. Proceedings of the International Conference on Artifical Intelligence (IC-AI), Las Vegas, USA, CSREA Press, 328-334

[9] Extoxnet:, http://pmep.cce.cornell.edu/profiles/extoxnet/TIB/extoxnetglossary.html, Extension Toxicology Network, Extoxnet Glossary

[10] http://toxnet.nlm.nih.gov

[11] http://www.cdc.gov/niosh/idlh/107028.html

[12] http://www.epa.gov/iris/subst/0364.htm

[13] http://www.ilo.org/public/english/protection/safework/cis/products/icsc/dtasht/_icsc00

[14] http://www.meddb.info/index.php.en?cat=13

[15] http://www.acdlabs.com

[16] http://www.osc.edu/ccl/pallas.html

[17] DSSTox: http://www.epa.gov/nheerl/dsstox/

[18] Richard, A. M., Williams, C. R., 2002. Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network: A Proposal, *Mutation Research*, vol. 499:27-52

[19] Helma, C., 2005. *In silico* predictive toxicology: the state-of-the-art and strategies to predict human health effects, *Current Opinions in Drug Discovery & Development*, 8:27-31

[20] http://www.enotes.com/public-health-encyclopedia/vivo-vitro-testing/

[21] http://www.pesticides.gov.uk

[22] http://accelrys.com/products/discovery-studio/toxicology/

[23] http://lhasa.harvard.edu

[24] http://oasis-lmc.org

[25] Eriksson, L. et al. 2003. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification and Regression–Based QSARs, *Environmental Health Perspectives*, vol. 111, no. 10, 1361-1375

[26] http://contamsites.landcareresearch.co.nz

[27] Malazizi, L., Neagu, D., Chaudhry, Q., 2006. Investigating Data Quality Assessment In Predictive Toxicology, *7th Informatics Workshop for Research Students*, University of Bradford, 129-132, ISBN 1-85143-2329

[28] Wald, L., 1999. Some Terms of Reference in Data Fusion, *IEEE Transactions on Geoscience and Remote Sensing,* Institute of Electrical and Electronics Engineers, vol 37:33, 1190-1193

[29] Roth, M. A., Wolfson, D. C., Kleewein, J. C. Nelin, C. J., 2002. Information Integration: A New Generation of Information Technology , *IBM systems journal*, vol. 41:4, 563-577

[30] Kambhampati, S. Knoblock, C. A., 2003. Information Integration on the Web, *IEEE Intelligent System, 14-15*

[31] Buchroithner, M.,1998. Geodata interrelations: Inventory and structuring attempt of taxonomic diversity, *In Proc. 2nd Conf. Fusion of Earth Data: Merging Point Measurements, Raster Maps and Remotely Sensed Images*, 11-15.

[32] Wald, L., 1997. Data fusion: An overview of concepts in fusion of Earth data. *In Proc. EARSeL Symp*, *Future Trends in Remote Sensing*. 385-390.

[33] Leser, U. et all, 2005. COLUMBA: Multidimensional Data Integration of Protein Annotations, BMC Bioinformatics, Germany, 6:81
[34] UCI Data Repository, http://kdd.ics.uci.edu
[35] http://us3.php.net/manual/en/class.com.php
[36] Preece J., Rogers Y., Sharp H., Benyon D., Holland S. and Carey T.,1994. Human Computer Interaction. England: Pearson Education Limited
[37] Gilmore J., 2004. PHP 5 and MySQL: From Novice to Professional. USA: Springe-Verlag.
[38] McDermid D., 1990. Software Engineering for Information Systems. England: BlackWell Scientific Publications.
[39] Connolly T., Begg C. and Strachan A., 1999. Database Systems. 2nd ed. England: Pearson Education Limited
[40] Hunter, A. How to act on inconsistent news: Ignore, resolve, or reject, Data & Knowledge Engineering, Elsevier, no.57, 221–239,
[41] Naumann, F., Roker, C., 2000. Assessment Methods for Information Quality Criteria, *Proceedings of the International Conference on Information Quality* (IQ2000), Cambridge, 148-162
[42] Anokhin, P., Motro, A., 2003. Fusionplex: Resolution of Data Inconsistencies in the Integration of Heterogeneous Information Sources, Technical Report ISE-TR-03-06, Information and Software Engineering Dept., George Mason Univ, Virginia , ISSN:1566-2535, 176-196
[43] Yang, L., Strong, D. and Wang, R., 2002. AIMQ: A Methodology for Information Quality Assessment, *Information and Management*, vol. 40:2, 133-146
[44] Tap, R., 1999. A Methodology for Establishing and Maintaining Quality in Data Context, Conference on Information Quality*, MIT Sloan School of Management*, 209-219
[45] Angeles, P., Mackinnon, L. M., 2007. Detection and Resolution of Data Inconsistencies and Data Integration using Data Quality Criteria
[46] Naumann, F., Leser, U., 1999. Quality Driven Integration of Heterogeneous Information Systems. *Proceedings of the International Conference on Very Large Databases*, VLDB'99, Edinburgh. Morgan Kaufmann. pp. 447-458.
[47] Xiong, H., Pandey, G., Steinbach, M., Kumar, V., 2006. Enhancing data analysis with noise removal*. Data & Knowledge Engineering, IEEE*, 18/3
[48] Ennet C., Frize M., Walker C., 2001. Influence of Missing Value on Artificial Neural Network Performance*, Medinfo*
[49] Pearson R., The Problem of Disguised Missing Data. SIGKDD Exploration. 8/1, 83-92
[50] Yuan Y. Multiple Imputation for Missing Data: Concepts and New Development. *SAS Institute Inc*, Rockville, 267-25
[51] Maloof, M.A., 2003. Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown. *ICML-2003: Workshop on Learning from imbalanced data sets II*
[52] Ertekin, S., Huang, J., Bottou, L., Giles C.L., 2007. Learning on the Border: Active Learning In Imbalanced Data Classification. *CIKM07*, Lisbon, Portugal
[53] Kubat, M., Matwin, S., 1997. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In*: 14th International Conference on Machine Learning,* 179-186. Morgan Kaufmann, San Francisco
[54] Guo, H., Viktor, H.L., 2007. Learning From Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach. *Sigkdd Explorations*, 6, 30-39
[55] Nickerson, A.S., Japkowicz, N., Milios, E., Using Unsupervised Learning to Guide Re-sampling in Imbalanced Data Sets. *In: 8th International Workshop on Artificial Intelligence and Statistics*
[56] Chawla, N.V., Lazarevic, A., O.Hall, L., Bowyer, K., 2003. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. *In: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 107-119. Cavtat-Dubrovnik, Croatia
[57] Japkowicz, N., 2000. Learning from Imbalanced Data Sets: A Comparison of Various Strategies. Technical Report, *The AAAI Workshop*
[58] Japkowicz, N., 2000. The Class Imbalance Problem: Significance and Strategies. *In: 2000 International Conference on Artificial Intelligence*, 111-117, IC-AI
[59] http://www.cerc.usgs.gov/
[60] http://www.demetra-tox.net
[61] http://www.acdlabs.com
[62] http://www.osc.edu/ccl/pallas.html
[63] http://www.cs.waikato.ac.nz/ml/weka
[64] http://www.spss.com/
[65] http://www.talete.mi.it/help/dragon_help/index.html?IntroducingDRAGON

[66] Malazizi, L., Neagu, D. Chaudhry, Q., 2006. Investigation, Assessment and Identification of Possible Data Quality Criteria in Predictive Toxicology, UKCI, Leeds, 229-236
[67] Malazizi, L. Neagu, D. Chaudhry, Q., 2006. A Data Quality Assessment Algorithm with Application in Predictive Toxicology, *AAIA*, Poland, ISSN: 1896-7094, 131 – 140
[68] Malazizi, L. Neagu, D. Chaudhry, Q., 2006. A Data Quality Assessment Algorithm with Application in Predictive Toxicology, Extended Version, *TASK Quarterly (Scientific Bulletin of Academic Computer Center in Gdansk)*, Poland , ISSN 1428-6394, volume 11, number 1-2/2007
[69] United Estates Environmental Protection Agency, Quality Assurance Division, Washington
[70] Kaiser, K.L.E., 1983. QSAR in Environmental Toxicology. Holland: D. Reidel Publishing Company
[71] McGarry, K., 2005. A Survey of Interestingness Measures for Knowledge Discovery. *The knowledge Engineering Review*, vol. 00:0, 1-24. Cambridge University Press
[72] Bhattacharyya G. k., Johnson R. A., 1977. Statistical Concepts and Methods. John Wiley and sons
[73] Herbin, M., Bonnet, N., Vautrot, P., Estimation of the Number of Clusters and Influence Zones. Pattern Recognition Letters, *Elsevier Science*, 22, 1557-1568
[74] Witten, I.H., Frank, E., 2005. Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann
[75] Frank, E., Hall, M., 2003. Visualizing Class Probability Estimators
[76] Melville, P., Mooney, R.J., 2005. Creating Diversity in Ensembles Using Artificial Data. Information Fusion, 6:1 99-111
[77] Malazizi, L. Neagu, D. Chaudhry, Q., 2008. Improving Imbalanced Multidimensional Dataset Learner Performance with Artificial Data Generation: Density-based Class-Boost Algorithm, ICDM conference, Germany, 165-176, ISBN 978-3-540-70717-2

GLOSSARY OF TERMS

Predictive Toxicology: predicting the toxic properties of chemicals based on the knowledge of their structures.
Descriptor: an element that describes a specific property of the compound.
Chemical compound: is a substance formed from two or more elements, with a fixed ratio that determines its composition.
Endpoints: a biological effect used as an index of the effect of a chemical on an organism.
CAS Identifier: Chemical Abstract Service Registry number
In silico: analysis performed using computers in conjunction with informatics capabilities.
***In Vitro*:** testing or action outside an organism (e.g. inside a test tube or culture dish.)
***In Vivo*:** testing or action inside an organism.
MOA (Mechanism of Action/Mode): the way a chemical compound interacts with a living system.
Dose: a measured amount of a chemical compound. Dose is often expressed in milligrams per kilogram (mg/kg) or parts per million (ppm).
LD_{50}: the amount of a chemical that is lethal to one-half (50%) of the experimental animals exposed to it. LD_{50}s are usually expressed as the weight of the chemical per unit of body weight (mg/kg).
LC_{50}: the amount of a chemical that is lethal to one-half (50%) of the experimental animals exposed to it. LD_{50}s are usually expressed as the weight of the chemical per unit of body weight (mg/kg) for aquatic species
Acute (short term): the short-term effects of a one-time exposure to a chemical substance.
Sub-chronic (mid-term): intermediate between acute and chronic toxicities;
Chronic (long-term): toxic effects resulting generally from long term exposure at low doses
Acrolein: unsaturated aldehyde, formula CH 2 CHCHO, formed as an oxidation product of butadiene, which is a common emission from automobiles.
ACD: Advanced Chemistry Development, Inc., (ACD/Labs) is a chemistry software company offering solutions that integrate chemical structures with analytical chemistry information.
PALLAS: software predicting pKa,logP, logD values and metabolites based on structural formulae of compounds.
Q(SAR): Quantitative structure activity relationship. A mathematical relationship between biological activity of a compound and computed (or measured) properties that depend on the molecular structure.
Data Quality: refers to the features and characteristics that ensure data are accurate and complete and that they convey the intended meaning.
Multidimensional Database: a multidimensional database (MDB) is a type of database that is optimized for data warehouse and online analytical processing OLAP applications
Toxnet: The National Library of Medicine's (NLM's) TOXNET provides access to a cluster of databases on toxicology, hazardous chemicals, and related areas.
DSSTox: Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network

IRIS: EPA's Integrated Risk Information System, an electronic data base containing the Agency's latest descriptive and quantitative regulatory information on chemical constituents. UCI:
DEMETRA: is a project funded by the European Commission to address the eco-toxicity evaluation of pesticides in a way suitable for the Directive 91/414 on pesticides.
Pesticide: Substances intended to repel, kill, or control any species designated a "pest" including weeds, insects, rodents, fungi, bacteria, or other organisms. The family of pesticides includes herbicides, insecticides, rodenticides, fungicides, and bactericides.
TOPKAT: is a toxicity prediction program. TOPKAT uses Kier & Hall electrotopological states (E-states) as well as shape, symmetry, MW, and logP as descriptors to build statistically robust Quantitative Structure Toxicity Relationship (QSTR) models for over 18 endpoints.
DEREK: is an expert knowledge base system that predicts whether a chemical is toxic in humans, other mammals and bacteria.
OASIS: laboratory of mathematical chemistry is a tool for predicting toxicity of chemicals resulting from their metabolic activation.
Good Laboratory Practice (GLP): written codes of practice designed to reduce to a minimum the chance of procedural or instrument problems which could adversely affect a research project or other laboratory work.
LogP: the logarithm of the partition coefficient.
LogD: the logarithm of the distribution coefficient.
Data integration: is the problem of combining data residing at different sources and providing the user with a unified view of these data.
Data fusion: is generally defined as the use of techniques that combine data from multiple sources and gather that information in order to achieve inferences, which will be more efficient and potentially more accurate than if they were achieved by means of a single source. Data fusion is integration followed by reduction or replacement.
PHP: an open-source, server-side scripting language used to create dynamic web pages.
MySQL: a database management system which is available for both Linux and Windows.
HTML: short for HyperText Markup Language, the authoring language used to create documents on the World Wide Web.
Risk Assessment: the overall process of using available information to predict how often hazards or specified events may occur and the magnitude of their consequences.
ITER: International Toxicity Estimated for Risk is a free Internet database of human health risk values and cancer classifications for over 600 chemicals of environmental concern from multiple organizations worldwide.
HSDB: Hazardous Substances Data System is HSDB is a factual database focusing on the toxicology of potentially hazardous chemicals.
ECOTOX: ECOTOXicology database provides single chemical toxicity information for aquatic and terrestrial life.
USGS: is the acute toxicity database which summarizes the results from aquatic acute toxicity tests conducted by the USGS CERC located in Columbia, Missouri.
Winsorized mean: involves the calculation of the mean after replacing given parts of a probability distribution or sample at the high and low end with the most extreme remaining values.

Weka: is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization.

APPENDIX1

1. DATA STORAGE AND ACCESS PROTOTYPE

The motivation behind the design and implementation of this prototype was to provide a sample data integration system for toxicity data. Since the variation in data representation on different sources causes data accessibility and processing problem, our prototype would encourage unifying structure and representation throughout.

The solution developed for the system involves using a PHP5 [35] and HTML front end connected to MySQL Server (database) both which run on the university server. These two tools provide environment and all the facilities needed to perform various tasks in the system. Functionalities for the system are implemented in the front-end supported by back-end database. This implies that the following core objectives have been achieved:

- Selection of the suitable methods of analysis and design, database and a programming application.
- Identifying the system requirements for the system.
- Design and implement the database.
- Design and implement an interactive user interface considering human computer interaction issues that supports the user and administrative functions.
- Design and implement the code to store retrieve and manipulate data in the database.
- Test completed application using an appropriate testing strategy.

The program has been uploaded to the web space assigned for this work on the server and also is accessible from external sources. Following table shows the server specifications, which supports and run the project.

Table1: Server specification for the implemented system

System	Linux linux2
Server API	Apache 2.0 Handler
Apache Version	Apache/2.0.45 (Unix) PHP/5.0.0
Hostname: Port	linux2.inf.brad.ac.uk:0
HTTP_HOST	linux2.inf.brad.ac.uk:59333
SERVER_NAME	linux2.inf.brad.ac.uk
SERVER_ADDR	143.53.28.29

1.1 System Analysis

The objective of this phase is to confirm that the requirements specification is feasible in terms of being implemental. Requirements of the system have been gathered through research in order to identify the application goals and the scope of the information retrieval and management system. This is the first prototype designed and implemented.

1.1.1 Goals of Information Retrieval and Management System

Based on the initial research and discussion on the matter and considering the existing systems, the application's goals were identified as following:

-Provide functionalities for administration of the database through interface.

-Generating connection to MySql server and display updated database details to the user through interface.

-Providing interface for the user to connect to external databases.

-Provide functionalities in order to collect feedback from the user and sent to administrator through email.

-Provide help facilities through connection to technical manual.

-Provide search facilities in order to carry out search on the internal database.

-Provide site map in the system for usability (considering easy navigation and usability issues).

These goals and the scope of the system were identified in terms of the requirements (for this stage). The requirements listed below are the assumptions we made through discussion within the project. Some of these have been discussed with the customer at Central Science Laboratory. These requirements can be considered in three general categories:

A. **Usability requirements**: specify the acceptable level of user performance and satisfaction with the system.

B. **System requirements:** represent behaviour of the system in terms of what the system has to be capable of doing. It is directly interacted with user actions.

C. **Data requirements**: specify the structure of the system and the data that must be available for processing to be successful [36].

Task Analysis: is part of usability requirement process. The tasks associated with the system can be performed by employees and managements. These tasks can be

categorised in three modules: Master Maintenance, Files and Data Maintenance and System Maintenance.

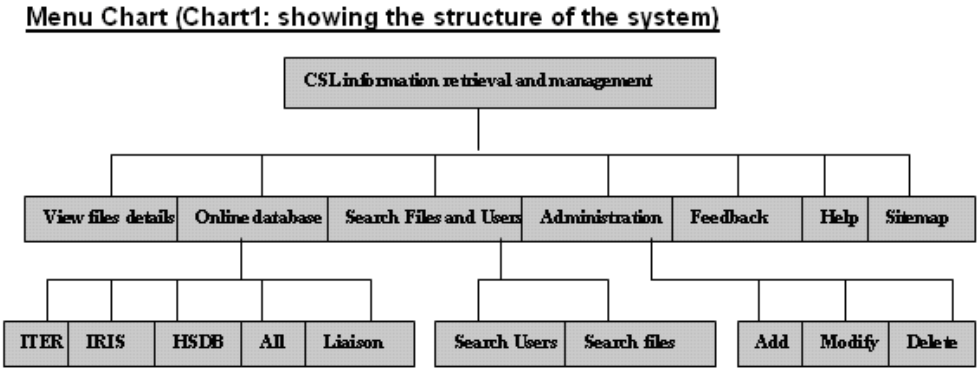


Figure63: The menu chart showing the structure of the system

-*Master Maintenance:* all data pertaining to "files" and "users" should be stored and maintained. The maintenance module should provide for addition, modification, viewing and deletion of these two databases information. Administration can perform addition, modification, deletion and viewing through application's main user interfaces. Employees are not authorised to edit any existing data in the related tables. These tasks are performed only through administration window by authorizing the user first. The files and any toxicity data stored in the system should be updated frequently to store authority and validity. Information on the system should be checked regularly to make sure the system quality and integrity. Administration is also responsible for updating the interfaces if anything needed updating. He is also the recipient of the user's feedback, which should be checked and dealt with regularly.

-*Files and data maintenance:* specific authorised users (main members of the project or managements) are responsible to provide the files and make sure of the files integrity.

-*System Maintenance:* includes backing up and restoring the data. Management and administrator operate all the related tasks. Employees are not authorized to perform these tasks.

1.2 System Design

This chapter focuses on system design issues related to front end design, screens or forms and database design and tables in terms of the input and presentation method

and structure. Since designing and implementing an interactive user interfaces that supports the users and administrative tasks was one of the core objectives for this project, the need for designing the interfaces for any user type even for those unfamiliar with the computers, became apparent.

1.2.1 Database Design

The database for the system was built on MySql server database, which runs on the university of Bradford Linux server. The database consists of two tables at the moment: "Users" table: which consist of information about the users of the system. This table is hypothetical and is for experimental purposes. The details stored in this table are the details of some members of the project.

"Files" table: this table contains information about number of Excel files stored on the server. These files contain toxicity data, which are provided by CSL.

A. Description of the Entities

Followings are the entities created in each table:

Type	Attributes	Null	Default	Action						
ID	smallint(5)	UNSIGNED	No	auto_increment	Change	Drop	Primary	Index	Unique	Fulltext
FileName	varchar(55)		No		Change	Drop	Primary	Index	Unique	Fulltext
FileAuthor	varchar(35)		No		Change	Drop	Primary	Index	Unique	Fulltext
CreationDate	varchar(35)		No		Change	Drop	Primary	Index	Unique	Fulltext
Comments	mediumtext		No		Change	Drop	Primary	Index	Unique	Fulltext

Keyname	Type	Cardinality	Action		Field
PRIMARY	PRIMARY	6	Drop	Edit	ID
FileName	INDEX	None	Drop	Edit	FileName

Type	Attributes	Null	Default	Action						
rowID	smallint(5)	UNSIGNED	No	auto_increment	Change	Drop	Primary	Index	Unique	Fulltext
lastname	varchar(35)		No		Change	Drop	Primary	Index	Unique	Fulltext
firstname	varchar(35)		No		Change	Drop	Primary	Index	Unique	Fulltext
email	varchar(55)		No		Change	Drop	Primary	Index	Unique	Fulltext
company	varchar(100)		No		Change	Drop	Primary	Index	Unique	Fulltext

Keyname	Type	Cardinality	Action		Field
PRIMARY	PRIMARY	5	Drop	Edit	rowID
email	UNIQUE	5	Drop	Edit	email
lastname	INDEX	None	Drop	Edit	lastname firstname

Figure64: All tables designed in SQL

1.2.2 Front End Design Issues

Ease of use of the system and simplicity has been considered as main issues in designing the interfaces. Followings are the main elements of the design considered for the system:

-*Simple navigation* is one of the important design elements. When running an application, the user should be able to navigate to the desired screen through main menu options. For that purpose the main options menu is implemented in all pages. Site map has also been implemented to provide quick accessibility to all the existing pages on the system.

-*Consistent layout* is another design element. Layout of the screens is consistent throughout. Different attributes in each page have been grouped in sections with relevant headings for easier recognition. The heading's font is also bigger than the item's font in each section. In each screen the size of the text boxes is same throughout, and they have same alignment. The background colour of all the pages is consistent. Each page has a guidance description about itself to help the user to understand what sort of tasks he can perform and what type of functionality the page provides. All the pages contain same header, which explains the system.

-*Help and explanatory* features have also been created to help user in case of confusion.

-*Minimal Input:* to prevent user errors, the number of inputs in different screens has been very limited. Most of the information is displayed by the system through table display or selection menu. Minimizing inputs also helps to ease memorising different elements on the screen. All together make the design easy to learn and understand.

1.2.3 Forms/Screens

There are eight main screens designed for this system to perform the various tasks for information management and retrieval. These are discussed below:

Home page: this page contains information about the whole system and also provides links to following pages: View File Details, Online Databases and Pesticides Information Resources, Search Internal Databases of Files and Users, Access to Administration Page, Feedback Form, Help and Site map.

View Files Details: displays a table of files details, which is created in MySql server database.

Online Databases and Pesticide Information Resources: provides links to five online and public databases as follow: ITER (International Toxicity Estimated for Risk),

IRIS (Integrated Risk Information System), HSDB (Hazardous Substances Data System), All Databases (search on all the above databases) and Liaison which is a database provided by CSL and subscription needed for access to the data.

Search Internal Databases of Files and Users: this page provides links to two other screens: Search Users Database and Search Files Database. These pages perform search on these two databases.

Access to Administration Page: provides links to three other pages: Add a new record, Modify record and Delete a record. The administration has access to these pages since access is through authentication.

Feedback Form: provides an input screen for the user to write his feed back about the system. This form would be sent to the administration through email.

Help: provides link to technical report.

Sitemap: provides a map of all the pages on the site and show the site structure with the link to each individual page.

Followings are the few screenshots of the main screens of the prototype.

Figure65: Home page of the prototype

Figure66: Onlinedatabases page of the prototype

Figure67: Searchinternal page of the prototype

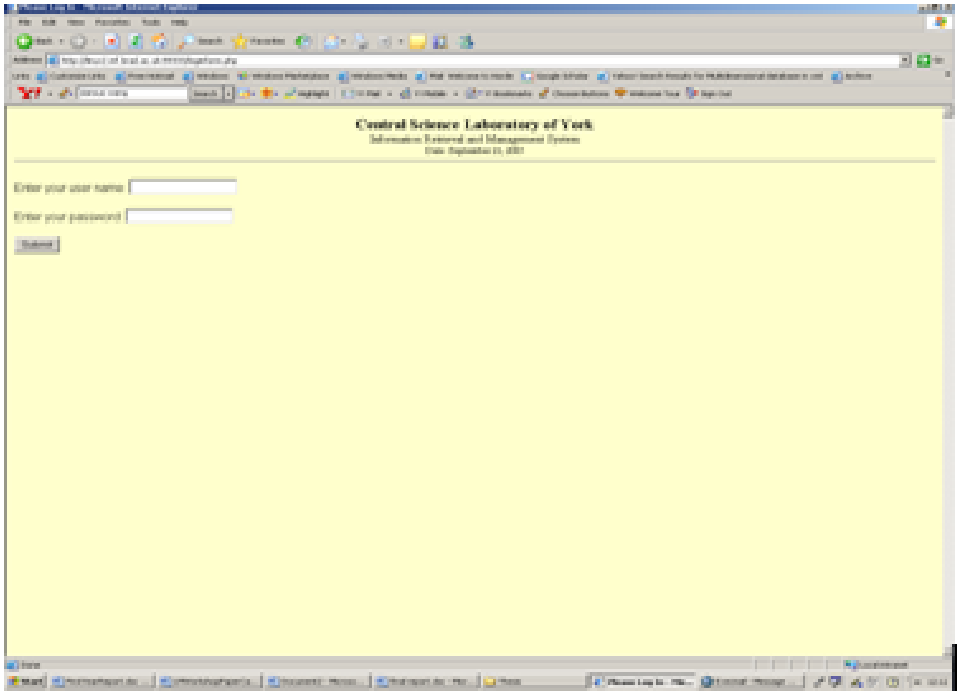


Figure68: Loginform page of the prototype

1.3 System Implementation

This chapter explains the type of the architecture, techniques and development tools that have been used to implement the system and also evaluating these tools and techniques in terms of their appropriation for this project.

1.3.1 Architecture

The information retrieval and management system is consisting of a front-end application, which is created, in PHP5 and a back end database created in MySQL [37] server. The system architecture is Client-Server.

-*Client-Server Architecture:* consists of one or more client applications communicating requests to another application, which is designated as the server [38].

-*Database Management System Architecture:* this architecture is the best way to support multi-user environment. Database Management System is a software system that enables users to define, create and maintain the database and provides controlled access to this database. It allows users to specify the data type and structures and the constraints on the data to be stored in the database [39].

1.3.2 Development Tools

Developments tools that have been used to implement the system are consisting of front-end interfaces for employees and managements and back-end database

accessible by management. The summary of these tools specifications and their users supporting features are discussed below:

-*Front-End Application:* front-end interfaces were created using PHP5 and HTML. PHP5 is very powerful programming language specially running on the server.

-*Back-End Database:* back-end database for this system was implemented in MySql server that is one of the leading database management systems available. It runs on the university server, which makes it easier to work parallel with PHP.

1.3.3 Testing Strategy

The main purpose of the testing strategy for this system was to confirm that executable code exactly meet its objectives and validate what has been completed is what the system goals specified. The followings are the main steps taken to check the implementation for conformance with the system specification:

1) Unit Testing: each individual component has been tested to verify they operated as it is specified in the system design. In this respect following components have been tested:

-*Testing Forms:* all the links and forms have been tested. At this moment the main functionality lays on administration tasks and also searches Excel files (which is implemented on local host). Due to the nature of the development with PHP and HTML, each page is run in the browser to assure the correct coding.

2) System Testing: the whole system functionality has been tested to make sure all the forms and database display and update accurate data accordingly. To satisfy this purpose, database was loaded with appropriate data, each form was tested to make sure this data can be “displayed”, “updated” and “deleted” also this data is imported properly and that each of the controls was bound to the proper database field.

1.4 Summary and Conclusions

The motivation behind the design and implementation of this prototype was to provide a sample data integration system for toxicity data. Since the variation in data representation on different sources causes data accessibility and processing problem.

For this prototype three main stages have been processed: analysis, design and implementation. Through analysis stage, issues relate to user requirement have been considered such as system usability. Tasks chart also has been designed to show the graphical view of the whole system. There are some designs issues form human computer interaction point of view, such as simple navigation, consistent layout and

help feature which have been considered at this stage. The tools and application used for implementation were MySql database and PHP. These applications have been chosen for their simplicity of implementation and process and ease of run on any server and platform. The prototype is accessible internally and externally.