

**REPUBLIC OF TURKEY
FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND
APPLIED SCIENCE**



**THE PERFORMANCE COMPARISON OF SUPPORT
VECTOR MACHINE CLASSIFICATION KERNEL
FUNCTIONS ON MEDICAL DATABASES**

HARDI MOHAMMED TALABANI

**Master Thesis
Department of Software Engineering
Supervisor: Prof. Dr. Engin AVCI**

JANUARY - 2019

REPUBLIC OF TURKEY
FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE


THE PERFORMANCE COMPARISON OF SUPPORT VECTOR MACHINE
CLASSIFICATION KERNEL FUNCTIONS ON MEDICAL DATABASES

MASTER THESIS

HARDI MOHAMMED TALABANI
(162 137 115)

Submission date: 26 December 2018

Thesis Presentation Date: 14 January 2019

Thesis Supervisor : Prof. Dr. Engin AVCI (F. U.) 

Other Jury Members : Prof. Dr. Asaf VAROL (F. U.)

: Assis. Prof. Dr. Ahmet Arif Aydın (I. U.) 

JANUARY - 2019

DEDICATIONS

First of all, I would like thank God who gave me the patience and ability to complete this humble work.

I would like thank my father, who gave me support in all stages of my life. He has always been considered a source of my trust and is my ideal for dedication and help.

I would like thank my mother, whom I consider a source of strength in my successes and the source of compassion and encouragement in the face of the difficulties of my life.

The duty of gratitude calls on me to extend my thanks and appreciation to the best guide and my supervisor, Professor Dr. Engin AVCI, who I was privileged to have supervise this study.

Last but not least, I would like to extend my thanks and gratitude to all of my colleagues and friends who have contributed even just one word or one piece of scientific advice in the support of this work.

Sincerely

Hardi Mohammed TALABANI

Elaziğ, 2019

TABLE OF CONTENTS

	Page No
DEDICATIONS.....	II
TABLE OF CONTENTS	III
ABSTRACT	V
ÖZET	VI
LIST OF FIGURES.....	VII
LIST OF TABLES.....	VIII
ABBREVIATIONS.....	IX
1. INTRODUCTION	1
2. DATA MINING AND METHODS	3
2.1. Prediction Method	4
2.2. Description Method	4
2.3. The Main Aims of Data Mining	5
3. KNOWLEDGE DISCOVERY IN DATABASE (KDD)	6
3.1. Data Discovery	6
3.2. Data Cleaning	6
3.3. Data Integration	6
3.4. Data Selection	6
3.5. Data Transformation	7
3.6. Data Mining	7
3.7. Pattern Evaluation.....	7
3.8. Knowledge Presentation	7
4. SUPPORT VECTOR MACHINE.....	8
4.1. Linearly Separable	10
4.2. Cross Validation	11
4.3. Kernel Functions.....	12
5. USED MEDICAL DATASETS	15
5.1. Autistic Children Dataset.....	16
5.2. Autistic Adolescent Dataset.....	18
5.3. Chronic Kidney Disease Dataset	19
5.4. Wart Treatment Datasets	21

5.4.1.	Cryotherapy	22
5.4.2.	Immunotherapy	23
6.	APPLICATIONS OF SVM CLASSIFIERS FOR PREDICTION OF USED MEDICAL DATASETS	25
6.1.	Impacts of Tuning Parameters on the NPK Classification Performance.....	26
6.2.	Impacts of Tuning Parameters on the PK Classification Performance.....	28
6.3.	Impacts of Tuning Parameters on the PUK Classification Performance.....	30
6.4.	Impacts of Tuning Parameters on the RBF Classification Performance	32
7.	CLASSIFICATION PERFORMANCE MEASUREMENTS OF THE USED KERNEL FUNCTIONS.....	35
7.1.	Confusion Matrix.....	35
7.2.	Accuracy	37
7.3.	Precision	37
7.4.	Sensitivity	37
7.5.	F-measure	38
8.	RESULTS AND DISCUSSION	39
8.1.	Confusion Matrixes of the SVM Kernel Functions Classifying the Medical Data .	39
8.1.1.	Confusion matrix outcome for the Disease datasets:.....	40
8.1.2.	Confusion matrix outcome for the treatment datasets:	40
8.2.	Classification Performance Measurements of the SVM Kernel Functions Classifying Medical Data	44
9.	CONCLUSIONS	49
	REFERENCES	50
	CURRICULUM VITAE	57

ABSTRACT

The Performance Comparison of Support Vector Machine Classification Kernel Functions on Medical Databases

In this research, an intelligent system framework was constructed by accurately comparing the classification performance of four different types of support vector machine; this included the SVM algorithm kernel functions (normalised polynomial kernel function (NPK), polynomial kernel function (PK), Pearson VII function-based Universal Kernel function (PUK), and the Radial Basis Function Kernel (RBF). This study used five different types of medical datasets (autistic children, autistic adolescents, chronic kidney failure, cryotherapy and immunotherapy), which differ from one another in terms of the quantity of the data and the medicinal and therapeutic content. The databases were extracted from the University of California Irvine machine learning repository.

The method of tuning the parameters was followed in order to obtain the best performance results for the kernel functions using the Weka workbench tool. We then compared the best result of each kernel with the other kernels in terms of familiar classification standards in the field of data mining, which consisted of the confusion matrix, accuracy, sensitivity, precision and error rate.

Keywords: Data Mining, Support Vector Machine, Kernel Functions, Medical Data Mining. SVM Parameters Selecting.

ÖZET

Destek Vektör Makine Sınıflandırma Çekirdeği Fonksiyonlarının Tıbbi Veri Setindeki Performans Karşılaştırması

Bu araştırmada, dört farklı destek vektörü makinesi çekirdek fonksiyonunun (normalleştirilmiş polinom çekirdek fonksiyonu (NPK), polinom çekirdek fonksiyonu (PK), Pearson VII fonksiyon tabanlı Evrensel Çekirdek fonksiyonu (PUK) ve Radyal Temel Fonksiyon Çekirdeği (RBF) doğru sınıflandırma performansları, 5 farklı tıbbi veri seti (otistik çocuklar, otistik ergenler, kronik böbrek yetmezliği, kriyoterapi ve immünoterapi) kullanılarak karşılaştırılmıştır. Bu tez çalışmasında kullanılan tıbbi veri tabanlarının tamamı, açık erişimli UC Irvine Machine Learning Repository' den elde edilmiştir. Bu tez çalışmasında yapılan bütün çalışmalar, WEKA paket program ortamında gerçekleştirilmiştir. Daha sonra, bu tez çalışmasından elde edilen doğru tanıma sonuçları, karmaşıklık matrisi, doğruluk, hassasiyet ve hata oranından oluşan veri madenciliği alanındaki bilinen sınıflandırma standartları açısından karşılaştırılmıştır.

Anahtar Kelimeler: Veri Madenciliği, destek vektör makinesi, çekirdek fonksiyonları, tıbbi veri madenciliği. SVM Parametreleri seçimi.

LIST OF FIGURES

	Page No
Figure 2.1. Searching for knowledge in our data.....	4
Figure 3.1. Knowledge Discovery in Database steps	7
Figure 4.1. SVM hyperplanes.....	9
Figure 4.2. The margin of the hyperplane	9
Figure 4.3. Multiplexed separating lines	10
Figure 4.4. Cross-Validation Diagram.....	12
Figure 4.5. Non linearly separable Data	13
Figure 4.6. Change the dimensions of the dataset using kernel tricks.....	14
Figure 5.1. Autistic Child	16
Figure 5.2. Autistic Adolescent	18
Figure 5.3. Chronic Kidney Disease.....	20
Figure 5.4. Types of Warts	22
Figure 5.5. Wart Treatment Using Cryotherapy Method	23
Figure 5.6. Wart Treatment Using the Immunotherapy Method.....	24
Figure 6.1. Effects of tuning parameters of NPK classification performance.....	27
Figure 6.2. Effects of tuning parameters of PK classification performance.....	29
Figure 6.3. Effects of tuning parameters of PUK classification performance.....	31
Figure 6.3. Effects of tuning parameters of RBF classification performance	33

LIST OF TABLES

	Page No
Table 5.1. Autistic Child dataset description	17
Table 5.2. Autistic Adolescent Dataset Description	19
Table 5.3. Chronic Kidney Disease dataset description	21
Table 5.4. Cryotherapy dataset description.....	23
Table 5.5. Immunotherapy Dataset Description	24
Table 6.1. Kernel Parameters.....	25
Table 6.2. The extent of the effect of the change in parameter C in the NPK in data classification	26
Table 6.3. The extent of the effect of the change in parameter C in the PK in data classification	28
Table 6.4. The extent of the effect of the change in parameter C in the PUK in data classification	30
Table 6.5. The extent of the effect of the change in parameter C in the RBF in data classification	32
Table 7.1. Confusion Matrix Outcomes.....	36
Table 7.2. Example of Confusion Matrix	36
Table 8.1. SVM kernel confusion matrixes classifying medical datasets	41
Table 8.2. SVM kernel Performance Measurements classifying Medical Data.....	44
Table 8.3. Kernel Function Error Rates for10-fold Cross Validation.....	46
Table 8.4. Comparisons Between Proposed Methods with Other Methods Using Autistic Children and Autistic Adolescent Datasets	47
Table 8.5. Comparisons Between Proposed Methods with Other Methods Using Chronic Kidney Disease Dataset	47
Table 8.6. Comparisons Between Proposed Methods with Other Methods Using Cryotherapy and Immunotherapy Datasets	48

ABBREVIATIONS

CKD	: Chronic Kidney Disease
DM	: Data mining
KDD	: Knowledge Discovery from Data
KFs	: Kernel Functions
NPK	: Normalized Polynomial Kernel
PK	: Polynomial Kernel
PUK	: Pearson VII function based Universal Kernel
RBF	: Radial Basis Function Kernel
SVM	: Support Vector Machine
SVN	: Support Vector Network

1. INTRODUCTION

The vast diversity of data in the modern world, in all its varieties and fields, and in accordance with the requirements and uses of human beings, is almost endless in its various forms (numbers, texts, pictures, etc.). This includes several dimensions, such as medical, electronic and so on [1]. This growing data collection has reduced or answered important questions to drive research toward the development of data mining techniques. The purpose of developing these techniques is to find information within a large range of data. Although data mining is a new field of medical informatics, its application may be one of the most successful methods of data analysis and thus allow for a better understanding of medical science [2]. The main idea in data mining is to extract information that is hidden in the any type of datasets by applying accurate classification techniques. Analyses of this vast amount of data have become more difficult and therefore more interesting to many researchers, especially in the areas of machine learning, data mining, and intelligent systems [3]. Like the other data types in different fields, medical data is characterised by its multiple forms (numbers, images, texts, recordings, etc.). In addition, most of the time, this type of data is objectively incomplete, such as where the information about a disease is incomplete in terms of the disease description. The reason for this is that at the phase of processing and data collection, negative effects can be caused as a fault in the diagnostic process by the physicians and due to the other factors that can lead to data incompleteness. These factors may increase the complexity of the data analysis that faces the researchers in this field [4]. In the field of machine learning, there are several algorithms applied for the purpose of data analysis and each algorithm has a specific method in this regard, each of which operates differently on a particular type of data according to the formula in which the medical data was formed as mentioned earlier [5]. The analysis of medical data of all kinds (classification, regression, pre-processing) has become a major concern over the past few years, which has become a challenge for many researchers in the fields of machine learning, data mining, and intelligent systems. They often work by choosing the method to obtain the best possible results [6].

One of the professional sectors that have started to benefit from the data mining concept is health care. With the growth in electronic health records more and more facilities are gathering huge amounts of digital patient data, so health care providers and researchers

can use data mining to detect previously unknown cognitive patterns. They can then use this information to build predictive models to improve diagnostic and health care outcomes [7]. One of the common algorithms in the field of machine learning, especially in the data classifications process, is the support vector machine algorithm, which is an important and well-performing tool in data analysis, which is very commonly used by researchers to categorise and recalibrate data types [8]. The support vector machine algorithm is a supervised learning algorithm that is applied to solve problems that can be encountered in both linear and nonlinear data classifications. The mechanism of the support vector machine algorithm is that it draws a line where it separates the data into two sections. Each section is called a class [9]. For example, if we look at the classification function, the goal of drawing this line, known as the decision line, is to enlarge the distance between all points located in the both classes. After completing this process (model), it is easy to model the target class for new cases or problems to come.

The remain sections in this research are as follows: Firstly, in the data mining and methods section, the basic principles of data mining and the basic methods that comprise it were explained. Later, in section Knowledge Discovery in Databases phases mentioned and explained. After that, an explanation of the working mechanism of the support vector machine (SVM) which is the used algorithm with its kernel functions applied in this research in order to data classifying in section of support vector machine (SVM). Next, describe all types of medical data used in this research in the used medical datasets section. Then, in the Application section, the changes in the classification performance of the classifiers were shown by changing the value parameters. Then mention the performance measurements for evaluating the performance of the SVM kernel functions. Finally, discuss all the results obtained in this research and compare them with the results obtained in other research and other works in the same field.

2. DATA MINING AND METHODS

Scientific progress and the widespread use of technology in various aspects of daily life has increased the ability to generate and collect data quickly in this era. This has contributed to the computerisation of most of the work done in the field of science, including the services offered daily around the world [10]. Technological advances have led to the emergence of new types of data such as text, images, video, multi-tasking systems and the Internet, which all contain vast amounts of data in all of its forms. All of this has led to an unprecedented inflation in the amount of data that is stored daily, demonstrating the urgent need for new technologies and intelligent tools that can help to transform this vast amount of data into useful information and knowledge. Cue the introduction of data mining tools [11].

Data mining aims to extract the information that is hidden in large data blocks, which is a modern technology that has strongly established itself in the information era. Its use provides companies and institutions in all fields, both civil and governmental, with the ability to explore and focus on the most important information contained within the large data blocks [12]. Data mining techniques also focus on sensing, building future predictions, and exploring patterns, relationships, behaviour and trends, which allows for the assessment of correct decisions, making them in a timely manner, developing appropriate solutions to problems and promoting planning, development and modernisation in all areas.

Data mining techniques can answer many questions and in record time, especially the kinds of questions that were difficult to answer, if not impossible, using classical statistical analysis techniques, which took time because of many analysis procedures [13]. The science of analysis and data mining is a relatively modern science and it is an extension of the science of statistical analysis and the main nerve of business intelligence in all its forms [14]. The science of data mining has emerged as a natural result of the great developments in the field of information systems and the large inflation in information, the widespread use of information systems and the accumulation of large amounts of data that has become common daily in many fields. This has led to the urgent need to answer many questions and to explore the knowledge, estimates and future predictions available [15].



Figure 2.1. Searching for knowledge in our data [16]

Data analysis and data mining is one of the priorities of the work in planning departments in companies and institutions globally. It is the best tool for the higher levels of management to use, particularly those that aspire to succeed and who seek to ensure their continuation strategically. This is because it provides the possibility to produce the real knowledge hidden in the large data blocks from the activities of each company or institution.

2.1. Prediction Method

Use the available data and apply specific techniques to achieve and predict successful future values [25]

2.2. Description Method

The process of describing the available data and their classification according to their existence and the relations between them through the simulation of human connections

(human interpretable). In other words, I take the links through natural interactions in order to explain the data [26].

2.3. The Main Aims of Data Mining

- In order to explain some phenomena. For example, what is the reason behind the growing phenomenon of smoking in the Arab countries?
- In order to ascertain a theory. For example, to check the theory that large families care more about health insurance than small families [27].
- In order to analyse the data behind new and unexpected relationships. Example:
How will public spending be if it is associated with fraudulent credit cards?

3. KNOWLEDGE DISCOVERY IN DATABASE (KDD)

Knowledge Discovery in Database (KDD) is not an easy process. Issues may happen during the process of data is collecting and managing, but rather, it extends to analysis and predicting what will happen in the future. Data mining is a part of knowledge discovery, and this process is the most comprehensive [17]. The Knowledge Discovery process includes eight steps as follows:

3.1. Data Discovery

At this stage, the data to be detected is collected. This includes detection, identification and the characterisation of available data [18].

3.2. Data Cleaning

This step is the data purification step, making it suitable and ready for operations. In other words, removing the irrelevant noise. In addition, conflicting data and inconsistent data is also deleted [19]. in another words, it is a level where trivial and clamor information are drive out from the cumulating

3.3. Data Integration

In this step, similar and relevant data are collected from multiple data sources and merged [20]. In another words, several information sources that orderly heterogeneous might be combined in a typical source at this level.

3.4. Data Selection

At this step, the appropriate data is selected and retrieved from the available data set [20]. In another words, several information sources that orderly heterogeneous might be combined in a typical source at this level.

3.5. Data Transformation

This step involves converting the data into custom templates suitable for search and retrieval procedures through a summary of the achievement or aggregation processes [21].

3.6. Data Mining

This is the stage of using intelligent methods to extract data patterns and useful models [22]. Data mining is a fundamental procedure where intense strategies are united to concentrate information designs. It is the fundamental step in which active strategies are united to concentrate designs.

3.7. Pattern Evaluation

At this stage, the really important patterns that represent the knowledge base are identified to use in important metrics or standards [22].

3.8. Knowledge Presentation

This is the final stage of knowledge discovery, which is the stage that the user interacted with the obtained results. This basic stage uses a visual method to help the user to understand and interpret the data extraction results [23].

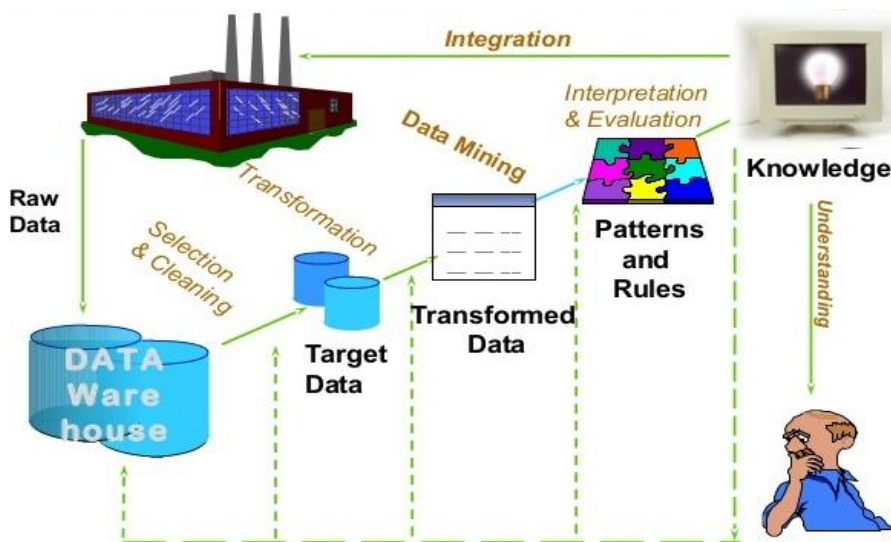


Figure 3.1. Knowledge Discovery in Database steps [24]

4. SUPPORT VECTOR MACHINE

Historically, the support vector machine (SVM) algorithm was invented in 1963 by Vladimir Vabnik and Alexey Shervonenikis. The approved algorithm currently used was formulated in 1993 by Corrina Cortes and Vabnik and published in 1995 [28]. The basic and fundamental meaning of the support vector machine (SVM) algorithm, which is also known as the support vector network (SVN) is that it analyses data through the classification and regression processes. The factor that makes the SVM algorithm prevalent in the data mining and machine learning fields is that it is not difficult in terms of understanding and application. The expression "support vector machine" is an astounding name for an algorithm used for analysis and prediction [29]. In fact, this name can be considered aptly associated with its powerful capabilities, which have achieved great success, especially in the data classification process. working mechanism which the SVM algorithm is based on is to delimit a boundary for a similar point area (belonging to a particular class) [30]. At the time when a limit is drawn (on the training sample), for any new points (test sample) to be classified, it is necessary to check whether the sample (test sample) falls within the boundary or not.

The important factor of the SVM algorithm, which makes it perform better than most other algorithms available in the data classification process, is that at the time of the creation of the boundary, the training data (except for data very close to or located on the boundary) are considered to be redundant, or it can be said that the algorithm does not rely on them to complete its work [31]. All it needs is a core set of points that can help to define and set the boundaries. The reason why these data points are called "support" is because they practically support the boundary. The reason why they are called a "vector" is because each data point is a vector: every line of data consists of the information for a set of attributes [32].

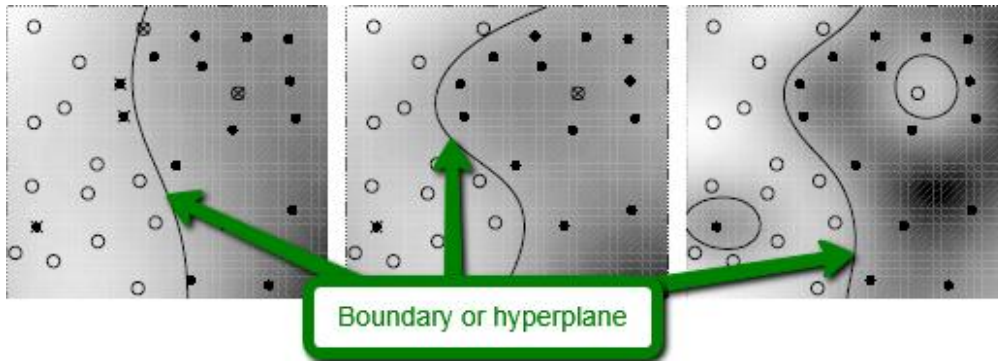


Figure 4.1. SVM hyperplanes [33]

In the simple cases where the data is two dimensions, the hyperplane appears as either a straight line or a curve (as shown in Figure 4.1.), and the data on both sides is in the form of two classes, each differs from the other in terms of its attributes. In the case of three-dimensional data, the hyperplane appears as a form of a complex surface or a plane considering the data to be categorised linearly. However, more than three-dimension data is very difficult to observe. Therefore, the name of the hyperplane is a general name for the boundaries that separates these types of data.

There are a number of hyperplanes that can be observed (as shown in Figure 4.1.). Can any of these hyperplanes be considered better? The correct answer here is that the hyperplane completely and correctly divided the classes [34]. In other words, a hyperplane with an equal distance (maximum distance) between each area of data on both sides is considered to be the best boundary line. This maximum distance between the boundary line and the closest pattern on both sides of the hyperplane is called the margin (as shown in Figure 4.2.).

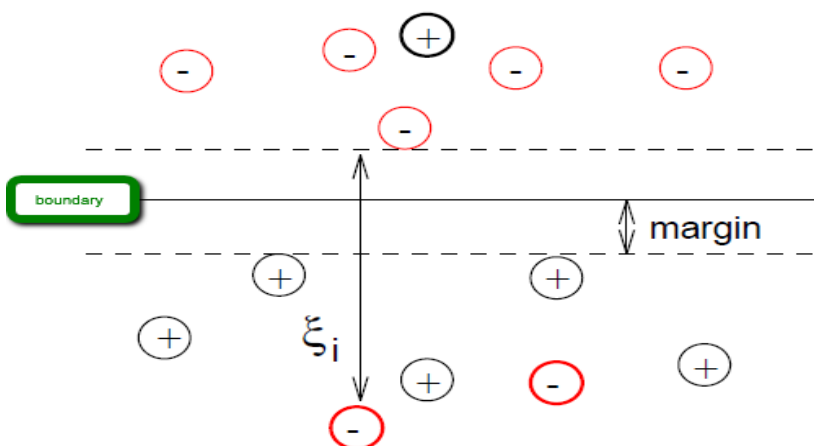


Figure 4.2. The margin of the hyperplane [33]

In the data separation processes, it is not possible to separate the two classes correctly and completely. It is also not possible to ensure that all patterns in a given class are actually part of this class. It is very rare to find data that can be linearly separable and some of the patterns may be located inside the margin (as shown in Figure 4.2.). Thus, the best hyperplane that can be taken into account is the one that contains the minimum patterns (points) possible within the distance of the margin.

4.1. Linearly Separable

As has been illustrated earlier, there are different types of data repositories available. In the case of the two class training datasets that can be separated, there will be more than one separator line (hyperplane). Or more precisely, there will be various separator lines (as shown in Figure 4.3.). All of which could be considered intermediate dividing lines that divide the data to a certain extent. The question that arises here is how to choose the most accurate line between the lines, and depending on which criteria? Comparatively, the separator line located in the space between the data points of the two classes appears to be better and more accurate than the separator line which approaches the data points of one or both classes [35].

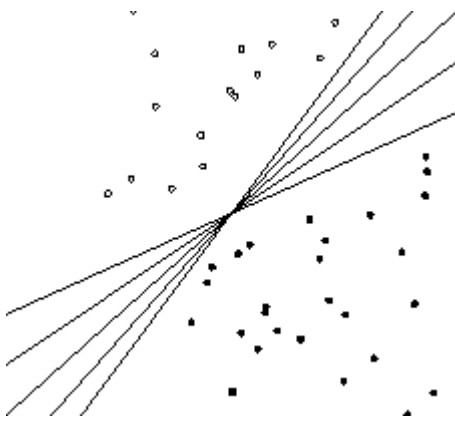


Figure 4.3. Multiplexed separating lines [33]

Several learning algorithms, such as Naive Bayes, select the best separate line based on specific criteria. Perception algorithms choose any of the lines approximately without any standards [36]. The criterion of the SVM algorithm for selecting the hyperplane is to be as far as from any of the data items of both classes as possible. Thus, the margin of the SVM

classifiers can be specified by calculating the gap between the hyperplane and the closest data item.

Following a technique like this means that the decision surface or the hyperplane is not determined by all of the data used, but it instead depends on a very small subset of the data (support vectors). This turn plays a key role in defining the location of the hyperplane (as shown in Figure 4.2.). Therefore, other data (no matter how many), except for the support vectors, plays no role and is not important in the process of choosing the hyperplane [37].

4.2. Cross Validation

One of the standard machine learning techniques used in Weka workbench (the third choice on its Classify panel with three other techniques in the same field but in a different style. The Training set, Percentage Split and Supplied test set are used for assessing the performance of the learning algorithms. Cross validation is a method that actually improves through repetition in order to reduce the differences in the process of estimating the classification algorithm [38].

The idea of cross-validation is as follows: it takes a specific data set and divides it into 10 separate parts for use in training and testing operations. The nine parts will be taken as training data and the tenth (the final) part is treated as the test data. This process continues until each part of the 10 parts is used as both training and testing data. Therefore, no data point is left in the used dataset until it is used 9 times for training and once for testing [39]. In other words, cross validation use periodical process to test and train itself on the rest of the data (as shown in Figure 4.4.).

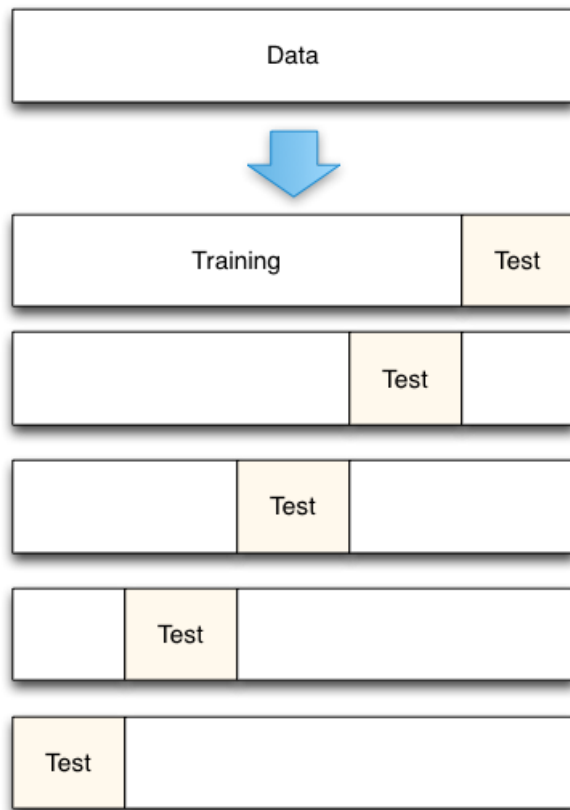


Figure 4.4. Cross-Validation Diagram [40]

Each of these 10 parts is called a fold, meaning that each data set consists of 10 folds and 10 sub-results. The final result is an average of the sub-results. The completion of the primitive division of the data in this form ensures that each partition or fold has obtained a correct percentage of the classes' values.

It is very well known in the field of data analysis that there are many techniques through which the data can be divided into several parts. The reason for the selection of the cross-validation technique is that it reduces the variance in the estimation a lot more than the other techniques [41]. This ensures that we gain the necessary estimations as well as in the performance of the classifier.

4.3. Kernel Functions

In the introduction of this chapter, the mechanics of the SVM algorithm and the advantages that it possesses in the context of the analysis of data has been explained and described as well as the implied parts (hyperplane, margin, support vectors and linearly

separable) on which it relies in the performance of data classification processes and the strategies of their use.

Another important expression called kernel functions is considered to be one of the more interesting techniques used to solve classification problems in the framework of the SVM algorithm. This play a key role in finding best results in the data classification operations [42]. This is especially so in relatively complex datasets that cannot be classified linearly, which will be addressed in this research, in addition to their effect on the data classification processes and the comparison of their classification performance on medical datasets. The time that it takes each type of used kernel to classify the datasets, which will be referred to and explained in detail in the results section, is also important. Therefore, as a general explanation, a kernel is a set of mathematical equations (that take any type of data as the input and transform it into its desired formation.) used by the SVM algorithm [43].

The study of the basic functions of data analysis (e.g. classification, regression, clustering etc.) is the primary objective of the principle of pattern recognition (pattern analysis) [44]. In order to perform one of these tasks by applying one of the purpose-oriented algorithms, the data must be converted from its original raw representation into a representation known as the feature vector representation.

In other words, kernel function is a technical trick that transforms the data that needs to be classified (which cannot be classified by a single straight line) by converting it from one-dimensional data into two-dimensional data [45]. For example, if we have a simple data set consisting of yellow and red balls distributed randomly on a particular surface that cannot be clearly split (by a straight line) into two classes, each class contains yellow balls and red balls evenly (as shown in Figure 4.5).

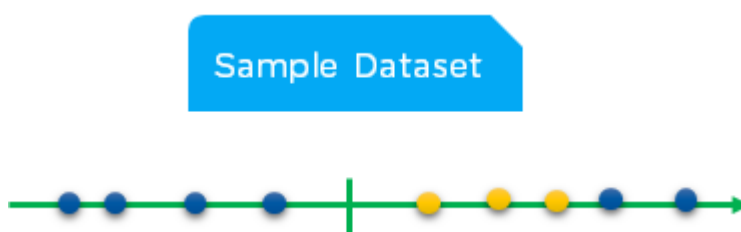


Figure 4.5. Non linearly separable Data [46]

Therefore, an effective way to solve such problems is to convert the data from a one-dimensional view to a two-dimensional view using the kernel functions (as shown in Figure 4.6).

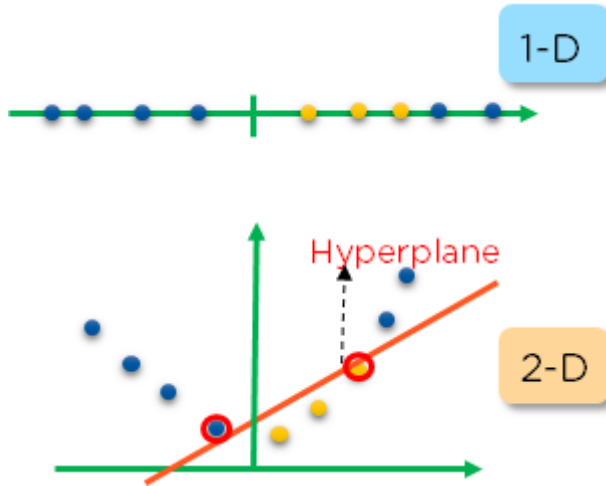


Figure 4.6. Change the dimensions of the dataset using kernel tricks [46]

In this study, four kernels were used for the SVM algorithm [47]:

- **Normalized Polynomial Kernel (NPK)**

$$k(x, y) = (x \cdot y + c) \quad 4.1$$

- **Polynomial Kernel (PK)**

$$k(x, y) = \frac{(x \cdot y + c)^d}{\sqrt{x^{T+1} + y^{T+1}}} \quad 4.2$$

- **Radial Basis Function Kernel (RBF)**

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \quad 4.3$$

- **Pearson VII function based Universal Kernel (PUK)**

$$k(x, y) = \frac{1}{\left[1 + \left(\frac{\sqrt{\|x - y\|^2 \sqrt{2^{(1/\omega)} - 1}}}{\sigma}\right)^2\right]^\omega} \quad 4.4$$

5. USED MEDICAL DATASETS

Advanced analysis techniques and data mining are among the most important tools used in the medical and pharmaceutical fields, especially in the field of exploration and in the evaluation of the prevailing health conditions, the investigation of the causes of disease and the investigation of pathological behaviour in the community in order to contribute to the development of appropriate medical and health plans and policies. Where medical and health data is available, analytical and prospecting techniques can be used to study, analyse and explore everything that will contribute to improving the overall health status of the community, improving the performance of health institutions and reducing the risk. The development of policies and procedures for health education are important for both the individual and the family [48]. Exploration techniques also assist in the exploration and characterisation of the most common diseases in specific areas, times and conditions, with a view to developing appropriate solutions and preventive measures to reduce the spread of disease. The study of medicines, medical treatments and ways to develop, upgrade them, and raise their efficiency, effectiveness, validity and ability to treat is important.

Public health and health insurance are areas that rely heavily on data mining techniques, and the use of data mining techniques in these areas is due to the increase in the population, the increase in health problems in societies and the accumulation of a huge amount of data. This has led to the urgent need to analyse and explore knowledge that can provide solutions to health problems and establish an intelligent health insurance system based on predictive exploration techniques. This is in order to develop the best plan for a health insurance system that benefits all members of society at the lowest cost. This is easily achievable using the data mining techniques of this era [49].

All of the data used in this research was downloaded from the data repository of the University of California, Irvine (UCI) [50]. The data included information on various diseases, and an attempt to cover a wide range in terms of specialisation, including the data on autism for two age groups (children and adolescents). There was also data on Chronic Kidney diseases, which are considered to be an esoteric disease. Finally, there was the data on two methods (Cryotherapy and Immunotherapy) to treat warts, which is considered to be a skin disease.

5.1. Autistic Children Dataset

This disease is new to the world, or has been newly discovered. In the past, no one had ever heard about children with autism or knew anything about it. Autism causes neurological dysfunctions and a lack of contact with those around the patient. It is noted that the infected child cares about specific things only and does not raise their attention to anything else.

Parents pay attention to the problems of their child around the age of three years old. At first, they may think that the reason is the child's small age or that he needs more time to learn and mature his mind. But when parents see other children at their child's age or even younger learning more skills and speaking better, it will make them feel suspicious of the child's actions and they will try to ascertain whether the behaviours are normal or caused by a disease. Parents should be more aware of all of the diseases that can affect children at a young age so then they can deal with the child and resort to treatment if necessary [51]. One of the most common diseases that have begun to spread among children in the current era is autism.



Figure 5.1. Autistic Child [52]

When the symptoms of autism appear at first, parents may not notice them because they are minor symptoms. With time, the symptoms become more difficult and the child becomes more difficult to deal with in turn. However, the sooner the parents discover the symptoms of autism, the better the child will be treated [53]. The child will face many

difficulties in his or her life when they grow up and they will need a lot of help from others when they want to do anything. Autism affects male children more than female children at a rate of three or four times. It is potentially a genetic disorder, so another family member's autism can cause the transmission of the disease through it being heredity. It can be due to a problem in the formation of the brain and nervous system. It may also be due to a spontaneous mutation.

The data used in this study, which is relevant information on the disease, consists of 292 records (children from different countries aged between 4-11 years) and 20 attributes + class (1 Numeric + 19 Nominal) per record [54]. The information on the procedures carried out by the family to check the condition of their child (infected or not) with the disease has been clarified in more detail in Table 5.1.

Table 5.1. Autistic Child dataset description

N	Attribute name	Attribute type	Description
1	A1_Score	Nominal (0,1)	1 st question of behavioural features (AQ-10-Child)
2	A2_Score	Nominal (0,1)	2 st question of behavioural features (AQ-10-Child)
3	A3_Score	Nominal (0,1)	3 st question of behavioural features (AQ-10-Child)
4	A4_Score	Nominal (0,1)	4 st question of behavioural features (AQ-10-Child)
5	A5_Score	Nominal (0,1)	5 st question of behavioural features (AQ-10-Child)
6	A6_Score	Nominal (0,1)	6 st question of behavioural features (AQ-10-Child)
7	A7_Score	Nominal (0,1)	7 st question of behavioural features (AQ-10-Child)
8	A8_Score	Nominal (0,1)	8 st question of behavioural features (AQ-10-Child)
9	A9_Score	Nominal (0,1)	9 st question of behavioural features (AQ-10-Child)
10	A10_Score	Nominal (0,1)	10 st question of behavioural features (AQ-10-Child)
11	Age	Numeric (4-11)	Years
12	Gender	Nominal	Male or female
13	Ethnicity	Nominal	Different types of Ethnicities
14	Jundice	Nominal (yes or no)	Whether the child had jaundice in his/her born
15	Autism	Nominal (yes or no)	Whether One of the family members is infected with PDD
16	Contry_of_res	Nominal	List of countries
17	Used_app_before	Nominal (yes or no)	Whether the screening app has been used
18	Result	Numeric	Screening app result
19	Age_desc	Nominal (4-11 years)	Screening app has four chooses based on age category. 0=toddler, 1=child, 2= adolescent, 3= adult
20	Relation	Nominal	1= parent , 2= relative , 3= self , 4= health care, 5= other
21	Class/ASD	Nominal (yes or no)	Whether the child has autism or not

5.2. Autistic Adolescent Dataset

One's teenage years certainly carry more feelings of tension, anxiety, and confusion for any adolescent, more so if he or she has autism disorder. It is the nature of adolescence that it varies from individual to individual, depending on the differences in each person, and from one geographic environment to another. It also varies according to the culture of the society in which the teenager lives.

Therefore, we have a male or female who has reached puberty and who has feelings that they cannot easily express, such as sexual feelings and functions. However, they are often unable to know the socially acceptable behaviour or method of expression. They will not know how to satisfy their psychological desires in the presence of problems and obstacles related to the nature of their disorder such as limited language, poor communication, social embarrassment, hypersensitivity to failures in time management, and a loss of control of their emotions [55].



Figure 5.2. Autistic Adolescent [56]

The data used in this study, which is the relevant information on this disease, consists of 104 patient records (adolescents from different countries aged between 12-16 years) and 20 attributes + class (2 Numeric + 19 Nominal) per record [57]. This is in addition to the

information about the procedures carried out by the family to check the condition of their child (infected or not) and other information. This has been clarified in more detail in Table 5.2.

Table 5.2. Autistic Adolescent Dataset Description

N	Attribute Name	Attribute Type	Description
1	A1_Score	Nominal (0,1)	1st question of behavioural features (AQ-10-Child)
2	A2_Score	Nominal (0,1)	2st question of behavioural features (AQ-10-Child)
3	A3_Score	Nominal (0,1)	3st question of behavioural features (AQ-10-Child)
4	A4_Score	Nominal (0,1)	4st question of behavioural features (AQ-10-Child)
5	A5_Score	Nominal (0,1)	5st question of behavioural features (AQ-10-Child)
6	A6_Score	Nominal (0,1)	6st question of behavioural features (AQ-10-Child)
7	A7_Score	Nominal (0,1)	7st question of behavioural features (AQ-10-Child)
8	A8_Score	Nominal (0,1)	8st question of behavioural features (AQ-10-Child)
9	A9_Score	Nominal (0,1)	9st question of behavioural features (AQ-10-Child)
10	A10_Score	Nominal (0,1)	10st question of behavioural features (AQ-10-Child)
11	Age	Numeric (12-16)	Years
12	Gender	Nominal	Male or female
13	Ethnicity	Nominal	Different types of Ethnicities
14	Jundice	Nominal (yes or no)	Whether the adolescent had jaundice in his/her born
15	Autism	Nominal (yes or no)	Whether One of the family members is infected with PDD
16	Country_of_res	Nominal	List of countries
17	Used_app_before	Nominal (yes or no)	Whether the screening app has been used
18	Result	Numeric	Screening app result
19	Age_desc	Nominal (12-16 years)	Screening app has four chooses based on age category. 0=toddler, 1=child, 2= adolescent, 3= adult
20	Relation	Nominal	1= parent , 2= relative , 3= self , 4= health care, 5= other
21	Class/ASD	Nominal (yes or no)	Whether the adolescent has autism or not

5.3. Chronic Kidney Disease Dataset

Chronic kidney disease is a condition where kidney damage occurs due to several possible causes, leading to the gradual and permanent loss of kidney function over time. The main causes of chronic kidney disease (CKD) include diabetes, hypertension, and obesity.

Other conditions that can cause chronic kidney disease include glomerulonephritis and genetic diseases, such as polycystic kidney disease.

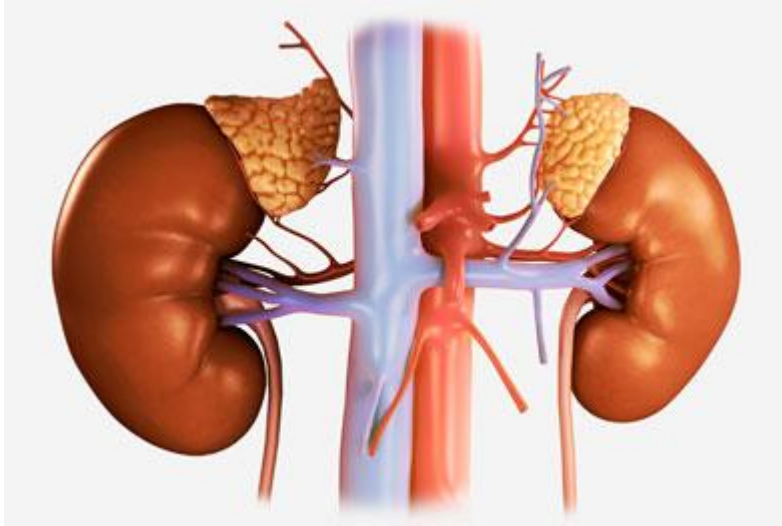


Figure 5.3. Chronic Kidney Disease [58]

It occurs when the kidneys are severely damaged. They cannot purify toxins from the blood and put the waste where it should go, leading to the accumulation of toxins in the body. This causes complications that affect human health. The kidneys are made up of nephrons, which make up the functional structural unit of the kidney [59]. Chronic kidney disease targets these units and affects them gradually over months and years.

The data used in this study is relevant information on this disease and consists of 400 records (individual aged between 2-90 years) and 25 attributes (11 numeric, 14 nominal) per record [60]. This has been clarified in more detail in Table 5.3.

Table 5.3. Chronic Kidney Disease dataset description

N	Attribute Name	Value Range	Description
1	Age	Numeric (2, ..., 90)	Age
2	Bp	Numeric (50, ..., 180)	blood pressure
3	Sg	Nominal (1.005,1.010,1.015,1.020,1.025)	specific gravity
4	Al	Nominal (0,1,2,3,4,5)	Albumin
5	Su	Nominal (0,1,2,3,4,5)	Sugar
6	Rbc	Nominal (2.1, ..., 8)	red blood cells
7	Pc	Nominal (normal,abnormal)	pus cell
8	Pcc	Nominal (present,notpresent)	pus cell clumps
9	Ba	Nominal (present,notpresent)	Bacteria
10	Bgr	Numeric (22, ..., 490)	blood glucose random
11	Bu	Numeric (1.5, ..., 391)	blood urea
12	Sc	Numeric (0.4, ..., 76)	serum creatinine
13	Sod	Numeric (4.5, ..., 163)	Sodium
14	Pot	Numeric (2.5, ..., 47)	Potassium
15	hemo	Numeric (3.1, ..., 17.8)	Haemoglobin
16	Pcv	Numeric (9, ..., 54)	packed cell volume
17	Wc	Numeric (2200, ..., 26400)	white blood cell count
18	Rc	Numeric (2.1, ..., 8)	red blood cell count
19	Htn	Nominal (yes, no)	Hypertension
20	Dm	Nominal (yes, no)	diabetes mellitus
21	Cad	Nominal (yes, no)	coronary artery disease
22	appet	Nominal (good,poor)	Appetite
23	Pe	Nominal (yes, no)	pedal oedema
24	Ane	Nominal (yes, no)	Anaemia
25	class	Nominal (ckd,notckd)	Class

5.4. Wart Treatment Datasets

A wart is a rough, non-painful jagged bump that usually appears on the fingers or hands, individually or as a group. Warts are common in children and young adults. A wart is an abnormal growth of the skin caused by the Human Papillomavirus. The virus has more than 60 strains, some of which cause skin warts by stimulating the growth of the outer layers of the skin. In most cases, warts appear on the fingers, near the fingernails or on the back of the hand.



Figure 5.4. Types of Warts [61]

Specific types of virus infect the genitals and facilitate the transmission of infection to the injured or cracked skin for easy access to the skin layers. The wart virus is passed from person to person through touching or by touching the personal tools of the casualty. The virus enters the body through surface skin lesions [62]. Generally, the two types of dataset used for this disease include the two most common methods for treating this skin disease: Cryotherapy and Immunotherapy.

5.4.1. Cryotherapy

Cryotherapy is performed in doctor's office by placing liquid nitrogen on the wart, by spraying it or using a cotton swab. Treatment with nitrogen can temporarily cause warts around the wart itself, although the dead tissue disappears in about a week. Cryotherapy can stimulate the immune system to fight the viral strains. It is worth mentioning that the treatment of cooling may be painful and may require the doctor to anesthetise the area before the start of the treatment. The treatment may require several sessions repeated weekly, up to nearly two weeks until the disappearance of the wart(s) [63].



Figure 5.5. Wart Treatment Using Cryotherapy Method [64]

This treatment method dataset consists of the information about 90 records (patients aged from 15 to 67 years) with 7 nominal attributes for each patient. The patients (male and female) in this data were infected with two types of warts (Common and Plantar), and Cryotherapy was used to treat them [65]. This has been clarified in more detail in Table 5.4.

Table 5.4. Cryotherapy dataset description

N	Feature Name	Feature Type	Feature Values
1	Sex (male or female)	Nominal	47 man and 43 women
2	Age	Nominal	15 – 67 (year)
3	Time	Nominal	0 – 12 (month)
4	number of warts	Nominal	1 – 12
5	Type	Nominal	1. Common (54) 2. Plantar (9) 3. Both (27)
6	Area	Nominal	4 – 750 (mm ²)
7	Results of treatment	Nominal	Yes or No

5.4.2. Immunotherapy

Immunotherapy involves several methods, including application of a chemical called Diphenycprone on warts to stimulate immune system to fight them. After application, this can cause a slight allergic reaction, causing wart's disappearance later [66].



Figure 5.6. Wart Treatment Using the Immunotherapy Method [67]

This dataset is kind of high imbalance dataset ((data points number is not evenly distributed over the classes). This treatment method dataset consists of the information about 90 records (patients aged from 15 to 65 years) with 8 nominal attributes for each patient. Patients (male and female) in this dataset were infected with two types of warts (Common and Plantar) and the method of Immunotherapy was used to treat them [68]. This has been clarified in more detail in Table 5.5.

Table 5.5. Immunotherapy Dataset Description

N	Feature Name	Feature Type	Feature Values
1	Sex (male or female)	Nominal	41 man and 49 women
2	Age	Nominal	15 – 56 (year)
3	Time	Nominal	0 – 12 (month)
4	number of warts	Nominal	1 – 19
5	Type	Nominal	1.Common (54) 2.Plantar (9) 3.Both (27)
6	Area	Nominal	6 – 900 (mm ²)
7	induration diameter	Nominal	5 – 70
8	Result of treatment	Nominal	Yes or No

6. APPLICATIONS OF SVM CLASSIFIERS FOR PREDICTION OF USED MEDICAL DATASETS

In the Weka workbench, when using the SVM algorithm, it can be noticed that each of the SVM kernel functions has special parameters. When the parameters are changed and modified gradually, the performance of the used kernels is significantly affected. In order to obtain the best model of the SVM algorithm in terms of accuracy in their classification, it is necessary to replicate the parameters gradually. Each kernel has different types of parameters, as has been clarified in more detail in Table 6.1.

Table 6.1. Kernel Parameters

Kernel Types	Parameters
NPK	1. C 2. Exponent
PK	1. C 2. Exponent
PUK	1. C 2. Omega 3. Sigma
RBF	1. C 2. Gamma

Parameter C directs the working mechanism of the SVM algorithm to the following direction; to what extent do you want to avoid misclassification through each training process based on the margin between the hyperplane and the nearest data points to it. When choosing a large value for parameter C, a hyperplane will be chosen with a small margin so then in most cases, the percentage of misclassification is big. Thus, the occurrence percentage of over-fitting will be significant. On the other hand, if a small value is specified for parameter C, then a large margin hyperplane will be chosen even if there are some misclassifications, thus avoiding an over-fitting case. This description is confirmed by our work, in tuning the value of parameter C in trying to obtain the most accurate percentage of the performance of the kernel in the data classification process [69]. For each kernel, the value of parameter x has been changed (adjusted +10) in parallel with the other variables (gamma (+0.01), (omega, and sigma) (+1)). This change continued for all types of data used in this research. This repeated process showed that when the large values of the parameters

has been reached, and then the performance of the kernels will decrease in the accuracy of its classification. In contrast, in all cases, the performance of the kernel is at its most accurate when the values of the parameters are small.

6.1. Impacts of Tuning Parameters on the NPK Classification Performance

When applying the NPK function as part of classifying all types of medical data used in this research and when there is the gradual increase in the variables (C, Exponent) times each (C by 10 and Exponent by 1), there is a remarkable change in the accuracy of the classification performance of the kernel in the process of the data classification. This has been clarified in more detail in Table 6.2.

Table 6.2. The extent of the effect of the change in parameter C in the NPK in data classification

NPK Parameters		Autistic Children Dataset	Autistic Adolescent Dataset	Chronic Kidney Disease Dataset	Cryotherapy Dataset	Immunotherapy Dataset
C	Exponent	Accuracy %	Accuracy %	Accuracy %	Accuracy %	Accuracy %
1	0	51.7123	60.5769	62.5	53.3333	78.8889
10	2	94.863	91.3462	97.75	95.5556	78.8889
20	3	95.5479	95.1923	98.5	95.5556	78.8889
30	4	94.863	97.1154	98.75	95.5556	78.8889
40	5	93.4932	97.1154	98.5	88.8889	78.8889
50	6	93.1507	95.1923	98.25	81.1111	78.8889
60	7	91.7808	95.1923	98.75	74.4444	78.8889
70	8	91.0959	96.1538	98.75	73.3333	78.8889
80	9	90.0685	92.3077	98.75	67.7778	78.8889
90	10	89.0411	93.2692	99.25	66.6667	78.8889
100	11	89.0411	93.2692	99.5	64.4444	78.8889
110	12	88.0137	93.2692	99.5	63.3333	78.8889
120	13	87.6712	93.2692	99.5	60	78.8889
130	14	86.3014	92.3077	99.5	60	78.8889
140	15	84.589	92.3077	99.75	57.7778	78.8889
150	16	84.589	91.3462	99.75	56.6667	78.8889
160	17	83.9041	79.8077	99.75	56.6667	78.8889
170	18	82.8767	67.3077	99.75	56.6667	78.8889
180	19	81.8493	64.4231	99.75	55.5556	78.8889
190	20	81.8493	62.5	99.75	55.5556	78.8889
200	21	80.137	61.5385	99.75	55.5556	78.8889
210	22	78.7671	60.5769	99.75	55.5556	78.8889
220	23	78.7671	60.5769	99.75	55.5556	78.8889

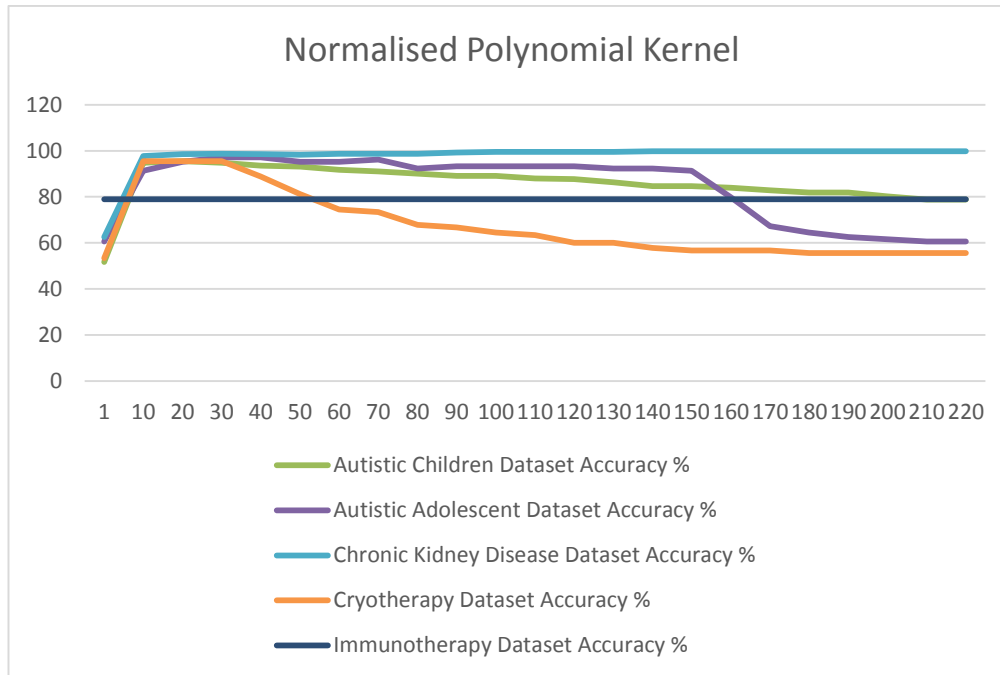


Figure 6.1. Effects of tuning parameters of NPK classification performance

As shown in Table 6.2. There is a disparity in the NPK performance in the classification of the medical data from one type to another depending on two main factors; increasing the value of the parameters and the type of dataset used. In order to provide a summary of the contents of Table 6.2 to measure the effectiveness of these two factors, we divided the classification performance of NPK (from the best to the worse) into five stages in terms of performance accuracy depending on the type of the dataset used as follows.

- The first best performance of NPK was in classifying the Chronic Kidney Disease Dataset, which reached 99.75% accuracy when ($C= 140$, Exponent = 15). The worst classification performance of NPK for the same data set is 62.5% accuracy when ($C= 1$, Exponent = 0).
- The second best performance of NPK was in classifying the Autistic Adolescent Dataset, which reached 97.1154 % accuracy when ($C= 30$, Exponent = 4). The worst classification performance of NPK for the same data set is 60.5769 % accuracy when ($C= 1$, Exponent = 0).
- The third best performance of NPK was in classifying the Cryotherapy Dataset, which reached 95.5556 % accuracy when ($C= 10$, Exponent = 2). The worst classification performance of NPK for the same data set is 53.3333 % accuracy when ($C= 1$, Exponent = 0).

- The fourth best performance of NPK was in classifying the Autistic Children Dataset, which reached 95.5479 % accuracy when (C = 20, Exponent = 3). The worst performance of NPK for the same data set is 51.7123 % accuracy when (C= 1, Exponent = 0).
- Finally, the fifth best performance of NPK was in classifying the Immunotherapy Dataset which reached 78.8889% accuracy. This remained unchanged by changing the parameters because of the different distribution of data points of the immunotherapy dataset classes (imbalance dataset).

6.2. Impacts of Tuning Parameters on the PK Classification Performance

When applying the PK function for classifying all of the types of medical data used in this research and the gradual increase of the variables (C by 10 and Exponent by 1), there was a remarkable change in the accuracy of the classification performance of the kernel in the process of data classification. This has been clarified in more detail in Table 6.3.

Table 6.3. The extent of the effect of the change in parameter C in the PK in data classification

PK Parameters		Autistic Children Dataset	Autistic Adolescent Dataset	Chronic Kidney Disease Dataset	Cryotherapy Dataset	Immunotherapy Dataset
C	Exponent	Accuracy %	Accuracy %	Accuracy %	Accuracy %	Accuracy %
1	1	100	89.4231	97.75	93.3333	81.1111
10	2	99.3151	92.3077	99	96.6667	78.8889
20	3	98.2877	92.3077	98.75	95.5556	78.8889
30	4	97.6027	93.2692	98.75	93.3333	78.8889
40	5	96.2329	92.3077	98.5	84.4444	78.8889
50	6	94.5205	92.3077	98.5	77.7778	78.8889
60	7	92.1233	90.3846	98.25	67.7778	78.8889
70	8	51.7123	87.5	98.25	67.7778	78.8889
80	9	89.3836	84.6154	98.5	67.7778	78.8889
90	10	86.3014	81.7308	99.25	67.7778	78.8889
100	11	85.6164	78.8462	99.25	61.1111	78.8889
110	12	84.589	74.0385	99.25	57.7778	78.8889
120	13	83.5616	70.1923	98.5	56.6667	78.8889
130	14	81.8493	66.3462	98.5	56.6667	78.8889
140	15	81.5068	65.3846	99	56.6667	78.8889
150	16	80.137	63.4615	98.75	55.5556	78.8889
160	17	79.1096	63.4615	99.25	55.5556	78.8889
170	18	76.7123	63.4615	98.75	55.5556	78.8889
180	19	75.3425	61.5385	98.75	55.5556	78.8889
190	20	73.9726	76.9231	99	55.5556	78.8889
200	21	79.4521	88.4615	97.75	55.5556	78.8889
210	22	88.0137	66.3462	97	55.5556	78.8889
220	23	59.2466	39.4231	62.5	55.5556	78.8889

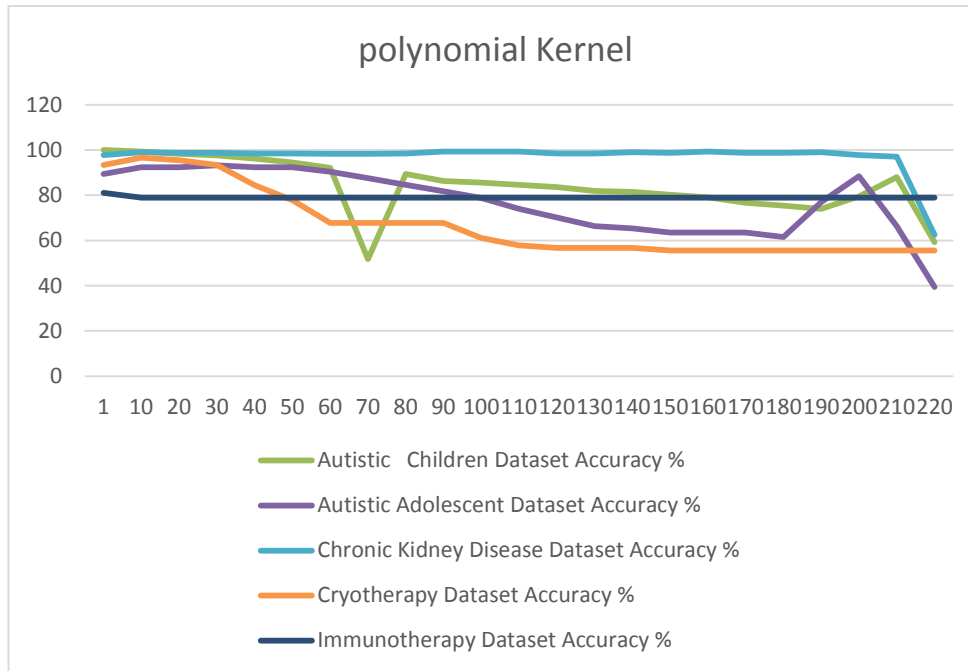


Figure 6.2. Effects of tuning parameters of PK classification performance

As shown in Table 6.3, there is a disparity in the PK performance in relation to the classification of the medical data from one type to another depending on two main factors; an increase in the value of the parameters and the type of dataset used. In order to provide a summary of the contents of Table 6.3 and to measure the effectiveness of the two factors, we divided the classification performance of PK (from the best to the worse) into five stages in terms of performance accuracy depending on the type of the dataset used as follows.

- The first best performance of PK was in classifying the Autistic Children Dataset, which reached 100% accuracy when (C= 1, Exponent = 1). The worst performance of PK for the same data set was 59.2466 % accuracy when (C= 220, Exponent = 23).
- The second best performance of PK was in classifying the Chronic Kidney Disease Dataset, which reached 99.25% accuracy when (C= 90, Exponent = 10). The worst performance of PK for the same data set was 62.5% accuracy when (C= 220, Exponent = 23).
- The third best performance of PK was in classifying the Cryotherapy Dataset, which reached 96.6667% accuracy when (C= 10, Exponent = 2). The worst performance of PK for the same data set was 55.5556 % accuracy when (C= 150, Exponent = 16).
- The fourth best performance of PK was in classifying the Autistic Adolescent Dataset which reached 93.2692% accuracy when (C= 30, Exponent = 4). The worst performance of PK for the same data set was 39.4231% accuracy when (C= 220, Exponent = 23).

- Finally, the fifth best performance of PK was in classifying the Immunotherapy Dataset, which reached 81.1111% accuracy when (C= 1, Exponent = 1). The worst performance of PK for the same data set was 78.8889 % accuracy when (C= 10, Exponent = 2). remained unchanged because immunotherapy is an imbalanced dataset

6.3. Impacts of Tuning Parameters on the PUK Classification Performance

When applying the PUK function while classifying all of the medical data used in this research, the gradual increase of the variables (C by 10 and each of Omega and Sigma by 1) shows a remarkable change in the accuracy of the classification performance of the kernel in the process of the data classification. This has been clarified in more detail in Table 6.4.

Table 6.4. The extent of the effect of the change in parameter C in the PUK in data classification

PUK Parameters			Autistic Children Dataset	Autistic Adolescent Dataset	Chronic Kidney Disease Dataset	Cryotherapy Dataset	Immunotherapy Dataset
C	Omega	Sigma	Accuracy %	Accuracy %	Accuracy %	Accuracy %	Accuracy %
1	1	1	93.1507	74.0385	98.75	81.1111	78.8889
10	2	2	93.4932	92.3077	99	88.8889	78.8889
20	3	3	94.863	95.1923	99	94.4444	78.8889
30	4	4	96.5753	96.1538	98.75	95.5556	77.7778
40	5	5	96.9178	94.2308	98.75	96.6667	75.5556
50	6	6	98.2877	93.2692	98.75	96.6667	74.4444
60	7	7	98.9726	93.2692	98.75	97.7778	74.4444
70	8	8	99.3151	90.3846	98.75	96.6667	77.7778
80	9	9	99.3151	90.3846	98.75	96.6667	80
90	10	10	99.3151	90.3846	98.75	96.6667	78.8889
100	11	11	99.6575	90.3846	98.75	96.6667	78.8889
110	12	12	99.6575	90.3846	98.5	96.6667	78.8889
120	13	13	100	90.3846	98.5	96.6667	78.8889
130	14	14	100	91.3462	98.5	96.6667	78.8889
140	15	15	100	91.3462	98.25	96.6667	78.8889
150	16	16	100	91.3462	98.25	96.6667	78.8889
160	17	17	100	91.3462	98.25	94.4444	78.8889
170	18	18	100	91.3462	98.25	94.4444	81.1111
180	19	19	100	91.3462	98.25	94.4444	81.1111
190	20	20	100	91.3462	98.25	94.4444	81.1111
200	21	21	100	91.3462	98.25	94.4444	81.1111
210	22	22	100	91.3462	98.25	94.4444	81.1111
220	23	23	100	90.3846	98.25	94.4444	81.1111

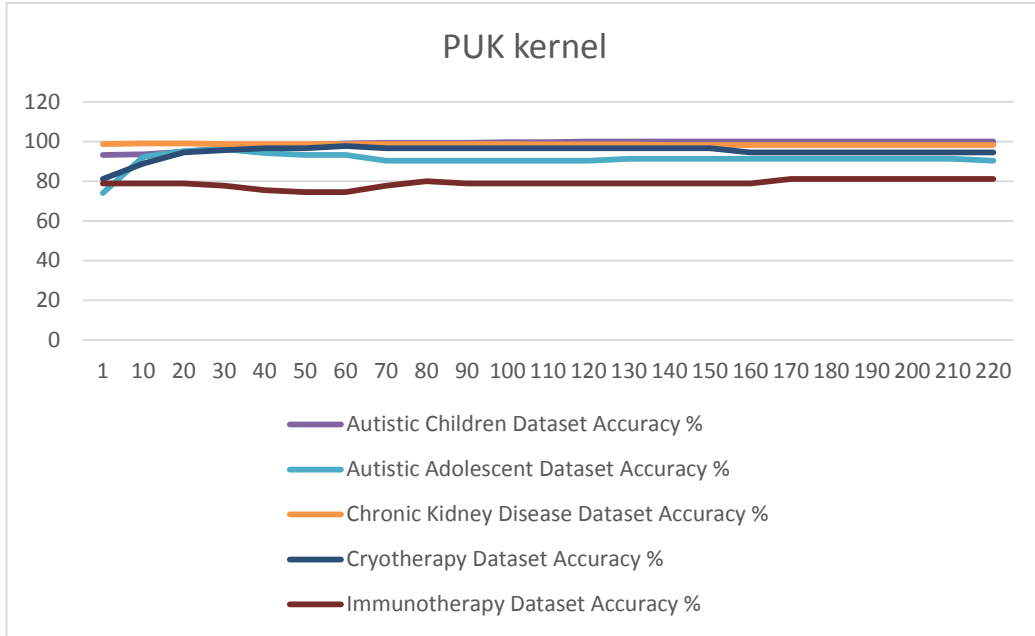


Figure 6.3. Effects of tuning parameters of PUK classification performance

As shown in Table 6.4, there is a disparity in the PUK performance in the classification of medical data from one type to another depending on two main factors; the increase in the value of the parameters and the type of used dataset. In order to provide a summary of the contents of Table 6.4 and to measure the effectiveness of the two factors, we divided the classification performance of PUK (from the best to the worse) into five stages in terms of performance accuracy depending on the type of the dataset used as follows.

- The first best performance of PUK was classifying the Autistic Children Dataset, which reached 100% accuracy when ($C= 120$, $\Omega = 13$, $\Sigma = 13$). The worst classification performance of PUK for the same data set is 93.1507% accuracy when ($C= 1$, $\Omega = 1$, $\Sigma = 1$).
- The second best performance of PUK was classifying the Chronic Kidney Disease Dataset, which reached 99% accuracy when ($C= 10$, $\Omega = 2$, $\Sigma = 2$). The worst classification performance of PUK for the same data set is 98.25% accuracy when ($C= 140$, $\Omega = 15$, $\Sigma = 15$).
- The third best performance of PUK was classifying the Cryotherapy Dataset, which reached 97.7778% accuracy when ($C= 60$, $\Omega = 7$, $\Sigma = 7$). The worst classification performance of PUK for the same data set was 81.1111% accuracy when ($C= 1$, $\Omega = 1$, $\Sigma = 1$).
- The fourth best performance of PUK was classifying the Autistic Adolescent Dataset, which reached 96.1538% accuracy when ($C= 30$, $\Omega = 4$, $\Sigma = 4$).

The worst classification performance of PUK for the same data set was 74.0385% accuracy when (C= 1, Omega = 1, Sigma = 1).

- Finally, the fifth best performance of PUK was classifying the Immunotherapy Dataset, which reached 81.1111% accuracy when (C= 170, Omega = 18, Sigma = 18). The worst classification performance of PUK for the same data set was 78.8889% accuracy when (C= 1, Omega = 1, Sigma = 1).

6.4. Impacts of Tuning Parameters on the RBF Classification Performance

When applying the RBF function for classifying all of the types of medical data used in this research, the gradual increase of the variables (C by 10 and Gamma by 0.01) showed a remarkable change in the accuracy of the classification performance of the kernel in the process of data classification. This has been clarified in more detail in Table 6.5.

Table 6.5. The extent of the effect of the change in parameter C in the RBF in data classification

RBF Parameters		Autistic Children Dataset	Autistic Adolescent Dataset	Chronic Kidney Disease Dataset	Cryotherapy Dataset	Immunotherapy Dataset
C	Gamma	Accuracy %	Accuracy %	Accuracy %	Accuracy %	Accuracy %
1	0.01	96.5753	78.8462	94.25	54.4444	78.8889
10	0.02	96.9178	91.3462	98.25	95.5556	77.7778
20	0.03	99.3151	90.3846	97.75	96.6667	78.8889
30	0.04	99.3151	92.3077	98.25	96.6667	78.8889
40	0.05	98.6301	93.2692	98.5	96.6667	77.7778
50	0.06	98.2877	95.1923	98.75	97.7778	76.6667
60	0.07	98.2877	96.1538	98.75	97.7778	74.4444
70	0.08	98.2877	96.1538	98.75	96.6667	74.4444
80	0.09	98.2877	96.1538	98.5	96.6667	74.4444
90	0.1	97.9452	96.1538	98.5	96.6667	74.4444
100	0.11	97.2603	96.1538	98.5	96.6667	74.4444
110	0.12	96.9178	96.1538	97.75	96.6667	75.5556
120	0.13	96.9178	96.1538	98.25	96.6667	75.5556
130	0.14	96.9178	96.1538	98	96.6667	75.5556
140	0.15	96.9178	96.1538	98.25	96.6667	76.6667
150	0.16	96.5753	96.1538	98.5	96.6667	76.6667
160	0.17	96.5753	96.1538	98.5	95.5556	76.6667
170	0.18	96.2329	96.1538	98.75	95.5556	76.6667
180	0.19	96.2329	96.1538	98.75	95.5556	77.7778
190	0.2	96.2329	96.1538	98.75	95.5556	78.8889
200	0.21	95.8904	97.1154	98.75	95.5556	78.8889
210	0.22	95.8904	97.1154	98.75	95.5556	78.8889
220	0.23	95.5479	97.1154	99.75	95.5556	78.8889

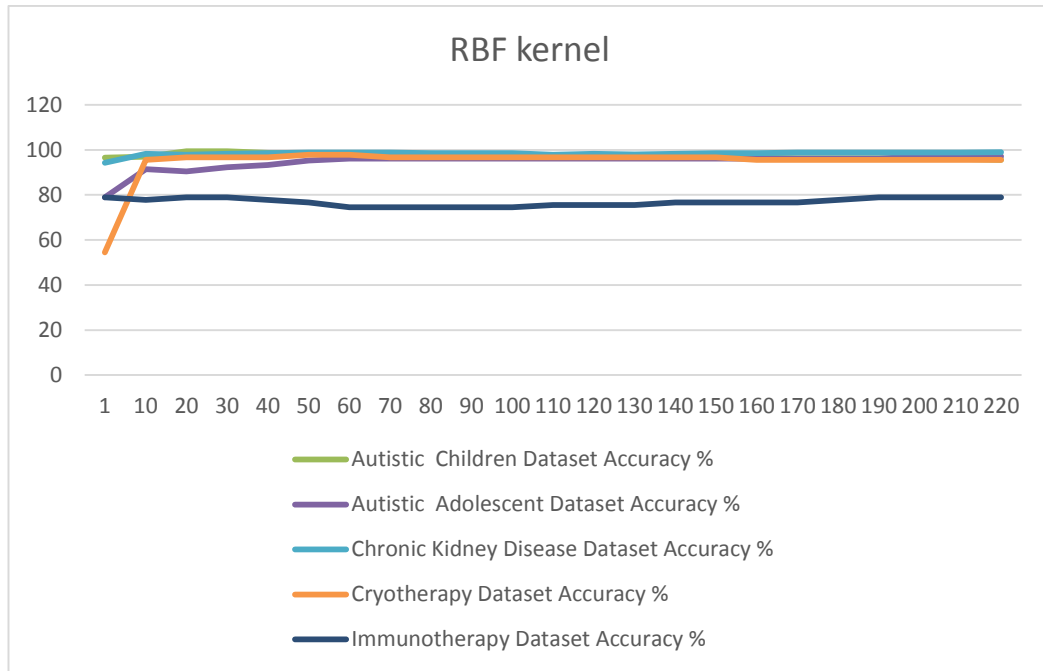


Figure 6.3. Effects of tuning parameters of RBF classification performance

As shown in Table (6.5.), there is a disparity in the RBF performance in the classification of medical data from one type to another depending on two main factors; the increase in the value of the parameters and the type of dataset used. In order to provide a summary of the contents of Table 6.5 and to measure the effectiveness of the two factors, we divided the classification performance of RBF (from the best to the worse) into five stages in terms of the performance accuracy depending on the type of dataset used as follows.

- The first best performance of RBF is in classifying Chronic Kidney Disease Dataset which reached 99.75% accuracy when ($C= 220$, $\text{Gamma} = 0.23$) and the worst classification performance of RBF for the same data set is 94.25% accuracy when ($C= 1$, $\text{Gamma} = 0.01$).
- The second best performance of RBF was classifying Autistic Children Dataset which reached 99.3151% accuracy when ($C= 20$, $\text{Gamma} = 0.03$). The worst classification performance of RBF for the same data set was 95.8904% accuracy when ($C= 200$, $\text{Gamma} = 0.2$).
- The third best performance of RBF was classifying the Cryotherapy Dataset which reached 97.7778% accuracy when ($C= 50$, $\text{Gamma} = 0.06$). The worst classification performance of RBF for the same data set was 54.4444% accuracy when ($C= 1$, $\text{Gamma} = 0.01$).

- The fourth best performance of RBF was classifying the Autistic Adolescent Dataset which reached 97.1154% accuracy when ($C= 200$, $\text{Gamma} = 0.21$). The worst classification performance of RBF for the same data set was 78.8462% accuracy when ($C= 1$, $\text{Gamma} = 0.01$).
- Finally, the fifth best performance of RBF was classifying the Immunotherapy Dataset, which reached 78.8889% accuracy when ($C= 1$, $\text{Gamma} = 0.01$). The worst classification performance of RBF for the same data set was 74.4444% accuracy when ($C= 60$, $\text{Gamma} = 0.07$). The classifiers did not perform well in classifying this dataset because of the different distribution of data points of the immunotherapy dataset classes (imbalance dataset).



7. CLASSIFICATION PERFORMANCE MEASUREMENTS OF THE USED KERNEL FUNCTIONS

Reference should be made to a set of specific criteria that can be used to distinguish the differences in the performance of the algorithms depending on the values obtained based on special calculation equations for each criterion. Through these classification measurements, it is possible to observe the difference in the performance of each algorithm [70].

In this study, four types of SVM kernel functions were applied to five types of medical data. Each of dataset is specific to information on a specific disease (the information for each dataset indicated in detail in section 5). These classification performance measurements were used to measure and evaluate each kernel's effectiveness in the datasets as follows.

7.1. Confusion Matrix

In the field of data mining, especially in the case of data classification (statistical classification), there is a common standard (table) to describe the performance of a specific algorithm in terms of accuracy in the classification process known as the confusion matrix.

In the table confusion matrix (two dimensions, actual and predicted), the classes that represent the contents of the data set have been divided into two parts: the class containing the predicted instances which represents the column of the confusion matrix table and the class containing the actual instances in the dataset which represents the confusion matrix row or vice versa [71]. This has been clarified in more detail in Table 7.1.

In other words, we can describe the confusion matrix as a simple and powerful tool to find out the impact of the classification system that we live in daily. For example, medical examinations for patients (sick or not-sick), feeling emotions about people around us (like them or not) etc.

Table 7.1. Confusion Matrix Outcomes

		Prediction		Total
		Yes	No	
Actual	Yes	TP	FN	TP + FN
	No	FP	TN	FP + TN
Total		TP + FP	FN + TN	

Confusion matrix have four outcomes:

- TP (True Positive)
- FN (False Negative)
- FP (False Positive)
- TN (True Negative)

Since medical data has been used in this research, we will employ the example of a medical examination for patients (sick or not-sick) in an attempt to illustrate the outcomes of the confusion matrix as it has been clarified in more details in Table 7.2.

Table 7.2. Example of Confusion Matrix

		Prediction	
		Sick	Not-Sick
Actual	Sick	TP	FN
	Not-Sick	FP	TN

- TP (True Positive): The number of patients who are actually sick and they have been classified as sick as well.
- FN (False Negative) The number of patients who are actually sick but they have been classified as a not-sick.
- FP (False Positive): The number of patients who are actually not-sick but they have been classified as sick.
- TN (True Negative): The number of patients who are actually not-sick and they have been classified as not-sick as well.

7.2. Accuracy

It is a common metric used for evaluating and measuring the algorithm classification performance, which can be described as the total correctness of the classifier calculated as the sum of the correct classification instance divided by the total number of classification instances [72].

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FN + TN + FP)} \quad 7.1$$

7.3. Precision

It is another common measurement used for evaluating the algorithm classification performance which is known as positive predictive value, representing the number of instances that have been positively classified. Mathematically, it can be described as the number of true positive values divided by the aggregate of the true positive and false positive values [73].

$$\text{Precision} = \frac{(TP)}{(TP + FP)} \quad 7.2$$

7.4. Sensitivity

Sensitivity is the criterion used for measuring the percentage of actual positives that are correctly specified by the used classification algorithm. For example, the proportion of sick individuals who are correctly identified as having the condition. This is called the probability of detection, the recall and true positive rate in some fields [74]. Mathematically, it can be described as the number of true positive values divided by the aggregate of true positive and false negative values.

$$\text{Sensitivity} = \frac{(TP)}{(TP + FN)} \quad 7.3$$

7.5. F-measure

In the F- measure, it is standard that both false positives and false negatives will be taken into account. Thus, it can be said that the F-measure represents the weighted average of precision and sensitivity. In some cases, the F-measure is more helpful than accuracy [75]. Exceptionally, in cases where there are datasets with an uneven class distribution, it is not considered easy to understand like the other measurements of classification performance.

$$\mathbf{F\ Measure} = 2 * \frac{\mathbf{precision*sensitivity}}{\mathbf{precision+sensitivity}} \quad 7.4$$

7.6. Classification Error Rate

Represents the percentage of instances that have been incorrectly classified by the used classifier.

$$\mathbf{Error\ Rate} = 1 - \mathbf{Accuracy} \quad 7.5$$

8. RESULTS AND DISCUSSION

Through our attempts to obtain good results in the process of the classification of the medical data used in this research, we used several types of classifier representing the four types of SVM kernel functions (NPK, PK, PUK, RBF). We found that each kernel has its own and different effect apart from the rest of the other kernels in terms of the familiar standards used to evaluate the performance of an algorithm (accuracy, precision, sensitivity, and F-measure).

In general, we can conclude that most of the kernels used in this study had a relatively good level of performance when classifying the five different types of medical data, which included three types of disease (autistic children, autistic adolescent and chronic kidney disease) and the two types of wart treatment (Cryotherapy and Immunotherapy). Each of the datasets contained a completely different data form and content.

In this section, the best results were selected for each kernel through the results obtained by changing the kernel parameters in section (Applications of SVM Classifiers for Prediction of Used Medical Datasets)

8.1. Confusion Matrixes of the SVM Kernel Functions Classifying the Medical Data

Different medical data was used in terms of the content and type (class type). In terms of content, all of the types of data used have been explained in detail in the "Used Medical Datasets" section. It should be pointed out that the data used in this study was divided into two types of data in terms of the dataset class type: data for three types of disease (autistic children, autistic adolescents, and chronic kidney disease). The classes in this type were (infected and not-infected). The data for the two types of wart treatments were (Cryotherapy and Immunotherapy). The classes in this type were (treated, and not-treated). Thus, the outcomes of the confusion matrix will vary.

8.1.1. Confusion matrix outcome for the Disease datasets:

- TP (True Positive): Individuals who were actually infected with the disease and who were also classified as infected cases.
- FN (False Negative): Individuals who were actually infected with the disease but who were classified as non-infected cases
- FP (False Positive): Individuals who were actually not infected with the disease but who were classified as infected cases.
- TN (True Negative): Individuals who were actually not infected with the disease and who were also classified as not-infected cases.

8.1.2. Confusion matrix outcome for the treatment datasets:

- TP (True Positive): Individuals who have actually been treated using this method and who were also classified as treated cases.
- FN (False Negative): Individuals who have actually been treated using this method but who were classified as non-treated cases.
- FP (False Positive): Individuals who have not actually been treated using this method but who were classified as treated cases
- TN (True Negative): Individuals who have not actually been treated using this method and who were also classified as non-treated cases.

Table 8.1. SVM kernel confusion matrixes classifying medical datasets

Used Datasets	Kernel Functions	Actual	Prediction	
			Yes	No
Children Autistic	NPK	Yes	142	9
		No	4	137
	PK	Yes	151	0
		No	0	141
	PUK	Yes	151	0
		No	0	141
	RBF	Yes	150	1
		No	1	140
Adolescent Autistic	NPK	Yes	63	0
		No	3	38
	PK	Yes	58	5
		No	2	39
	PUK	Yes	62	1
		No	3	38
	RBF	Yes	63	0
		No	3	38
Chronic Kidney Disease	NPK	Yes	249	1
		No	0	150
	PK	Yes	247	3
		No	0	150
	PUK	Yes	246	4
		No	0	150
	RBF	Yes	249	1
		No	0	150
Cryotherapy	NPK	Yes	39	3
		No	1	47
	PK	Yes	40	2
		No	1	47
	PUK	Yes	41	1
		No	1	47
	RBF	Yes	41	1
		No	1	47
Immunotherapy	NPK	Yes	0	19
		No	0	71
	PK	Yes	6	13
		No	4	67
	PUK	Yes	6	13
		No	4	67
	RBF	Yes	0	19
		No	0	71

As shown in Table 8.1, in the process of the classification of the Autistic Children dataset, it was found that PK and NPK had the best performance because the distributions and equations of PK and NPK kernels appropriate to autistic adolescent dataset. Therefore,

the results by using PK and NPK kernels for autistic children dataset are higher than other kernel functions. The distribution of confusion matrix cases using PK and NPK for Autistic Children dataset was as follows: the individuals who were actually infected with the disease and who were also classified as infected cases totalled (151) cases, the individuals who were actually infected with the disease but who were classified as non-infected cases totalled (0) cases, the individuals who were actually not infected with the disease but who were classified as infected cases totalled (151) cases and the individuals who were actually not infected with the disease and who were also classified as non-infected cases totalled (0) cases.

Next, in the process of the classification of the Autistic Adolescent dataset, it was found that NPK and RBF had the best performance because the distributions and equations of NPK and RBF kernels appropriate to autistic adolescent dataset. Therefore, the results by using NPK and RBF kernels for autistic adolescent dataset are higher than other kernel functions. The distribution of confusion matrix cases using NPK and RBF kernels for Autistic Adolescent dataset was as follows: the individuals who were actually infected with the disease and who were also classified as infected cases totalled (63) case, the individuals who were actually infected with the disease but who were classified as non-infected cases totalled (0) cases, the individuals who were actually not infected with the disease but who were classified as infected cases totalled (3) cases and the individuals who were actually not infected with the disease and who were also classified as non-infected cases totalled (28) cases.

In the process of the classification of the Chronic Kidney Disease dataset, it was found that NPK and RBF had the best performance because the distributions and equations of NPK and RBF kernels appropriate to Chronic Kidney Disease dataset. Therefore, the results by using NPK and RBF kernels for Chronic Kidney Disease dataset are higher than other kernel functions. The distribution of confusion matrix cases using NPK and RBF kernels for Chronic Kidney Disease dataset was as follows: the individuals who had actually been infected with the disease and who were also classified as infected cases totalled (249) cases, the individuals who were actually infected with the disease but who were classified as non-infected cases totalled (1) case, the individuals who were actually not infected with the disease but who were classified as infected cases totalled (0) cases and the individuals who were not actually infected with the disease and who were also classified as non-infected cases totalled (150) cases.

After that, in the process of the classification of the Cryotherapy treatment dataset, it was found that PUK and RBF had the best performance because the distributions and equations of PUK and RBF kernels appropriate to Cryotherapy treatment dataset. Therefore, the results by using PUK and RBF kernels for Cryotherapy treatment dataset are higher than other kernel functions. The distribution of confusion matrix cases using PUK and RBF for Cryotherapy treatment dataset was as follows: the individuals who had actually been treated using this method and were also classified as treated cases totalled (41) cases, the individuals who had actually been treated using this method but who were classified as not-treated cases totalled (1) case, the individuals who had not actually been treated using this method but who were classified as treated cases totalled (1) case and the individuals who had not actually been treated using this method and who were also classified as non-treated cases totalled (47) cases.

Finally, in the process of the classification of the Immunotherapy treatment dataset, it was found that PK and PUK had the best performance (comparatively). The reason why all the kernel functions did not perform well in classifying immunotherapy datasets Compared to other data sets is that immunotherapy dataset considered an imbalanced dataset where the number of data points is distributed unevenly on both dataset classes. The distribution of confusion matrix cases using PK and PUK for Immunotherapy treatment dataset was as follows: individuals who had actually been treated using this method and who were also classified as treated cases totalled (6) cases; the individuals who had been treated using this method but who were classified as not-treated cases totalled (13) cases and the individuals who had not been treated using this method but who were classified as treated cases totalled (4) case. The individuals who had not been treated using this method and who were also classified as non-treated cases totalled (67)

8.2. Classification Performance Measurements of the SVM Kernel Functions Classifying Medical Data

As mentioned above, good results were obtained when evaluating the performance of the used kernels according to the common classification measurements used in the field of data mining. In this section, we will discuss the results obtained and compare the performance of each kernel. We will also indicate the best and worst classification performance of the kernels according to the performance classification measurements that can be seen in Table 8.2.

Table 8.2. SVM kernel Performance Measurements classifying Medical Data

Used Datasets	Kernel Functions	Classification Performance measures			
		Accuracy (%)	Precision	Sensitivity	F-measure
Children Autistic	NPK	95.5479	0.973	0.940	0.956
	PK	100	1.000	1.000	1.000
	PUK	100	1.000	1.000	1.000
	RBF	99.3151	0.993	0.993	0.993
Adolescent Autistic	NPK	97.1154	0.955	1.000	0.977
	PK	93.2692	0.967	0.921	0.943
	PUK	96.1538	0.954	0.984	0.969
	RBF	97.1154	0.955	1.000	0.977
Chronic Kidney Disease	NPK	99.75	1.000	0.996	0.998
	PK	99.25	1.000	0.988	0.994
	PUK	99	1.000	0.984	0.992
	RBF	99.75	1.000	0.996	0.998
Cryotherapy	NPK	95.5556	0.975	0.929	0.951
	PK	96.6667	0.976	0.952	0.964
	PUK	97.7778	0.976	0.976	0.976
	RBF	97.7778	0.976	0.976	0.976
Immunotherapy	NPK	78.8889	0.000	0.000	0.000
	PK	81.1111	0.600	0.316	0.414
	PUK	81.1111	0.600	0.316	0.414
	RBF	78.8889	0.000	0.000	0.000

According to Table 8.2 and in the case of accuracy in the performance of the kernel functions in the processes of classifying the used medical datasets, PK and PUK had the best performance in the accuracy of Autistic Children dataset classification (Because of the extreme balance of this dataset, where all data points are distributed almost equally on dataset classes) by a percentage which at best reached (100%) and at worst was (81.1111%) in the Immunotherapy dataset classification in the first rank compared to the performance of the other kernels in the classification of all of the other datasets. Each of the NPK and RBF kernels came out as having the second-best performance in relation to the accuracy of the Chronic Kidney Disease dataset classification by a percentage which at best reached (99.75) and at worst was (78.8889%) in the Immunotherapy dataset classification.

After that, in the case of the precision of the performance measurements of the kernel functions in the processes of classifying the used medical datasets, each of the PK and PUK had the best performance in relation to the precision of the Autistic Children and Chronic Kidney Disease datasets by a percentage that at best reached (1.000) and at worst (0.600) in the Immunotherapy dataset classification in the first rank compared to the performance of the other kernels in the classification of all the other datasets. Each of the NPK and RBF kernels came out as having the second-best performance in the precision terms of the Chronic Kidney Disease dataset classification by a percentage which at best reached (1.000) and at worst was (0.000) in the Immunotherapy dataset classification.

In the case of the sensitivity of the performance measurements of the kernel functions in processes of classifying the used medical datasets, each of the PK and PUK had the best performance in terms of sensitivity in the Autistic Children dataset classification by a percentage which at best reached (1.000) and at worst was (0.316) in the Immunotherapy dataset classification. This was in the first rank compared to the performance of the other kernels in the classification of all of the other datasets. Each of the NPK and RBF kernel had the second-best performance in terms of the sensitivity of the Autistic Adolescent dataset classification by a percentage which at best reached (1.000) and at worst was (0.000) in the Immunotherapy dataset classification.

Finally, in the case of the F-Measures of the performance measurements of the kernel functions involved in the processes of classifying the used medical datasets, as can be seen, each of the PK and PUK had the best performance in the Autistic Children dataset classification by a percentage which at best reached (1.000) and at worst (0.414). Then, each of the NPK and RBF kernels came out as the second-best performance in terms of the F-

measure in the Chronic Kidney Disease dataset classification by a percentage which at best reached (0.998) and at worst (0.000) according to the Immunotherapy dataset classification.

There are other important criteria that contribute to the assessment of the classification performance when building a 10-fold cross-validation model by the classifier, including the error rate of the algorithm, the time that it needs to complete its classification tasks, and the number of data points that were correctly and incorrectly classified, all of which are referred to in Table 8.3.

Table 8.3. Kernel Function Error Rates for 10-fold Cross Validation

Used Datasets	Kernel Functions	Correctly predicted instances	Incorrectly predicted instances	Error rates %	Overall time taken to build the model (in seconds)
Autistic Children	NPK	279	13	4.45	0.04
	PK	292	0	0	0.03
	PUK	292	0	0	0.03
	RBF	290	2	0.68	0.02
Autistic Adolescents	NPK	101	3	2.88	0.01
	PK	97	7	6.73	0.01
	PUK	100	4	3.84	0.01
	RBF	101	3	2.88	0.01
Chronic Kidney Disease	NPK	399	1	0.25	0.02
	PK	397	3	0.75	0.01
	PUK	396	4	1	0.02
	RBF	399	1	0.25	0.02
Cryotherapy	NPK	86	4	4.4	0.01
	PK	87	3	3.33	0.01
	PUK	88	2	2.22	0.01
	RBF	88	2	2.22	0.01
Immunotherapy	NPK	71	19	21.11	0.01
	PK	73	17	18.88	0.01
	PUK	73	17	18.88	0.01
	RBF	71	19	21.11	0.01

According to Table 8.3, the time that the kernels needed to classify any type of medical data used did not exceed 0.04 seconds and was no less than 0.01 seconds. The highest error rate was recorded in the NPK and RBF kernels in the case of classifying the Immunotherapy dataset, which reached 21.11%. Both PK and PUK performed a complete percentage of accuracy (100%) in the case of the autistic child dataset classification; therefore, they did not record an error rate 0%.

The error rate is inversely proportional to the number of data points that have been correctly classified, and this is further inverse with the number of data points that have been incorrectly classified. The higher value of error rate of the kernel classification performance, the lowest number of data points that have been correctly classified. Vice versa with the number of data points that have been incorrectly classified.

Table 8.4. Comparisons Between Proposed Methods with Other Methods Using Autistic Children and Autistic Adolescent Datasets

Used methods for Autistic Children	Accuracy %	Used methods For Autistic Adolescent	Accuracy %
Proposed method with NPK	95.5479	Proposed method with NPK	97.1154
Proposed method with PK	100	Proposed method with PK	93.2692
Proposed method with PUK	100	Proposed method with PUK	96.1538
Proposed method with RBF	99.3151	Proposed method with RBF	97.1154
Ref [76]	99.7	Ref [77]	98.27

As shown in Table 8.4. the proposed method with PK and PUK kernel functions are superior than other method using Autistic Children dataset. The obtained accuracy (%) for proposed methods using PK and PUK is 100%. In case of Autistic Adolescent Dataset, the other method with accuracy 98.27 is superior than proposed methods.

Table 8.5. Comparisons Between Proposed Methods with Other Methods Using Chronic Kidney Disease Dataset

Used methods for Chronic Kidney Disease	Accuracy %
Proposed method with NPK	99.75
Proposed method with PK	99.25
Proposed method with PUK	99
Proposed method with RBF	99.75
Ref [78]	99

As shown in Table 8.5. the proposed method with NPK, PK and PUK kernel functions are superior than other method using Chronic Kidney Disease. The obtained accuracy (%) for proposed methods using NPK and RBF is 99.75 and for PK is 99.25%.

Table 8.6. Comparisons Between Proposed Methods with Other Methods Using Cryotherapy and Immunotherapy Datasets

Used methods for Cryotherapy	Accuracy %	Used methods for Immunotherapy	Accuracy %
Proposed method with NPK	95.5556	Proposed method with NPK	78.8889
Proposed method with PK	96.6667	Proposed method with PK	81.1111
Proposed method with PUK	97.7778	Proposed method with PUK	81.1111
Proposed method with RBF	97.7778	Proposed method with RBF	78.8889
Ref [79]	80.7	Ref [79]	83.33

As shown in Table 8.6. the proposed methods with all kernel functions are superior than other method using Cryotherapy dataset. In case of Immunotherapy Dataset, the other method with accuracy 83.33 is superior than proposed methods.

9. CONCLUSIONS

In this study, the emphasis was placed on the medical aspect representing three types of diseases (Autistic Children, Autistic Adolescents, and Chronic Kidney failure) and two of the common treatments used to treat warts. All of the datasets were extracted from the UCI data depository. The approach of this study was to try to build a system through which we could know whether the cases to be diagnosed are considered sick or not sick, treated or not treated depending on the used datasets. Four types of the support vector machine kernel functions were used (normalised polynomial kernel function (NPK), polynomial kernel function (PK), Pearson VII function based Universal Kernel function (PUK), and Radial Basis Function Kernel (RBF)). A comparison was made of these classifiers in terms of data classification based on the familiar performance standards (confusion matrix, accuracy, sensitivity, precision and error rate) in the field of data mining to evaluate the extent of each kernel's impact on the used medical datasets.

In order to obtain the best performance for each kernel, an increase in the value of parameters that makes the kernel results being adjusted for each dataset in WEKA workbench tool using 10 fold cross-validation to divide the data into two parts: the testing data and the training data. The parameters vary from kernel to kernel, and generally include C, Omega, Sigma, and Gamma. It was concluded that the gradual and systematic increase in the value of parameters, especially the parameter (C), is an important step in the best practice in the use of SVM kernel functions as well as to tell the kernel optimization how much you want to avoid misclassifying data points According to this study, the best result obtained in the accuracy of the data classification was for both kernels (PK) and (PUK) in classifying the autistic children dataset, which reached 100% in the accuracy of performance. As a future work, it will be about how to handle the classification of imbalanced datasets as (Immunotherapy) by applying other types of SVM kernel functions or by using other classification algorithms.

REFERENCES

- [1] Boyd, C.M., Landefeld, C.S., Counsell, S.R., Palmer, R.M., Fortinsky, R.H., Kresevic, **D., Burant, C. and Covinsky, K.E.**, 2008. Recovery of activities of daily living in older adults after hospitalization for acute medical illness. *Journal of the American Geriatrics Society*, 56(12), pp.2171-2179.
- [2] **Palaniappan, S. and Awang, R.**, 2008, March. Intelligent heart disease prediction system using data mining techniques. In *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on* (pp. 108-115). IEEE.
- [3] **De'ath, G. and Fabricius, K.E.**, 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11), pp.3178-3192.
- [4] **Lavanya, D. and Rani, K.U.**, 2011. Performance evaluation of decision tree classifiers on medical datasets. *IJCA) International Journal of Computer Applications*, 26(4).
- [5] **Mazurowski, M.A., Habas, P.A., Zurada, J.M., Lo, J.Y., Baker, J.A. and Tourassi, G.D.**, 2008. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks*, 21(2-3), pp.427-436.
- [6] **Brameier, M. and Banzhaf, W.**, 2001. A comparison of linear genetic programming and neural networks in medical data mining. *IEEE Transactions on Evolutionary Computation*, 5(1), pp.17-26.
- [7] **Soni, J., Ansari, U., Sharma, D. and Soni, S.**, 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8), pp.43-48.
- [8] **Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M. and Haussler, D.**, 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10), pp.906-914.
- [9] **Cristianini, N. and Shawe-Taylor, J.**, 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.
- [10] **Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P.**, 1996. From data mining to knowledge discovery in databases. *AI magazine*, 17(3), p.37.

- [11] **Bock, H.H. and Diday, E. eds.**, 2012. *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Springer Science & Business Media.
- [12] **Linoff, G.S. and Berry, M.J.**, 2011. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.
- [13] **Liu, H. and Kešelj, V.**, 2007. Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. *Data & Knowledge Engineering*, 61(2), pp.304-330.
- [14] **Vercellis, C.**, 2011. *Business intelligence: data mining and optimization for decision making*. John Wiley & Sons.
- [15] **Peng, Y., Kou, G., Shi, Y. and Chen, Z.**, 2008. A descriptive framework for the field of data mining and knowledge discovery. *International Journal of Information Technology & Decision Making*, 7(04), pp.639-682.
- [16] **Pujari, A.K.**, 2001. *Data mining techniques*. Universities press.
- [17] **Ester, M., Kriegel, H.P., Sander, J. and Xu, X.**, 1996, August. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).
- [18] **Zhu, X. ed.**, 2007. *Knowledge Discovery and Data Mining: Challenges and Realities: Challenges and Realities*. Igi Global.
- [19] **Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P.**, 1996. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11), pp.27-34.
- [20] **Levy, A.Y.**, 1999. Combining artificial intelligence and databases for data integration. In *Artificial Intelligence Today*(pp. 249-268). Springer, Berlin, Heidelberg.
- [21] **Hadavandi, E., Shavandi, H. and Ghanbari, A.**, 2010. Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowledge-Based Systems*, 23(8), pp.800-808.
- [22] **Liu, H. and Motoda, H.**, 2012. *Feature selection for knowledge discovery and data mining* (Vol. 454). Springer Science & Business Media.
- [23] **Bhatt, G.D.**, 2001. Knowledge management in organizations: examining the interaction between technologies, techniques, and people. *Journal of knowledge management*, 5(1), pp.68-75.

- [24] <https://www.slideshare.net/amritanshumehra/knowledge-discovery-and-data-mining> data mining and knowledge discovery in database (Accessed on 8 Dec. 2018.)
- [25] **Delen, D., Walker, G. and Kadam, A.**, 2005. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2), pp.113-127.
- [26] **Ge, Z., Gao, F. and Song, Z.**, 2011. Batch process monitoring based on support vector data description method. *Journal of Process Control*, 21(6), pp.949-959.
- [27] **Lei-da Chen, T.S. and Frolick, M.N.**, 2000. Data mining methods, applications, and tools. *Information systems management*, 17(1), pp.67-68.
- [28] **Schölkopf, B., Smola, A.J. and Bach, F.**, 2002. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.
- [29] **Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M. and Haussler, D.**, 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1), pp.262-267
- [30] **Hastie, T., Rosset, S., Tibshirani, R. and Zhu, J.**, 2004. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5(Oct), pp.1391-1415.
- [31] **Wang, L. ed.**, 2005. *Support vector machines: theory and applications* (Vol. 177). Springer Science & Business Media.
- [32] **Hong, X., Chen, S. and Harris, C.J.**, 2008. A forward-constrained regression algorithm for sparse kernel density estimation. *IEEE Transactions on Neural Networks*, 19(1), pp.193-198.
- [33] <http://www.simafore.com/blog/bid/112816/When-do-support-vector-machines-trump-other-classification-methods> When do support vector machines trump other classification methods (Accessed on 8 Dec. 2018.)
- [34] **Chen, D.R., Wu, Q., Ying, Y. and Zhou, D.X.**, 2004. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, 5(Sep), pp.1143-1175.
- [35] **Yang, Y. and Liu, X.**, 1999, August. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 42-49). ACM.

- [36] **Pham, B.T., Bui, D.T., Pourghasemi, H.R., Indra, P. and Dholakia, M.B.**, 2017. Landslide susceptibility assessment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. *Theoretical and Applied Climatology*, 128(1-2), pp.255-273.
- [37] **Subasi, A.**, 2013. Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders. *Computers in biology and medicine*, 43(5), pp.576-586.
- [38] **Kohavi, R.**, 1995, August. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*(Vol. 14, No. 2, pp. 1137-1145).
- [39] **Arlot, S. and Celisse, A.**, 2010. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4, pp.40-79.
- [40] <https://www.ritchieng.com/machine-learning-cross-validation/> Cross-validation (Accessed on 8 Dec. 2018)
- [41] **Hall, M.A. and Holmes, G.**, 2003. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data engineering*, 15(6), pp.1437-1447.
- [42] **Muller, K.R., Mika, S., Ratsch, G., Tsuda, K. and Scholkopf, B.**, 2001. An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks*, 12(2), pp.181-201.
- [43] **Amari, S.I. and Wu, S.**, 1999. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6), pp.783-789.
- [44] **Byun, H. and Lee, S.W.**, 2002. Applications of support vector machines for pattern recognition: A survey. In *Pattern recognition with support vector machines* (pp. 213-236). Springer, Berlin, Heidelberg.
- [45] **Eichhorn, J. and Chapelle, O.**, 2004. Object categorization with SVM: kernels for local features. *Advances in Neural Information Processing Systems (NIPS)*, 89
- [46] <https://www.quora.com/What-is-the-purpose-of-the-support-vector-in-SVM> What is the purpose of the support vector in SVM (Accessed on 8 Dec. 2018.)
- [47] **Trivedi, S.K. and Dey, S.**, 2013. Effect of various kernels and feature selection methods on SVM performance for detecting email spams. *International Journal of Computer Applications*, 66(21).

- [48] **Prather, J.C., Lobach, D.F., Goodwin, L.K., Hales, J.W., Hage, M.L. and Hammond, W.E.**, 1997. Medical data mining: knowledge discovery in a clinical data warehouse. In *Proceedings of the AMIA annual fall symposium* (p. 101). American Medical Informatics Association.
- [49] **Xing, Y., Wang, J. and Zhao, Z.**, 2007, November. Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In *Convergence Information Technology, 2007. International Conference on*(pp. 868-872). IEEE.
- [50] <https://archive.ics.uci.edu/ml/index.php> (Accessed on 29 Nov. 2018.)
- [51] **Lovaas, O.I.**, 1987. Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of consulting and clinical psychology*, 55(1), p.3.
- [52] <http://www.mediamaxnetwork.co.ke/430436/double-jeopardy-for-parents-with-autistic-children/> Double jeopardy for parents with autistic children (Accessed on 8 Dec. 2018.)
- [53] **Happé, F.G.**, 1994. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, 24(2), pp.129-154.
- [54] <https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++> (Accessed on 29 Nov. 2018.)
- [55] **Hollander, E., Phillips, A., Chaplin, W., Zagursky, K., Novotny, S., Wasserman, S. and Iyengar, R.**, 2005. A placebo controlled crossover trial of liquid fluoxetine on repetitive behaviors in childhood and adolescent autism. *Neuropsychopharmacology*, 30(3), p.582.
- [56] <http://hesed.info/blog/avoiding-eye-contact-autism.abp> Avoiding Eye Contact Autism (Accessed on 8 Dec. 2018.)
- [57] <https://smjournals.com/ebooks/chronic-kidney-disease/index.php> Chronic Kidney Disease (Accessed on 8 Dec. 2018.)
- [58] <https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Adolescent+++> (Accessed on 29 Nov. 2018.)

- [59] **Go, A.S., Chertow, G.M., Fan, D., McCulloch, C.E. and Hsu, C.Y.**, 2004. Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. *New England Journal of Medicine*, 351(13), pp.1296-1305.
- [60] https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease (Accessed on 29 Nov. 2018.)
- [61] <https://airfreshener.club/quotes/different-kinds-of-warts.html> Different Kinds Of Warts (Accessed on 8 Dec. 2018.)
- [62] **Wiley, D.J., Douglas, J., Beutner, K., Cox, T., Fife, K., Moscicki, A.B. and Fukumoto, L.**, 2002. External genital warts: diagnosis, treatment, and prevention. *Clinical Infectious Diseases*, 35(Supplement_2), pp.S210-S224.
- [63] **Bruggink, S.C., Gussekloo, J., Berger, M.Y., Zaaijer, K., Assendelft, W.J., de Waal, M.W., Bavinck, J.N.B., Koes, B.W. and Eekhof, J.A.**, 2010. Cryotherapy with liquid nitrogen versus topical salicylic acid application for cutaneous warts in primary care: randomized controlled trial. *Canadian Medical Association Journal*, 182(15), pp.1624-1630.
- [64] <http://removewartsfast.com/cryotherapy/> Everything About Cryotherapy (Accessed on 8 Dec. 2018.)
- [65] <https://archive.ics.uci.edu/ml/datasets/Cryotherapy+Dataset+> (Accessed on 29 Nov. 2018.)
- [66] **Johnson, S.M., Roberson, P.K. and Horn, T.D.**, 2001. Intralesional injection of mumps or Candida skin test antigens: a novel immunotherapy for warts. *Archives of dermatology*, 137(4), pp.451-455.
- [67] http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0365-05962012000400011 Treatment of common warts with the immune stimulant Propionium bacterium parvum (Accessed on 8 Dec. 2018.)
- [68] <https://archive.ics.uci.edu/ml/datasets/Immunotherapy+Dataset> (Accessed on 29 Nov. 2018.)
- [69] **Cherkassky, V. and Ma, Y.**, 2004. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural networks*, 17(1), pp.113-126
- [70] **Bhardwaj, B.K. and Pal, S.**, 2012. Data Mining: A prediction for performance improvement using classification. *arXiv preprint arXiv:1201.3418*.

- [71] **Dangare, C.S. and Apte, S.S.**, 2012. Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, 47(10), pp.44-48.
- [72] **Hammond, T.O. and Verbyla, D.L.**, 1996. Optimistic bias in classification accuracy assessment. *International Journal of Remote Sensing*, 17(6), pp.1261-1266.
- [73] **Loh, W.Y.**, 2009. Improving the precision of classification trees. *The Annals of Applied Statistics*, pp.1710-1737.
- [74] **Fule, P. and Roddick, J.F.**, 2004, January. Detecting privacy and ethical sensitivity in data mining results. In *Proceedings of the 27th Australasian conference on Computer science-Volume 26* (pp. 159-166). Australian Computer Society, Inc..2004.
- [75] **Powers, D.M.**, 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.
- [76] **Wall, D.P., Kosmicki, J., Deluca, T.F., Harstad, E. and Fusaro, V.A.**, 2012. Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational psychiatry*, 2(4), p.e100.
- [77] **Schopler, E., Reichler, R.J., DeVellis, R.F. and Daly, K.**, 1980. Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS). *Journal of autism and developmental disorders*, 10(1), pp.91-103.
- [78] **Avci, E., Karakus, S., Ozmen, O. and Avci, D.**, 2018, March. Performance comparison of some classifiers on Chronic Kidney Disease data. In *Digital Forensic and Security (ISDFS), 2018 6th International Symposium on* (pp. 1-4). IEEE.
- [79] **Khozeimeh, F., Alizadehsani, R., Roshanzamir, M., Khosravi, A., Layegh, P. and Nahavandi, S.**, 2017. An expert system for selecting wart treatment method. *Computers in biology and medicine*, 81, pp.167-175.

CURRICULUM VITAE

Hardi Mohammed TALABANI

E-mail: harditalabani@gmail.com
Nationality: Iraq
Place of birth: Kirkuk
Date of birth: 18 / 03 / 1986
Marital status: single

EDUCATION

- 2017 – 2019 Master Degree, Software Engineering department, Technology Faculty, Fırat University, Elazığ, Turkey.
- 2006 – 2010 Bachelor Degree, Computer Science department, College of Science, Newroz University, Duhok, Iraq.

WORK EXPERIENCE

- I.T technician at Ministry of Finance, Charmo Bank, Iraq, Sulaymaniyah, 2010-2014.
- I.T technician at Ministry of Finance, Information Technology (I.T) Bank, Iraq, Sulaymaniyah 2014 - 2016.
- Research Assistance at Ministry of High Education Charmo University, Iraq, Sulaymaniyah 2016 – present.

PUBLICATIONS

- TALABANI, H., AVCI, E., (2018). *Performance Comparison of SVM Kernel Types on Child Autism Disease Database*, 3rd international conference on artificial intelligence and data processing (IDAP), 34 (5), 177 – 181.
- TALABANI, H., AVCI, E., (2018). *Impact of Various Kernels on Support Vector Machine Classification Performance for Treating Wart Disease*, 3rd international conference on artificial intelligence and data processing (IDAP), 107 (6), 566 - 571