# The role of data mining in diagnosing the diseases: A case study of detecting the thyroid disease

Nariman Khaled Hambeishi [a,1,*], Shaymaa Abed Hussein [b,2], Waleed Khalid AlAzzawi [b,3]

[a] Department of Information Sciences - Faculty of Arts and Humanities, King Abdul Aziz University, Jeddah, Saudi Arabia
[b] Department of Information Sciences - Faculty of Arts and Humanities, King Abdul Aziz University, Jeddah, Saudi Arabia
[c] Department of Medical Instruments, Medical Technical College, Al-Farahidi University, Baghdad, 10022, Iraq
[1] nhambishi@hotmail.com*; [2] shaimaa2021@uomanara.edu.iq; [3] waleed.khalid@alfarahidiuc.edu.iq
* corresponding author

## ARTICLE INFO

**Article history**
Received: 2021-12-05
Revised: 2022-02-20
Accepted: 2022-06-29
Published: 2022-12-20

**Keywords**
Data Mining
Thyroid disease
EMR
Health Care
Machine Learning
Decision Tree
Naïve Bayes
DNA
Support Vector Machines

## ABSTRACT

One of the most important tools used in the field of medicine is to search for data, especially in the field of exploration and knowledge of the prevailing health conditions, research into diseases and patient behavior in society, as well as contribute to knowing the effect of drugs and drugs on patients according to previous patient records. Data mining also helps to identify diseases spread in a specific area, which helps to take the necessary measures and awareness to control this disease, and works to develop and update the field of medicine and medicines and increase their spread. Efficiency and processing capacity. Data mining techniques are a recent method in the medical field, and have become increasingly reliable in diagnosing diseases especially in diagnosing and detecting thyroid diseases. Since such techniques help to properly analyze and accurately predict the disease, they contribute to helping physicians, providing appropriate medical care, and reducing the incidence or development of side effects of the disease. Today, data mining techniques are important in the medical field as they help detect diseases and epidemics that are spreading around the world. The purpose of this paper is to know the role of data mining techniques in the diagnosis of thyroid disease through the survey of a number of relevant studies. The critical appraisal tool of previous studies in this area has been used. The study has also found the importance of early diagnosis of thyroid diseases, providing proper treatment for patients. It also agreed on the importance of data mining tools such as neural networks, Machine and decision tree learning and their disease diagnosis potential. The paper recommends the need for extensive systematic reviews of studies on the use of data mining techniques in the diagnosis of thyroid diseases. Also, we concluded that Prospecting tools in medical devices have had an enormous and important impact on the health care industry and the country. It should be remembered that the medical data that began to multiply in a large amount must be contained herein a lot of useful information that greatly affects the improvement of the level of medical services and detection in Characteristics of many diseases and epidemics and find solutions to many difficult diseases.

## 1. Introduction

Presently, clinical foundations for the most part use EMR to record a patient's condition, including analytic data, strategies performed, and treatment results. EMR has been broadly perceived as a significant asset for investigation. Be that as it may, EMR has the attributes of variety, impediments, excess, and particularity, which make it hard to straightforwardly perform information extraction and investigation. In this way, it is important to pre-measure the source information to further develop information quality and further develop information extraction results. Various sorts of information require distinctive handling strategies. Most organized information commonly needs exemplary pre-preparing methods, including information purifying, information reconciliation, information change, and information decrease. For semi-organized or unstructured information, like a clinical text, that contains more wellbeing data, it requires more mind boggling and testing preparing strategies. The data extraction task for restorative texts for the most part incorporates NER (Identification of Named Entities) and RE (Relationship Extraction). This paper centers around the EMR handling measure and for the most part investigates the fundamental innovations. Furthermore, we are leading an inside and out study on applications created dependent on text mining alongside open difficulties and exploration issues for future work.

Thyroid gland is currently one of the most common diseases, especially among women. Iodine deficiency is a cause of this disease. It is widely spread the United States and the world, and statistical data indicate that women are five to eight times more likely to have thyroid problems than men [1]. The disease is caused by a breakdown in the hormones secreted by the thyroid gland, resulting in hypothyroidism where glands cannot produce enough thyroid hormones, or hyperthyroidism, where there is an increase in thyroid hormone secretion. Symptoms of impairment include constipation, weight gain, slow heart rate, fatigue, and depression, while symptoms of hyperthyroidism are weight loss, arrhythmias, nervousness and difficulty of sleeping.

Diagnosing the disease is a difficult and highly complex task, but timely identification contributes to the proper treatment. Studies indicate that early detection of the disease and rapid medical care can provide effective treatment for the majority of the injured [2]. It is therefore important to take advantage of modern technological capabilities and developments in support of early diagnosis and the provision of appropriate health care. Data mining applications have enormous potentials in medical services. However, the efficiency of healthcare information extraction techniques depends on the availability of duplicate health care data. Machine learning (ML) and data extraction (DM) have upgraded symptomatic capacities in the clinical field, and help in recognizing and identifying endemic infections and tumors, as the conclusion under typical information might contain abandons or may not be precise because of the presence of huge clinical data in any event, for a clinical subject matter expert, so the ML and DM can assist with extricating information based reactions from loose information, which helps in the early location of straightforward illnesses and works on tolerant consideration.

It is evident that 60% of patients are not diagnosed correctly due to defects in the diagnostic mechanism, or it is not accurate due to the presence of a large amount of medical information. Therefore, the inaccuracy of diagnosis causes an increase in the incidence of thyroid cancer and a high mortality rate due to delaying appropriate treatment. Therefore, the use of data mining techniques helps reduce the rate of error in diagnosis, as statistics indicate that it may decrease by 12% as a result of relying on machine learning tools and techniques, automating decision-making functions and successfully solving complex problems. The reliance on the use of these techniques greatly contributes to the diagnosis of the disease [1].

This paper therefore seeks to review the literature and relevant studies to investigate the role of data mining in the diagnosis of thyroid diseases. It aims to answer the following question: What are the most important data mining techniques used in diagnosing thyroid diseases? What is the role of data mining techniques in managing thyroid diseases? The importance of the paper is that it seeks to define the relationship between the data mining techniques in extracting information related to thyroid disease and the most important of these techniques, in addition to highlighting the most important studies on proposed models that contribute to the detection and diagnosis of the disease. Despite the importance of data mining, the advantages of group acquisition should be chosen in line with the proposed algorithms, when used in the diagnosis of diseases, to obtain more accuracy in diagnosis, as the data may contain ineffective or redundant features that affect the vitality of accurate diagnosis [1]. The studies in this paper were selected on the basis of their direct link to the topic, while studies on the subject of thyroid disease were excluded in general. This paper contributes to the importance of early diagnosis of thyroid disease in providing medical care and reducing complications of the disease .The problem of the study is the following question: What is the role of data techniques applications in diagnosing thyroid gland diseases? What is the most important technique? This paper is divided into five sections: First: Introduction, second: Literature review, third: Method of research used, fourth: Results and discussion, fifth: Conclusion and recommendations.

## 2. Method

### 2.1. Genetic Engineering In The Medical Field

Utilizations of data mining to bioinformatics incorporate quality discovering, protein work area discovery, work theme recognition, protein work derivation, infection finding, illness anticipation, sickness therapy advancement, protein and quality association network remaking, information purifying, and protein sub-cell area forecast [3]. Genetic engineering and bioinformatics, which relies mainly on data mining techniques, so that some exploration methods have been specially developed for the purposes of bioinformatics research in a way that has made exploration and bioinformatics science two paths of integrated science, mutually developing each other. The bioinformatics environment is also considered fertile for prospecting techniques due to the vast amount of almost infinite data it contains, represented by chains of amino acids and gene chains that number in the billions. Who contributed to understanding the nature of diseases, discovering new and effective drugs for a large number of them, developing existing drugs and raising their efficiency and effectiveness. In general, data mining techniques contribute to raising the efficiency of research and vital studies in all its forms and enhance the ability

of researchers to advanced and in-depth analysis of vital information in new and unprecedented ways. Discovering many genetic diseases, their causes, and methods of treating them, such as: sickle cell disease, production of many medical hormones such as growth hormone and insulin, in addition to making and modifying vaccinations in order to reduce their side effects on the human body [4].

This paper relied on the critical review approach to search for data mining and its role in detecting thyroid diseases. This approach is used to review, analyze and criticize relevant research, analyze studies that can be consulted and present findings and recommendations that have been reached. (IEEE, Google Scholar, Pub Med) have been consulted for research on relevant studies and studies have been reviewed from 2017-2020, to determine the relevance of studies to this paper and come up with findings and recommendations. The Descriptive phenomenology utilized in this article is the fitting plan of the review in case there is very little or nonexistent accessible writing on the wonder of interest. This remains constant in Medicine where imaginative educational plan models have just barely begun to create over the most recent ten years and is clearly the justification behind 12 examinations appropriating exploratory or distinct phenomenology as an exploration plan in their review. (Polit and Beck 2006) report that graphic phenomenological concentrates regularly incorporate four stages: organizing (whereby the specialist keeps individual convictions and assessments from tainting information), intuiting (in which the analyst opens up to new implications), dissecting (through which new topics are recognized) and portraying (when the creator comprehends and endeavors to characterize the wonder). In the entirety of the 12 examinations inspected, just (Nickerson and Resick 2010) notice organizing being thought about in arranging the review while (Child and Langford 2011) and (Ritchie et al. 2010) report on an autonomous questioner and exploration aides directing the separate meetings in their investigations. Instinct isn't unequivocally expressed in any of the articles yet it tends to be inferred from the analysts' earnest point in consolidating new models and utilizing a subjective way to deal with deciding its prosperity.

### 2.2. Applications of Genetic Engineering

#### Medical applications

The applications of genetic engineering have entered the medical fields, as its applications have been used in several sections that include understanding the causes of diseases. Accordingly, drugs, new treatments, research and diagnostic methods have been developed, in addition to the development of clinical devices, and genetic engineering scientists also search for the locations of genes on the chromosome. Creates significant opportunities in the ongoing understanding of genetically related diseases and genetics among family members, as well as individual therapies.

### Industrial applications

The applications of genetic engineering have also entered the industrial field, as many chemical commodities have been produced that were mainly dependent on living organisms to be produced, such as enzymes, for example, and specialized chemicals using biotechnology applications, and this process is important as it works to raise the efficiency of industrial processes, and reduce Its negative environmental impacts, and it is reported that biotechnology worked on the use of corn as an alternative to oil, and the fermentation of sugar to produce acids that can be used later in other industrial processes, in addition to using them in the textile industry to produce cotton in a bio-based way that has higher specifications than industrial cotton.

### Animal applications

Applications of genetic engineering have been used in the field of genetic modification of animals, and the goal of this use is to produce genetically modified animals that meet human needs in various products and shapes, as these applications allow farmers to produce the desired breeds of livestock with the least possible time and the lowest cost, which allows for obtaining It is based on more food products aimed at improving the general health of the human being, and the principle of these applications is based on inserting the desired genes into the genome of cattle, subsequent in a greater amount of nutrients, in addition to increasing the nutritional value of them, and the possibility of increasing the proportion of a specific component of the nutritional value of these Substances to meet market demand, such as increasing the proportion of omega-3 acids in fish, and reducing the risk of cardiovascular disease for people who eat these fish.

## 2.3. The Importance of Genetic Engineering

Genetic engineering is concerned with various living organisms, including plants, microorganisms, animals and humans, of course, and work is carried out on studying DNA, separating chromosomes separately and inserting it into the bodies of living organisms to find out the characteristic for which this gene is responsible, and then control and control it [5].

The technique of genetic modification and engineering was used many years ago, as bacteria were the first organisms to work on in 1973, and after that work on mice was done in 1974, and finally humans began to reap the fruits of their labor when insulin was produced and sold in 1982 AD and developed later in 1994 AD.

Genetic engineering is called genetic modification, and it is an indirect manipulation by humans in order to change the physical characteristics that they will appear on. Genes enable us to increase the amount of material produced in the human body, and supply it with the deficient in order to avoid disease.

The last and important development in the world of genetic engineering has reached to isolate the genetic material and copy the material that relates to the trait that we want to modify, and then generate a structure that contains the genetic elements in order to obtain a correct constructed genetic change, and then implant it in the host's body (the human).

Genetic engineering is involved in many fields, such as medicine, chemistry, biology, biochemistry, and physical sciences of the body. It also represented an unparalleled development in pharmacology, as important drugs such as insulin and insecticides were manufactured that protect plants and crops from insect attack.

The goal of man in his development of genetic engineering and genetics has become to make life easier for human beings, reduce costs, and increase the quality and efficiency of the manufactured material, as in the medicinal protein that was manufactured and sold in 2009, which is genetically modified.

Experiments came out of the theoretical scope to the practical arena in 1986, when the United States of America and France engineered tobacco and made it resistant to herbicides, as well as the People's China marketed genetically modified plants to become resistant to viruses, and also the European Union genetically modified tomatoes and sold them, and it met with great success.

## 3. Recent Related Work

Mankind Is Passing Through The Stage Of The Supremacy Of Science And Technology, And It Is Not Clear To Anyone The Impact That This Superiority Has Had On Human Life, And There Is No Doubt That Societies That Have Acquired Information And Knowledge Are Now Ruling The World And Leading The Scene Of Advanced And Modern Countries. One Of The Forms Of Advancement And Technology And One Of Its Most Prominent Effects Is The Huge Information Inflation, Which Is Happening Steadily, As Data Stores And Databases On Which Most Systems And Applications Depend Have Become Filled With A Large And Huge Amount Of Random And Historical Data, And We Can Really Say That The Data Does Not Sleep And Does Not Stop. The Growth. According To The Latest Statistics, The Number Of Internet Users Increased In 2017 By 7.5% Compared To 2016, To Reach 3.7 Billion Users. The Social Site Twitter, For Example, Includes A Large Number Of Active Users Around The Clock, And Every Minute It Generates About 456,000 Tweets That Contain Texts, Pictures, Videos, Links And Hashtags, And You Can Imagine How Much Data This Results In In Just One Day.

With This Growth In Data, We Are Facing A Great Challenge, Which Is How To Make Use Of This Data Stored In Its Repositories, And How To Perform The Process Of Extracting Useful Information From It, And This Challenge Has Become Unsolvable By Traditional Methods Of Data Analysis, So We Need Unconventional Techniques That Accommodate Handling. With The Huge Amounts Of Data In Addition To The Ability To Deal With Different Types And Forms Of Data, This Means Technologies Of A High Level Of Efficiency And Intelligence. Through Research Into The Intellectual Production Of Data Mining And Diagnosis Of Thyroid Diseases In Available Databases, It Has Been Found That Studies Have Discussed This Topic Using Different Methodologies, Models And Methods. The Aim Of Such Studies Is To Find Out How Effective These Tools And Methods Are In Diagnosing Thyroid Diseases And The Presentation Of Studies Has Been Arranged From The Latest To The Oldest And The Following Is A Presentation Of These Studies:

**In The Work entitled: “Prediction Of Thyroid Disease Using Decision Tree Ensemble Method"** [6], This Paper Analyzes The Large And Complex Thyroid Data Set By Means Of The Meta-Data Classification Algorithm: Boosting, Bagging, Stacking, And Voting Using A New Group Model That Is Contrary To The Comparison Of Classification Accuracy And Quality Sensitivity.

The Study Adopted The Analytical Descriptive Method For Describing Data And Algorithms. The Study Was Done In 12204 Cases, Collected From The Chandan Diagnosis Siba Jaunpur Centre, Where It Contains 30 Patient Data Sets From Several Hospitals. The Sample Was Divided Into Two Groups, Including The First Group, With 6,480 Cases Belonging To The Negative Category And 5,724 Cases In The Positive Category. The Thyroid Dataset Is Based On Three Categories Of Hypothyroidism, Hyperthyroidism, And Thyroid Condition. The Study Is Based Solely On Analysis Of The Thyroid Deficiency Data Set, Associated Symptoms Such As Cold Feeling, And Increased Heartbeat. In The Proposal Form, The Assembly Method, Which Includes Bagging, Ada Boostm1 And Stacking, Is Used. After Using The Three Classification Methods, Voting Is Used To Measure Thyroid Symptoms In Women. This Model Provides Support In The Pre-Treatment That Helps The Physician “See Fig.1”. Whatever Methodology Is Used, A Key Challenge In Making Any Kind Of Prediction Is To Establish The Extent To Which The Past Is Likely To Be An Accurate Guide To The Future. There Are Different Levels Of Predictability That May Be Taken To Apply, Expressed Here In Terms Of The Structure And Parameter Estimates Of A Mathematical Model Fitted To A Previous FMD Epidemic But Now Having To Be Adapted To Make Predictions About A New Epidemic.

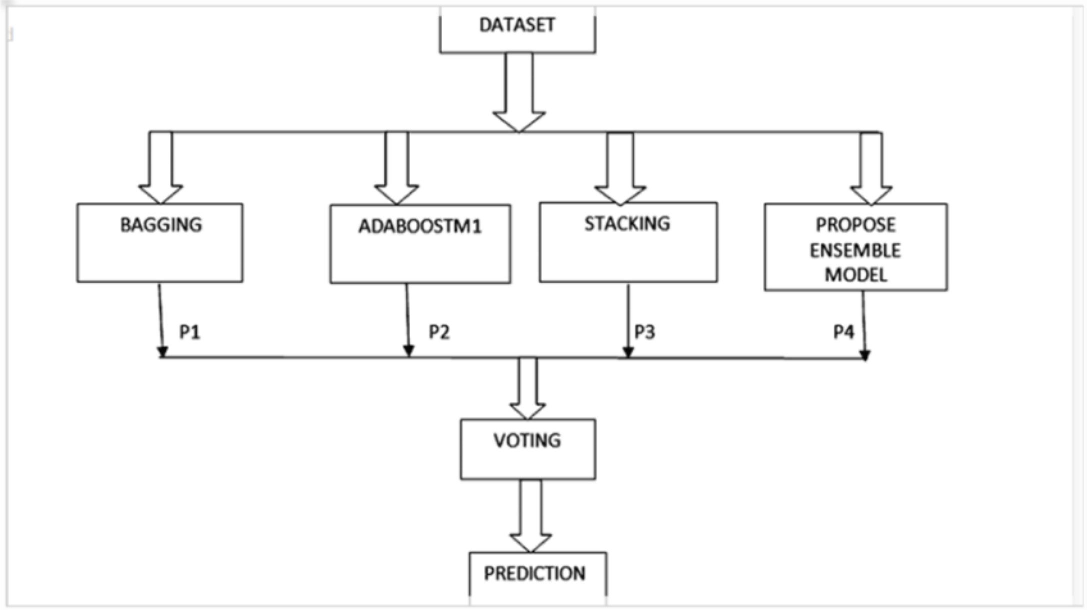

**Fig. 1.** Proposed thyroid forecasting model [7]

**Regarding study entitled "A Survey of Thyroid Detection using Dating Techniques"** [8]**,** this paper reviews some classification techniques that help detect thyroid problems, such as Naïve Bayes, Decision Tree, Artificial neurons, and Support Vector Machines. Different types of classification algorithms are used to predict thyroid problems and each algorithm has produced different levels of accuracy. This study concluded, after discussing several classification algorithms, that (J48) is used to develop the decision tree and that it is a continuation of the ID3 algorithm, providing the most accurate and the least time-consuming algorithm. This study recommends that (J48) can be combined with any other advanced methodology to give maximum accuracy and less time-consuming. It also found the importance of early detection of the thyroid problem, which will result in early recovery.

We assure that the diagnosis of genetic diseases goes through many stages, starting from the evaluation of the patient by doctors who are not specialized in genetic medicine, and then consulting the geneticist to perform a complete examination and take information about the patient's condition and the family's medical history, and then examinations of the chromosomes to ensure the safety of their number and composition, then moving to tests More precisely by analyzing the DNA of the responsible gene that causes the disease, and at other times the treating physician performs an analysis of the deficient enzymes by taking samples from the blood, skin or muscle of the affected person according to the type of disease and perhaps the problem that the doctors specializing in this field is the acute shortage of laboratories specialized in this field And a lot of samples are sent abroad to advanced laboratories around the world.

**With regards to the work entitled "Identification of Key MiRNAs in Papillary Thyroid Carcinoma Based on Data Mining and Bioinformatics Methods"**[9], this study verifies the presence of major RNA molecules in papilloma thyroid cancer (PTC) and has collected data and obtained the two independent data sets (GSE73182) and GSE113629 from the GEO database. The data, predicting miRNAs' main objectives by MIRWALK and its functional impact were analyzed by cluster Profiler R. Through analysis, miRNAs targeted genes were predicted, and thyroid cancer causes were identified, and the study revealed that based on data exploration and bioinformatics methods, a range of diagnostic indicators were reached and therapeutic targets for thyroid cancer were identified.

It is inferred that Bioinformatics is an exceptionally famous field for some; Because researchers can utilize their natural and computational abilities to foster standard models for bioinformatics research. Numerous researchers likewise observe bioinformatics to be an interesting new field of logical request with incredible potential to carry advantages to human wellbeing and society. The eventual fate of bioinformatics is reconciliation, as it will permit us to incorporate a wide scope of information sources, for example, clinical and genomic information, utilization of infection manifestations to foresee hereditary transformations as well as the other way around, just as coordinating geographic data framework (GIS) like guides and climate frameworks with crop wellbeing information and example information. Hereditary examinations will permit anticipating victories of farming analyses. One more future space of hereditary informatics is enormous scope similar genomic considers.

**The article entitled: "Genes Influenced by Obesity May Contribute to the Development of Thyroid Cancer through the Regulation of Insulin Levels"**[10], this article aims to identify some of the causes of the development of thyroid cancer, including obesity. To reach at the outcomes, writing-based TC quality information and corpulence hereditary relationship information was broadly extricated, through which stoutness qualities and TC qualities were distinguished and thought about. From that point forward, an enormous examination was done to test for qualities directed by heftiness yet not identified with TC. And afterward, writing based pathway investigation and gene group enrichment analysis (GSEA)) was performed to recognize the potential useful organization connecting the chose particles, TC, and their natural profile, and to find the component behind the TC-upgrading taste impact at the hereditary level. To achieve the purpose of the study, a knowledge-based algorithm was used to analyze data on the relationship between disease and genes, and 176 fat-regulating genes were detected and identified as TC-related, and used to build a common genetic background for the causes of obesity and TC. After analysis and exploration through analysis of literature-based pathways, the study found that the five obesity genes associated with TC were identified and that biological experiments are necessary to choose these relationships, and the discussion absorbed on the five genes that passed the criteria of importance in the analysis.

**The work entitled "Prediction of Thyroid Disease using Data Mining Techniques"** [11], aim to identify the thyroid diseases using various techniques at an early stage, and to find TSH, T3 and T4 connections to hyperthyroidism and hyperthyroidism, and to find a relationship between TSH, T3 and T4 with sex towards hyperthyroidism and hypothyroidism, by using data mining techniques, the decision tree and the Neural Network algorithm. The study relied on thyroid data from the UCI database site, where the database was formed from the records of thyroid patients and the records contained multiple features in the description of the data set and the data are then extracted through data mining techniques to predict thyroid disease. Figure (2) illustrates the proposed classification model for thyroid disease prediction. The study found the importance of using the data mining classification algorithms to eliminate thyroid problems and to give a different level of accuracy to each technique. It also helps to reduce the alarming data of each technique. The study confirmed the results of different algorithms on the speed, accuracy, model performance and cost of treatment. In addition, the classification of effective data also helps to find better-cost treatment for thyroid patients and facilitates management.

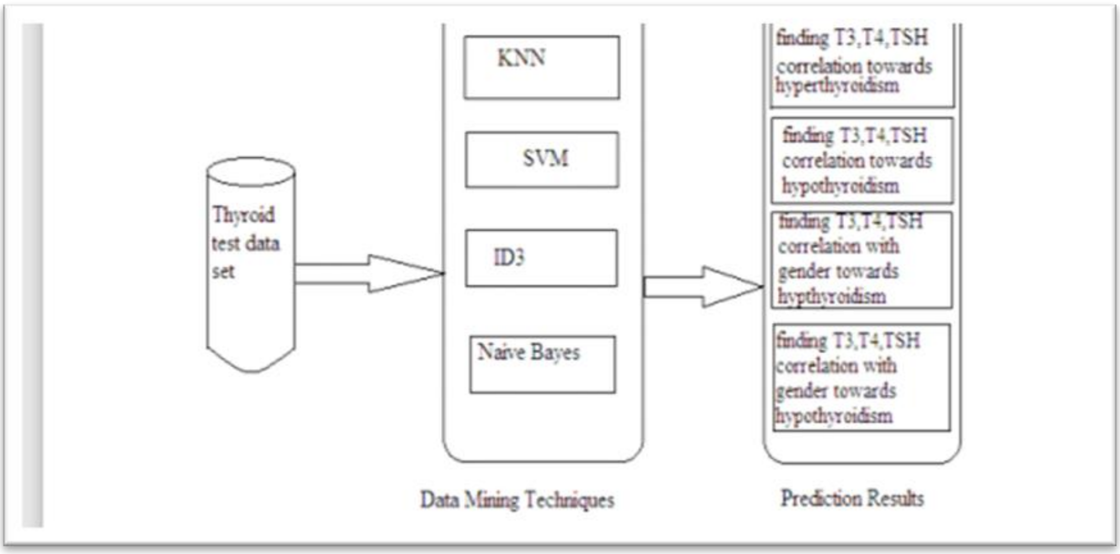
Thyroid test data set
KNN
SVM
ID3
Naive Bayes
finding T3,T4,TSH correlation towards hyperthyroidism
finding T3,T4,TSH correlation towards hypothyroidism
finding T3,T4,TSH correlation with gender towards hypthyroidism
finding T3,T4,TSH correlation with gender towards hypothyroidism
Data Mining Techniques
Prediction Results

**Fig. 2.**Proposed Classification Model to Predict Thyroid Disease Study [11]

There are several different methods of data mining. Choosing the right theme depends on the nature Data under study and volumetric blindness. Data mining can be performed in comparison with the data market A data cache. Some of these methods are: (they follow the function of each feature)

- Thinking and drawing conclusions and laws from a reasoning based-case
- Discovery Rule: Search for an exemption, an exempt growth, or a specific relationship in a large part of it.
- Processing Signal: Finding phenomena that are similar to each other
- Neural Nets: developing viable models to predict outcomes. These models are developed on my own, the foundations are deductive from the mindset of insufficiency.
- Fractals: Minimize big data in order to lose information

**In the Study entitled "Diagnosis through Data Mining Techniques"** [12], it aims to diagnose thyroid diseases using neural network methods, voting group and stacking group where classifiers are compared in terms of accuracy, recall and error rate. This study relies on three associated data-extraction algorithms to extract data, and on rapid Miner for data classification and forecast accuracy of each algorithm. To reach the results, the data of 807 thyroid patients in Kashmir has been used. The variants (traits) coming of age, sex, TSH, T3, and T4 as input amounts, and the class value (normal, thyroid deficiency as an output or labeling feature. Through these data, the three methods, the neural network, the voting group method, and the stacking method, have been used to diagnose thyroid diseases and compare these methods to reach whatever is best performed. The study found that the stacking group model showed higher accuracy and less error than the other two methods and showed better accuracy than some models.

It is concluded that application of smart algorithms in data mining and machine learning is to find out the extent to which cancer genes are affected by different types of developed drugs. For example, there is work in progress at the present time for experiments using Microarray technology to produce databases that include measurements of the action of cancer genes in colon cells after applying different types of chemotherapy. In these experiments, two types of chemotherapy are applied to cancer genes from colon cells, and then data-mining algorithms are used to obtain two gene networks, each of which represents the gene function after applying the advanced chemotherapy. The study recommended that the same methods of diagnosis of heart disease and cancer could be applied so as to achieve a high accuracy rate.

**Concerning the Study entitled "To Generate an Ensemble Model for Women Preparation using Data Mining Techniques"**[6], This paper aims to easily identify thyroid symptoms for treatment in women, and two key data mining techniques have been proposed. The first set and the second group are considered machine learning techniques. The first set of the decision tree was created through the installation and neural network and the second group resulting from Bagging and Boosting techniques. Finally, the proposed experiment was conducted by the first set versus the second group. The study data were collected from various sources, such as the Chandan Oncology Center, the Rahwal Thyroid Disease Center and the Rahza Thyroid Diagnosis Center. The total number of samples was 12,000, classified as follows, (5,628 hypothyroidism), (5,522 hyperthyroidism), (850 thyroid cases) and were considered targeted variables, while independent variables were cold intolerance, fatigue, skin, weight, facial swelling, heart rate, memory concentration, and blood pressure. After analyzing the variables, the study concluded that the second group model (Bagging +Boosting) is the best model among others. The study recommended that future research can be conducted using some models of groups with different computer methods and decision tree to collect and associate thyroid disease.

**The article entitled "Molecular Network-based Drug Prediction in Thyroid Cancer" (2019)**, This study aims to identify the impact of the drugs used for thyroid cancer, and has developed a computational pipeline to predict the response and potential target genes for thyroid cancer, by examining the disorders caused by the development of thyroid cancer compared to healthy patients including differential gene expression and joint expression of differential genes, so as to build weighted gene co-expression network analysis (WGCNA), in order to access drugs and genes in which their usage or interference can lead to reverse disorders imposed by thyroid cancer.

Data from 56 ordinary patients and 500 patients with thyroid cancer were examined. The study found that there could be an effect of certain medications on thyroid cancer and may have a real link to thyroid cancer and gland hormone levels, and that Papillary Thyroid Carcinoma (PTC) treatment is a simple and appropriate tool for the treatment of Ultrasound-Guided Percutaneous Ethanol Injection (UPEI). The study concluded that regulating thyroid hormone secretion may be closely related to inhibiting the proliferation of thyroid cancer cells.

**The work entitled "An Empirical Study Disease Diagnosis using data Mining techniques."(2018)**, This study aims to review the techniques of mining in the medical data for diagnosis and prognosis, such as diseases, diabetes, breast cancer, heart diseases and thyroid prognosis. The importance of the study lies in the fact that it will help medical practitioners and analysts diagnose diseases. It discusses some data mining techniques in analyzing medical data such as Artificial Neural Network (ANN), Logistic Regression (LR), Decision Tree (DT), Naïve Bayes (NB), Supporting Vector Machines (SVM), and their role in disease diagnosis. The researchers concluded that the techniques used contribute greatly to the diagnosis of diseases through which the patients' data were analyzed. However, decision trees and artificial neurons show better predictive results than other models, as the decision tree is working faster because it ignores the elements of unhelpful input and contributes to disease prediction because it produces more accurate results. The study recommended that we should study how mining tools operate in data and algorithms in order to analyze data in a way that gives better results

**In the Study entitled "an Intelligent System for Thyroid Diseases Classification and Diagnosis"** [13], This study aims to provide a way to classify and diagnose the thyroid disorders with a description of the disease in addition to providing medical advice, using the Machine (TSH). This study aims to provide a method for classifying and diagnosing patients' thyroid disorders with a description of the disease and also to provide medical advice. By using a TSH implement, "improving particle mobilization" is applied where the user is provided with a page to enter details such as SVM, the classification-supporting vectors. The proposed method is in two phase: The first phase is training, where the thyroid data set, the thyroid hyperactivity data and thyroid hypothyroidism data are taken so that the data set consists of 21 features and is numeric or logical, the most important step in the pre-treatment training phase. Records that do not represent values are also deleted from the data set. Phase 2 is classification and uses a (SVM) machine for classification, and to improve classification accuracy the use of particle-aggregation enhancement (PSO) can be used to improve parameters (SVM) the user interface is given in the selection phase, and the user can enter attribute values based on their test results. KNN is also used to approximate the values lost in the user input. The study found that this method helps significantly to diagnose the type of disease along with health advice. The proposed method can also be applied in the health care industry.

**In the Study entitled “Using K-dependence causal forest to mine the most significant decency relationship among clinical variables for thyroid disease diagnosis"**[14], This study aimed to identify some models of exploration in the data to build computer-assisted medical expert systems, including neural networks and has used three different types, including the multi-layered neural network, the probability neural network, and the neural network for the muzzle of vectors. The study relied on a data set of thyroid diseases from the University of California's. The UCI database contained 335 data sets, with 9,172 real historical cases. Each instance consists of 29 features, which can be classified into 20 categories. The study showed that the characteristics of the data associated with the thyroid disease data set are multi-variable. The domain and characteristics of the features contained are categorical and real, and the task associated with the data set is classification. The study found that Bayesian Network can describe the statement of conditional dependencies included in training data that proved to work efficiently in diagnosis and medical treatment, and the study concluded that a single data mining model cannot deal with all difficult and complex cases.

## 4. Results and Discussion

After presenting and analyzing previous studies, it was found that all studies had agreed that thyroid disease was a difficult disease to detect as easily as some physicians diagnosed the disease through associated symptoms. (Machine Learning, God Learning Algorithms, Classification Algorithms, Reading Tree, Naive Bayes, and Bayesian Compiler) have broken into the most accurate and difficult medical cases such as detection and diagnosis: Thyroid diseases and thyroid cancer and have come up with good results in helping physicians make the right decision at the right time. Data mining techniques have the ability to analyze and read data and patient test results in databases faster and more accurate. As a result, medical errors are reduced and patient treatment and care are improved and accelerated.

Some studies have been characterized by research on women such as study the fact that thyroid disease is more prevalent among women than men. Studies have shown that genes have a role in the emergence of thyroid disease, such as in a study [5][6], or a genetic link between obesity and thyroid disease, as in a study [15]. On the other hand, a showed a relationship between taking some medications and having thyroid disease [16]. The study stated that the decision tree and artificial neural networks could yield better results than other models when conducting data exploration analysis for a particular disease, as shown in figure 3 below[15].

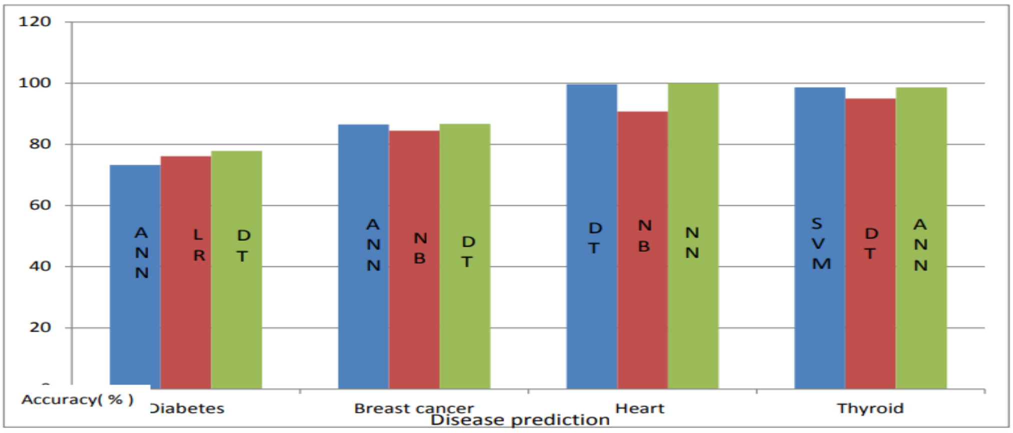

**Fig. 3.**Graphical representation of accuracy for each method for Various Case Prediction Study [15]

The data in the table show data on diabetes, cardiac disease, breast and thyroid cancer, cumulative result analysis, and the techniques used in the chart above are Artificial Neural Networks (ANN), logistical regression (LR), decision tree (DT), and Naïve Bayes. It can be inferred from the graph that the decision tree and artificial neural networks showed better results compared to other models, as the decision tree classifier works better and faster on the trained dataset because it literally ignores useless input features.

One of the most important findings was: There is a lack of studies on the side of data mining techniques in the field of thyroid disease, where a few studies on the topic have emerged from the literature review. On the other hand, it turns out the existence of several studies on data mining techniques and their role in the medical field. Such techniques have proved the effectiveness and efficiency of these proposed tools, systems and applications, in diagnosing the disease and its possible use in diagnosing and developing other different diseases to play a role in health care. Finally, studies have shown the importance of data mining techniques in diagnosing and describing the treatment by reading the data accurately. This helps the physician with analysis, predicting the correct treatment of the patient, and describing the appropriate doses as the treatment of thyroid disease is the precise description of the doses given to the patient.

## 5. Conclusion

It is better for health services to be available quickly and efficiently for everyone, and the turnout of millions to health service centers and the inability of some of them to bear the costs of treatment, which may be high, makes the provision of these services and their software applications not easy. In addition, others are unable to reach these centers in time, or have difficulties with movement. The prosperity of these applications ultimately depends on the extent of their use, so more awareness and education of these applications and how to use them can encourage more people to adopt them, and this will radically change the provision of health care. In the end, we would like to point out what is more important than increasing profits and reducing expenses, which is investing modern technologies in predicting epidemics, developing new treatments (as happened recently in China with the Coronavirus), and avoiding avoidable situations that may lead to death and work. On improving the quality of life in which we live.

This paper presented a critical assessment of a number of studies on “Data Mining and Diagnosis of Thyroid Diseases”, all of which were involved in the goal of determining the best method, the most accurate tool and method for diagnosing thyroid diseases. The studies also called for using the data mining techniques in the medical fields and diagnosing the difficult-to-diagnose diseases such as heart diseases and cancer, and developing these techniques to expand their future functions and be a main factor in the field of diagnosing diseases and providing health care. In light of the findings, the present study recommends that:

- Conducting extensive systematic reviews of studies on the use of data mining techniques in the diagnosis of thyroid gland diseases, particularly in Arabic.
- Making use of data mining techniques in the medical field, medical care, and clinical care fields, and specifically diagnosis neonatal diseases.
- Focus on the creation of warehouses and databases focusing on all data on thyroid patients, since the availability of data and the ability to treat them is an important factor in the diagnosis thyroid diseases when using Data Mining techniques.

It is proposed through this paper that future studies be undertaken to use, improve and develop the methods used to achieve more accurate results, and studies to analyze data mining techniques of breast cancer and vascular diseases.

Privacy and security is one of the most prominent challenges that are considered future work, as it is necessary to strike a balance between the flexibility of using data on the one hand and the privacy and confidentiality of that data on the other hand. And since privacy and confidentiality of information are among the basic rights of the patient, it is necessary to take into account the application of systems that control the quality, nature and users of that data. However, hackers are always targeting medical records, and it is said that you can make more money from stealing health data than from stealing credit card information! In February 2014, the largest health care data theft occurred, as hackers stole records related to eight million patients from a health insurance company in the United States. Then no risk.

## References

[1] S. J. Pasha and E. S. Mohamed, “Ensemble gain ratio feature selection (EGFS) model with machine learning and data mining algorithms for disease risk prediction,” in *2020 International Conference on Inventive Computation Technologies (ICICT)*, 2020, pp. 590–596.

[2] A. Sumathi, G. Nithya, and S. Meganathan, “Classification of thyroid disease using data mining techniques,” *Int. J. Pure Appl. Math.*, vol. 119, no. 12, pp. 13881–13890, 2018.

[3] K. Raza, “Application of data mining in bioinformatics,” *arXiv Prepr. arXiv1205.1125*, 2012.

[4] T. M. Lanigan, H. C. Kopera, and T. L. Saunders, “Principles of genetic engineering,” *Genes (Basel).*, vol. 11, no. 3, p. 291, 2020.

[5] A. B. M. S. Hossain and M. M. Uddin, “Genome as a Tool of Genetic Engineering: Application in Plant and Plant Derived Medicine,” *Int. J. Biotech Trends Technol.*, vol. 8.

[6] D. C. Yadav and S. Pal, “To generate an ensemble model for women thyroid prediction using data mining techniques,” *Asian Pacific J. cancer Prev. APJCP*, vol. 20, no. 4, p. 1275, 2019.

[7] D. C. Yadav and S. Pal, “Prediction of thyroid disease using decision tree ensemble method,” *Human-Intelligent Syst. Integr.*, vol. 2, no. 1, pp. 89–95, 2020.

[8] A. Vinitha, “A survey of thyroid detection using datamining techniques,” *Chief Ed.*

[9] J. Wang, L. Wu, Y. Jin, S. Li, and X. Liu, “Identification of key miRNAs in papillary thyroid carcinoma based on data mining and bioinformatics methods,” *Biomed. Reports*, vol. 12, no. 1, pp. 11–16, 2020.

[10] J. Chen, H. Cao, M. Lian, and J. Fang, “Five genes influenced by obesity may contribute to the development of thyroid cancer through the regulation of insulin levels,” *PeerJ*, vol. 8, p. e9302, 2020.

[11] A. Begum and A. Parkavi, “Prediction of thyroid Disease Using Data Mining Techniques,” in *2019 5th International Conference on Advanced Computing Communication Systems (ICACCS)*, 2019, pp. 342–345, doi: 10.1109/ICACCS.2019.8728320.

[12] U. Sidiq and S. M. Aaqib, “Disease diagnosis through data mining techniques,” in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 275–280.

[13] R. Sudirman, N. Tabatabaey-Mashadi, and I. Ariffin, “Aspects of a standardized automated system for screening children’s handwriting,” in *2011 First International Conference on Informatics and Computational Intelligence*, 2011, pp. 49–54.

[14] L. Wang, F. Cao, S. Wang, M. Sun, and L. Dong, “Using k-dependence causal forest to mine the most significant dependency relationships among clinical variables for thyroid disease diagnosis,” *PLoS One*, vol. 12, no. 8, p. e0182070, 2017.

[15] X. Wang, H. Xia, and Z. Wang, “The research of ear identification based on improved algorithm of moment invariant,” in *2010 Third International Conference on Information and Computing*, 2010, vol. 1, pp. 58–60.

[16] X. Xu *et al.*, “Molecular network-based drug prediction in thyroid cancer,” *Int. J. Mol. Sci.*, vol. 20, no. 2, p. 263, 2019.