MASTER'S THESIS

Interuniversity Master in Statistics and Operations Research UPC-UB

Title: Diagnostic biomarkers identification of arthritis in type 2 diabetic patients, using Artificial Intelligence classification techniques applied in Real World Data database based on GEO microarrays

Author: Helena Bartra

Advisor: José Manuel Mas

Department: Universitat de Barcelona (1004)

University: Universitat Politècnica de Catalunya – Universitat de Barcelona

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

Abstract

Type 2 diabetes mellitus has become an emerging health concern worldwide, affecting both adults and children. This disease can co-exist with other chronic conditions. In the ageing population, the co-existence of diabetes with arthritis is very frequent, exacerbating the health difficulties of these patients. At the same time, omics data applications are growing exponentially intending to determine biomolecular characteristics of specific diseases. Particularly, databases and repositories, including microarray data, are important sources of information that can be exploited to understand biological processes that in some cases could be initially identified as unrelated, such are diabetes and arthritis. The high-throughput data analysis is complex, but it is especially complex when data come from different experiments and objectives. In these cases, the use of Machine Learning approaches is essential to extract patterns that could help to identify biological reasons to link phenotypes unrelated. During this project, a set of diabetes samples obtained from the GEO database were analysed to compare patients with and without arthritis. The analysis was developed to identify common/uncommon patterns to help us determine biological factors associated to suffer arthritis by diabetic patients. A set of algorithms associated with feature reduction and others used as base classifiers to reach our objectives were employed during these analyses, which resulted in 5 proteins as potential classifiers for arthritis biomarkers. The classifiers obtained in a 10-fold cross-validation optimized for balanced accuracy > 80% were interleukin-18 (IL-18), tumour necrosis factor receptor superfamily member 1A (TNFRSF1A), osteopontin (SPP1), interleukin-8 (CXCL8) and interleukin-10 (IL-10); which the novel potential of the SPP1 protein as a new biomarker was highlighted due to its lack of previous reports in the scientific literature. Further studies are needed regarding the usage of these proteins as biomarkers of arthritis in diabetic patients due to their competent classification potential.

Keywords: Type 2 diabetes mellitus; Rheumatoid Arthritis; Biomarkers; Artificial intelligence; Data mining; GEO database; Real World Data.

Index

1. Introduction

Open-access medical research databases have necessarily emerged during the last few decades due to the big volume of omics data that is being collected in thousands of real-patient studies worldwide. Gene Expression Omnibus (GEO) is an example of a database repository of high-throughput gene expression data which gathers immeasurable rich information self-submitted by researchers. Each sample uploaded is classified with a tag related to the pathological condition under study: ill, treated, healthy control... The pathology labels are useful for studying that specific disease; however, they are not easy to tabulate since the patient description entry must be read in order to know the label, a laborious fact that might often require text-mining. In addition, gene expression data from one sample store an enormous amount of information that is not being fully utilized and could be properly exploited for other purposes, such as investigating other diseases that were not previously labelled and eventually optimize medical processes and management strategies. Data mining is used to extract knowledge from big datasets with inherited patterns which might seem hidden, at first sight, making them more potentially useful and profitable. The excellent performance of new patterns discovery has introduced new opportunities and prospects to clinical big-data research such as the identification of new biomarkers for early diagnosis [1].

The problem here is that, since the clinical data is self-submitted by each researcher with different classification standards, there is a high heterogeneity between samples that do not allow the integration of multiple studies. Although several cross-(multi)platform normalization algorithms (ComBat [2], UPC [3], YuGene [4], DBNorm [5], Shambhala [5]) have been recently developed for scaling gene expression data of multiple platforms together, they show generally poor performance. Anaxomics Biotech generated its own normalization method, called CuBlock [6], and is currently programming a new one pending of publication, which will be the algorithm used in this study.

The analyses for the discovery of new biomarkers through large volumes of data are time-consuming and highly costly. The detection and diagnosis of several diseases nowadays remain complicated, slow and sometimes unobjective; leading to adverse health and socioeconomic impacts. The discovery of novel biomarkers or techniques which allow moderately fast recognition of a certain disease has grown exponentially over the last few decades. However, many chronic conditions can occasionally be misdiagnosed or tardily detected in late stages even now, especially if these coexist with diverse comorbidities. A rapid and accurate diagnosis could prevent a wide range of disease-associated complications and optimally manage early-onset diseases, but also could help identify patients with a high risk of acquiring one or more diseases which have not yet started, perhaps avoiding the whole development.

Type 2 diabetes mellitus (T2DM or diabetes) is the most common type of diabetes, affecting 8% of adults around the world [7]. T2DM has spread almost epidemically worldwide with a prevalence constantly increasing; the number of diabetic patients is expected to reach 422 million according to the World Health Organization (WHO) in 2025 [8, 9]. This multi-factorial chronic disease is characterized by hyperglycaemia, insulin resistance and relative insulin deficiency in the liver, skeletal muscles and adipose tissues, among others. It is triggered by several genetic and/or environmental factors

[10, 11] and prevails in the elderly population, representing a major health concern due to the high number of comorbidities this subpopulation suffers from [12].

In the ageing population, T2DM often cooccurs with other chronic diseases, such as arthritis [13]. **Rheumatoid arthritis** (RA) is an autoimmune and inflammatory disorder that affects the small joints of the human body through persistent synovial inflammation, leading to cartilage damage and bone erosion [14]. Systemic inflammation and physical inactivity due to chronic pain caused by RA have resulted as triggering factors of diabetes according to accumulating evidence [15-17], a fact that enforces the connection between both conditions. The co-existence of these two conditions can conduct to very unstable and problematic lifestyle situations [18], a reason why early diagnosis of one/both of these diseases should be of great importance.

The use of diabetes **biomarkers** obtained in the blood is well-established. The current diagnosis for T2DM considers two types of biomarkers: traditional (e.i., the haemoglobin A1C test) and novel (miRNA and proteomic markers). Some of the molecular biomarkers that are being used at the present for T2DM classification are creatinine, interleukin-6 (IL-6), C-reactive protein (CRP), sOB-R, adipokines, ferritin… [19]. The diagnosis of RA consists of a scoring system developed by the American College of Rheumatology and European League Against Rheumatism (ACR/EULAR) called the 2010 RA Classification Criteria. The score takes into account the results of a physical exam, elements of the patient's history and, certainly, biomarkers [20]. The biomarkers currently used for RA diagnosis are rheumatoid factor (RF), autoantibodies against citrullinated proteins (ACPA), erythrocyte sedimentation rate (ESR), and CRP. These criteria of classification as RA have a sensitivity of 84% and specificity of 60% [21]. Indeed, in a patient with recent onset of symptoms, sensitivity is much lower (ranging from 35% to 50%) even if specificity remains high. Hence, RA would benefit from novel biomarker development for diagnosis where new biomarkers are still needed [22].

Considering the clinical necessities of rapid diagnosis, data mining techniques and a new normalization method have been used for the identification of diagnostic biomarkers of arthritis in diabetic patients, to demonstrate that integration of multiple GEO datasets can rapidly extract biological information, leading to optimization of disease management in clinical practices.

2. Objective

The aim of this work consists of reaching two main objectives:

- To provide a standard protocol which will allow us to characterize the clinic profile of patients using GEO database high-throughput data that have not been previously labelled or detailed by the researchers when submitting the microarrays to the public repository, and try to identify if these patients are suffering from other diseases.
- To check the plausibility of this protocol by applying the technique to diabetic patients from the GEO database with the aim to classify them into arthritis or not.

3. Methods

3.1. Generation of disease characterizations

Molecular characterization of a pathological condition consists of the definition of the main pathophysiological processes (or motives) which lead to the development of the condition, according to the general definitions used by the scientists studying the disease. These motives are formed by a set of proteins, called effector proteins, whose gain or loss of activity has been reported in the scientific literature to be involved in the pathophysiology of the condition. For instance, adiponectin, a protein contributing to peripheral insulin sensitivity, is inhibited (or downregulated) in type 2 diabetic patients [23]. A set of effector proteins working in the same biological process constitute a motive and a set of motives build the molecular characterization.

All the information gathered in a molecular characterization is archived in the Biological Effectors Database (BED) [1], which describes more than 300 clinical conditions or phenotypes as sets of genes and effectors proteins that can be active, inactive or neutral. In a genetic network, genes are active when they are expressed (experimentally detected as over-expressed) and inactive when they are repressed (experimentally detected as under-expressed). On the other hand, neutral proteins are neither active nor inactive for a particular phenotype.

The protocol followed to generate a characterization involves a curated review of the scientific literature regarding the pathogenesis of the disease. A specific search for reviews of the condition of interest is performed in the PubMed database. The results of this search are evaluated at the abstract level, and if molecular information describing the condition pathophysiology is found, it is thoroughly reviewed to identify protein/gene candidates to be condition effectors, i.e. proteins whose activity (or lack thereof) is functionally associated to the development of the condition.

If the evidence of the implication of a candidate in the condition is judged not consistent enough to be assigned as an effector, an additional PubMed search has to be performed specifically for the candidate, including all the protein names according to UniProtKB. If novel candidates are identified in this phase, they are included as effectors following the same criteria and protocol.

The specific searches used for the characterizations of diabetes and arthritis:

- ("Type 2 Diabetes" [Title] AND ("Molecular" [Title] OR "Pathogenesis" [Title] Or "Pathophysiology" [Title])) AND (English [Language] OR Spanish [Language]) AND Review[ptyp]
- ("Severe rheumatoid Arthritis" [Title] OR “Rheumatoid Arthritis" [Title]) AND ("Molecular" [Title] OR "Pathogenesis" [Title] OR "Pathophysiology" [Title])) AND (English [Language] OR Spanish [Language]) AND Review [Ptyp]

3.2. Description of the data science methods applied to the data mining process

Applications of big data as classification problems in health science imply the identification of biological elements to reach the classification objective. The gene-expression-based classification of unlabeled patients from the GEO database repository can be solved through the concept of data mining. Data mining refers to the process of extracting potentially useful knowledge hidden in a large amount of incomplete, noisy and random practical application data; by means of database technology, statistics, machine learning (ML) and pattern recognition. Data-mining technology is not used to replace traditional statistical analysis techniques, but to expand statistical analysis methodologies. The main analytical method used by data mining to train models and then predict outcomes is ML [24].

Data mining has two types of objectives: description and prediction. The goal of descriptive models is to find inherited patterns of associations in the data in order to generalize them; while predictive models aim to estimate unknown or future values of other variables of interest based on other variable values [24].

The mathematical expression of a classifier or feature is the discriminant function that allows to correctly classify the types of samples of the databases by using one or more independent variables. The available tool including all ML processes in Anaxomics encapsulates algorithms published and referenced for all the steps in Figure 1. The implementation of this algorithm has been developed in Matlab and R softwares and has been used in multiple scientific publications (Jorba et al., 2020 [25]; Bertran et al., 2022 [26]; Farres et al., 2021 [27]; Carcereny et al., 2021 [28]).

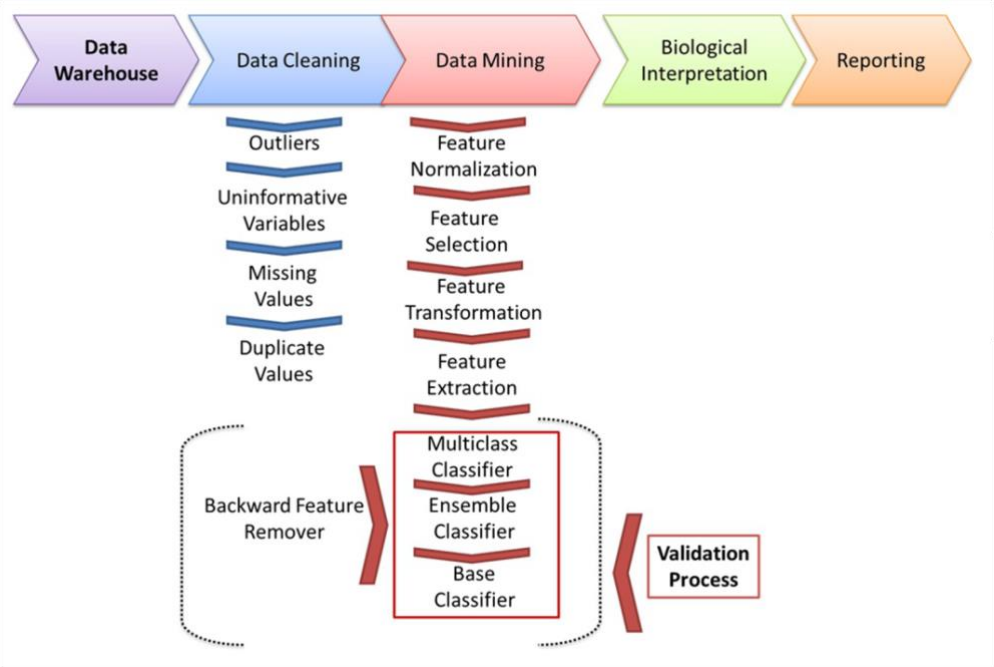

Figure 1. Scheme of the Anaxomics Data Science Strategy for classifiers with more than one element.

The data science analytical process has been divided into the following processes and subprocesses: database and patients' selection (3.3.1), data cleaning (3.3.2), delimitation of the dimensionality (3.3.3), data normalization (3.3.4), feature selection (3.3.5), data mining algorithms (3.3.6), model evaluation parameters (3.3.7) and cross-validation (3.3.8). The methodology used to follow these steps is thoroughly described in Jorba et al. 2020 [25].

3.2.1. Database and patients selection

Publicly available datasets containing high-throughput gene expression data can be found in several medical databases. Due to its numerous advantages and an immense amount of collected genomics data, the GEO database was selected as the source for finding samples for the data mining process. GEO functions with the self-submission of the researchers who performed the data collection in a determined study. It is structured in three levels: platforms record (sequencer or array used, GPLxxx), sample record (condition of individual samples, GSMxxx), and series record (gathers samples of the same study, GSExxx) [29].

Samples of patients labelled as T2DM or RA (accompanied by healthy control samples) were downloaded from this public repository, with the following criteria: the organism was "Homo Sapiens" and the study type was "expression profiling by array". The keywords used to detect datasets of the conditions of interest were "TYPE 2 DIABETES MELLITUS" AND "RHEUMATOID ARTHRITIS". Regarding diabetes, 246 samples were extracted from experiments datasets published by researchers, from which 123 samples were diabetic patients and 123 were healthy controls. In the case of arthritis, a total of 600 samples were obtained, from which 300 were samples of patients with arthritis and 300 were from control subjects.

3.2.2. Data cleaning

The main objective of the data cleaning process is to treat the original data and prepare it for the application of the data mining process with a better chance of success. This process includes some sub-analyse, such as outlier detection, uninformative variables detection, missing values treatment and duplicate information management. The data cleaning process produces a new dataset without missing values, which will be used to train the data mining models.

Different data mining processes, including different data cleaning treatments, were conducted for both conditions in order to find the best option in which the classifier would obtain the highest accuracy. Nevertheless, each model was treated with different methodologies in each treatment:

- Outliers detection. Some models used the Univariate Media Deviation (UMD) method, which was applied to variables with more than 10 samples and departing more than 3.00 times the variation of the median deviation. Other samples were not treated.
- Uninformative variables treatment was not applied thanks to the reduction of dimensionality problem that was afterwards conducted. It is explained in section 3.2.3.

- Missing values treatment. The following methods were tested in order to select the one providing higher accuracy: Remove all samples/variables with missing values and Greedy optimization of Samples and Variables (Sample Cost Remove weight 10, 2 and 1).
- Duplicate values treatment. Remove error clusters method acts via removing all samples with same in and different out.

3.2.3. Dimensionality reduction

A high number of variables accompanied by a limited number of patients is very frequent in clinical research, which leads to a dimensionality problem. Dimensionality reduction of the problem is a common procedure since a high number of features and a limited sample size can lead to the obtention of excessive false positives. It is known that most of the genes presented in microarray data sets are not relevant for a specific classification problem, and their presence often causes a loss in classification accuracy, which is known as a manifestation of the "curse of dimensionality" [30, 31]. In order to diminish the problems associated with the data dimensionality, a reduction of the number of variables from the dataset was achieved by removing all the genes from the microarray whose role in the disease was not previously reported in the scientific literature (effector protein included in the molecular characterization). Thus, the number of features considered in the data mining process was limited to effector proteins from the molecular characterization, decreasing the probability of obtaining false positives. The remarkable advantage of this procedure is that the probability of obtaining a good classifier will increase since this classifier will be beforehand associated with the pathophysiological process of the condition.

3.2.4. Feature selection procedure

Besides the reduction of the number of variables from the original dataset to the selection of only effector proteins, other feature selection techniques were developed to identify and remove irrelevant genes from data sets before training classification models. The feature selection methods that were employed are: Ridge regression Feature Weights [32], LASSO [33], Elastic Net [34], SF linear regression classifier, ReliefF [35], Entropy and uncorrelation of selected features [36 Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011] T-Student test and uncorrelation of features, Entropy [36], T-Student test [37], Wilcoxon rank sum test [38], Wilcoxon rank sum test plus features uncorrelation [38], Bhattacharyya [39], Random Forests [40] and GLM random sets [41].

3.2.5. Data normalization

Different processing and normalization procedures previously applied to public databases often impede data integration from multiple datasets in Big Data experiments. To overcome these challenges, Anaxomics' method for microarray data processing and normalization that allows the integration of previously-processed datasets has been used in this data mining study (in publication process). This method is intended to be published so the exact steps cannot be described in this thesis due to confidentiality concerns. Briefly, there is first a prediction and reversion of mathematical transformations to uniformize the orders of magnitude; and second, an application of a normalization

technique that considers the expression of House Keeping genes across the whole GEO database is utilized, to reduce the impact of sample-specific background corrections and then, readjust the different normalization techniques. This was performed by using regression algorithms from a 40 HK genes background dataset, which was built using the median value of expression of the 40 genes with the least cross-tissue variability as found in the literature. By applying this technique, we were able to combine microarray profiles from multiple sources independently of the initial processing.

3.2.6. Data mining methods

Data mining based on clinical big data can produce effective and valuable knowledge, which is essential for accurate clinical decision-making and risk assessment. Data-mining algorithms enable the realization of these goals. The data-mining method depends on whether or not dependent variables (labels) are present in the analysis. Predictions with dependent variables (labels) are generated through supervised learning, which can be performed by the use of linear regression, generalized linear regression, a proportional hazards model (the Cox regression model), a competitive risk model, decision trees, the random forest (RF) algorithm, support vector machines (SVMs) [24] and others.

3.2.7. Data mining classification algorithms

The Data Mining process uses the cleaned database created in the data cleaning process to identify the desired classification patterns. This process is the most computationally expensive in the data science pipeline since it is in charge of evaluating sets of mathematical models to identify the best classifier. There is a set of methods used to divide the dataset into the classes included in the analysis. Each method uses its own strategy and some of them are derived from others. These sets of methods cover quite well all possibilities for all types of data and data distribution.

The base classifiers most used in this study are support vector machines (SVM) [41 20(3):273-297, September 1995. [Vladimir,Vapnik] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.], Multilayer Perceptron (MLP) [42], linear regression + threshold and optimal quadratic threshold [43].

3.2.8. Model evaluation parameters

In order to assess model performance, commonly used metrics in machine learning such as accuracy, error rate, sensitivity, specificity, and predictive values were computed.

Accuracy and error rate. Classification accuracy is defined as the ratio between the total number of correctly classified samples and the total number of samples (Equation 1). From the accuracy, one can compute the error rate, which represents the number of misclassified samples (Error Rate = 1 – accuracy).

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

Sensitivity and specificity. Sensitivity and specificity are statistical measures of the performance of biomarkers using a binary classification test. This represents the number

of correctly classified samples in the positive and negative classes respectively. This provides a more informative metric when working with imbalanced data sets than accuracy [22].

$$Sensitivity = \frac{TP}{TP + FN}; \quad Specificity = \frac{TN}{TN + FP}$$

Precision and Negative Predictive Value. Precision, also known as positive predictive value (PPV), and negative (NPV) predictive values reflect on the number of correctly predicted samples for each class over all the predicted samples for that class [22].

$$Precision = \frac{TP}{TP + FP}; \qquad NPV = \frac{TN}{TN + FN}$$

An optimal biomarker would aim to achieve 100% sensitivity (i.e., predict all people with the condition) and 100% specificity (i.e., not predict anyone from the control group). For any biomarker, there is usually a trade-off between the measures and their impact, setting acceptable limits and allowing detection of false positives (lowering specificity), but limiting false negatives (increasing sensitivity) [22].

3.2.9. Validation of the model

The Validation process is designed to ensure that the biological conclusions reached with the available data will be generalized to new samples. Validation is used for estimating the performance of a predictive model. This process is performed against the original database, after the mathematical models are produced.

The final strategy selected for this DM was 10 K-FOLD cross-validation. This validation consists of the estimation of the threshold value for 10 random subsets of all available samples, and the final threshold is determined as the average of the all thresholds determined. In more detail, the model is given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset). The goal of cross-validation is to define a dataset to test the model in the training phase (i.e., the validation dataset) in order to limit problems like overfitting. In K-Fold Cross-Validation, the data are split into K roughly equal-sized parts. For the kth part, the model is trained on the other K- 1 parts of the data, and the prediction error of the fitted model is computed when using it to predict the kth part of the data [44].

By this, those candidates with higher generalization capability are selected. Cross-validated accuracy was used as classifier optimization and quality measure, together with cross-validated p-value.

4. Results

4.1. Characterization results

4.1.1. Type 2 Diabetes Mellitus characterization

Type 2 Diabetes Mellitus (T2DM) is a metabolic disorder which has become of the most common diseases worldwide. Its pathophysiology is mainly defined by a combination of defective insulin secretion by pancreatic β-cells and the non-response of insulin-sensitive tissues. Here, the molecular mechanisms working in the synthesis and release of insulin are dysregulated, leading to defects in the metabolic balance of our bodies [45].

T2DM usually occurs in adulthood, but young people can also suffer from this disease. Although T2DM patients are generally independent of exogenous insulin, they may need it when blood glucose levels are not well controlled with diet alone or with oral hypoglycaemic drugs. In addition, people with T2DM are often accompanied by complications, such as cardiovascular diseases, diabetic neuropathy, nephropathy, and rheumatoid arthritis. The reason why these comorbidities co-exist is due to high prevalence and shared risk factors [46].

Risk factors for the development and/or progression of T2DM include genetics/family history (DNA damage and alterations). It can also be both pre-and post-natal environmental factors, such as suboptimal intrauterine environment, low birth weight, obesity, inactivity, gestational diabetes and advancing age [47].

T2DM molecular characterization has been generated and it consists of 5 motives, comprising a total of 136 unique effector proteins (Table 2).

Table 1. Summary of T2DM defined motives and the effector proteins associated with each of them.

Motive ID	Motive name	# Proteins
1	B-cell dysfunction	61
2	B-cell destruction	18
3	Insulin resistance	54
4	A-cell dysfunction	21
5	Increased adiposity	8

Two major defects are associated with T2DM: **insulin resistance** and **β-cells failure** [48]. In the early stages of the disease, pancreatic β-cells adapt to insulin resistance by increasing mass and function. As nutrient excess persists, hyperglycaemia and elevated free fatty acids negatively impact β-cells function [49].

Table 2. Example of T2DM molecular characterization. Only 3 proteins of each motive are shown as representation. Causative effect: "+1" if the protein is activated/overexpressed in diabetic patients; "-1" if the protein is inactivated/underexpressed in diabetic patients.

Motive ID	Effector Protein Name	Gene name	Uniprot code	Causative Effect	Reference
1	Glucagon-like peptide 1 receptor	GLP1R	P43220	-1	[50]
1	Hemoglobin subunit alpha	HBA1	P69905	1	[51]
1	Hepatocyte nuclear factor 1-alpha	HNF1A	P20823	-1	[52, 53]
1	Other proteins identified as effectors of the motive 1.				
2	Insulin-like growth factor 1 receptor	IGF1R	P08069	-1	[54]

Motive ID	Effector Protein Name	Gene name	Uniprot code	Causative Effect	Reference
2	Interferon gamma	IFNG	P01579	1	[54]
2	Interleukin-1 beta	IL-1B	P01584	1	[54] [55, 56]
2	Other proteins identified as effectors of the motive 2.				
3	Insulin receptor substrate 2	IRS2	Q9Y4H2	-1	[57, 58]
3	Insulin-like growth factor I	IGF1	P05019	-1	[59]
3	Interferon gamma, IFN-gamma	IFNG	P01579	1	[60]
3	Other proteins identified as effectors of the motive 3.				
4	Glucagon receptor, GL-R	GCGR	P47871	1	{24736842;19749172
4	Growth hormone receptor, GH receptor	GHR	P10912	9	{24736842;19749172
4	Hematopoietically-expressed homeobox protein	HHEX	Q03014	-1	[23, 61]
4	Other proteins identified as effectors of the motive 4.				
5	Interleukin-6	IL-6	P05231	1	[62]
5	Interleukin-8	IL-8	P10145	1	[62]
5	Transmembrane protein 18	TMEM18	Q96B42	9	[23]
5	Other proteins identified as effectors of the motive 5.				

4.1.2. Rheumatoid arthritis characterization

Rheumatoid arthritis (RA) is a chronic autoimmune disease that involves the joints. It is characterized by synovial inflammation and hyperplasia, cartilage and bone destruction and systemic complications such as pulmonary, cardiovascular, psychological and skeletal disorders. RA is influenced by both genetic and environmental factors [63, 64].

RA affects approximately 0.5–1% of the population worldwide, occurring in females more frequently than males; it is diagnosed mainly in individuals aged 40–60 years [65]. The exact aetiology of rheumatoid arthritis is unknown, different factors are related: genetic predisposition, environmental factors and hormonal factors [66].

4 different motives were described for RA pathophysiology with a total of 158 non-duplicated effectors identified (Table 3).

Table 3. Summary of RA defined motives and the effector proteins associated to each of them.

Motive ID	Motive name	# Proteins
1	Autoimmune response	20
2	Synovial inflammation	97
3	Articular destruction	38
4	Bone erosion	37

RA is characterized by an **autoimmune response** and the **inflammation of the synovium**, followed by the **destruction of the joints and bone erosion**. The inflammatory process of RA is first initiated in peripheral lymphoid organs where dendritic cells present self-antigens to autoreactive T cells, which in turn activate autoreactive B cells via cytokines and co-stimulatory molecules [67]. This leads to the secretion of autoantibodies that can be detected in the serum of RA patients, of which rheumatoid factor (RF) and anticitrullinated protein antibodies (ACPA) are the most prominent. The inflammation drives bone resorption by promoting the differentiation of osteoclasts, resulting in a change in the delicate balance between bone resorption and bone formation [68].

Table 4. List of effector proteins included in RA characterization. Not all protein are listed here due to confidentiality concerns. Causative effect: "+1" if the protein is activated/overexpressed in RA patients; "-1" if the protein is inactivated/underexpressed in RA patients.

Motive ID	Effector Protein Name	Gene name	Uniprot code	Causative Effect	Reference
1	Antithrombin-III	ANT3	P01008	-1	[69]
1	C-C motif chemokine 20	CCL20	P78556	1	[70]
1	Apoptosis inhibitor expressed by macrophages	CD5L	O43866	1	[71]
1	Other proteins identified as effectors of the motive 1.				
2	Inhibitor of nuclear factor kappa-B kinase subunit epsilon	IKKε	Q14164	1	[72]
2	FAS-associated death domain protein	FADD	Q13158	-1	[73]
2	Allograft inflammatory factor 1	AIF-1	P55008	1	[74]
2	Other proteins identified as effectors of the motive 2.				
3	A disintegrin with thrombospondin motifs 18	ADAMTS18	Q8TE60	1	[75]
3	Cathepsin S	CATS	P25774	1	[76, 77]
3	Cathepsin K	CTSK	P43235	1	[68, 78]
3	Other proteins identified as effectors of the motive 3.				
4	Transforming growth factor beta-1 proprotein	TGF-β	P01137	1	[68, 74]
4	Tumour necrosis factor	TNF-α	P01375	1	[74]
4	Granulocyte-macrophage colony-stimulating factor	CSF2	P04141	-1	[79]
4	Other proteins identified as effectors of the motive 4.				

4.2. Data mining classification results for diabetes and arthritis

According to data mining results, a total of 246 and 292 different raw classification models were obtained, respectively for diabetes and arthritis. Statistically non-significant classifiers were discarded using a threshold of Cross-Validated Accuracy p-value < 0.05 [80 1995.]. Finally, after k-fold cross-validation (k=10) [44] was applied, the proteins were sorted by the balanced accuracy [81] of the classification, which resulted in 21 and 93 models for diabetes and arthritis.

Table 5 shows the models obtained for diabetes. Table 6 presents the models obtained for arthritis which will be used for the prediction (a complete list of arthritis classifiers can be found in Annex 1). For this study, only the 16 proteins (or pairs of proteins) with the highest balanced accuracy that were included in both characterizations were selected as best-classifier proteins.

Table 5. Models of diabetes classification statistically significant using a threshold of Cross-Validated Balanced Accuracy > 0.80. TP: True Positive; TN: True Negatives; FP: False Positives; FN: False Negatives; Balanced ACC: Cross-Validated Balanced Accuracy; p-value: Cross-validation p-value. PRE: Precision; SNS: Sensibility; SPC: Specificity; Outliers: Outliers detection method (NA or UMD). SVM: Support Vector Machines; MLP: Multi-Layer Perceptron.

Model	Feature	TP	TN	FP	FN	Balanced ACC	p-value	PRE	SNS	SPC	Outliers	Features removed	Samples removed	Base classifier
MD1	**CBL, GIP**	60	112	11	2	0.94	1.53E-34	85	97	91	NA	0	98	SVM
MD2	**CASP3, HSF1**	55	112	11	7	0.90	5.64E-28	83	89	91	NA	0	98	SVM
MD3	**ERN1, PLAGL1**	94	105	18	7	0.89	2.12E-35	84	93	85	UMD	88	153	SVM
MD4	**PRKCI, CBL**	88	113	10	18	0.87	3.49E-33	90	83	92	UMD	105	119	SVM
MD5	**CBL, IL-10**	87	113	10	19	0.87	2.17E-32	90	82	92	UMD	105	119	SVM
MD6	**CBL, CASP3**	91	108	15	15	0.87	1.91E-31	86	86	88	NA	21	66	SVM
MD7	**ATP2A3, HSF1**	50	112	11	12	0.86	6.31E-23	82	81	91	NA	0	98	SVM
MD8	**IL-10, IGF1R**	82	97	26	7	0.85	6.35E-27	76	92	79	UMD	88	153	SVM
MD9	**IL-10, HHEX**	86	110	13	20	0.85	4.39E-29	87	81	89	UMD	105	119	SVM

Model	Feature	TP	TN	FP	FN	Balanced ACC	p-value	PRE	SNS	SPC	Outliers	Features removed	Samples removed	Base classifier
MD10	**IL-10, JUN**	94	100	23	12	0.85	2.65E-28	80	89	81	UMD	88	153	SVM
MD11	**CASP3, TCF7L2**	87	98	25	10	0.85	1.23E-26	78	90	80	NA	21	66	SVM
MD12	**GHR, PRKCZ**	55	97	26	7	0.84	2.30E-19	68	89	79	NA	0	98	SVM
MD13	**CBL, P2RX4**	85	87	36	4	0.83	8.91E-25	70	96	71	NA	21	66	SVM
MD14	**ATP2A3, IL-18**	79	95	28	10	0.83	5.70E-23	74	89	77	NA	21	66	SVM
MD15	**P2RX5, TXNDC5**	80	87	36	4	0.83	1.20E-23	69	95	71	UMD	88	15	SVM
MD16	**TNFRSF1A,PLAGL1**	89	100	23	17	0.83	2.38E-24	79	84	81	UMD	88	15	SVM
MD17	**IL-10, TXNDC5**	94	87	36	7	0.82	3.45E-24	72	93	71	UMD	105	119	SVM
MD18	**CAP1, PPP1R3A**	54	116	7	25	0.81	5.48E-22	89	68	94	NA	0	98	MLP
MD19	**ATP2A3, P2RX7**	77	93	30	12	0.81	2.76E-20	72	87	76	NA	21	66	SVM
MD20	**FOXO1, HSF1**	52	95	28	10	0.81	8.23E-16	65	84	77	NA	21	66	SVM
MD21	**JUN, PLAGL1**	94	89	34	12	0.81	6.20E-22	73	89	72	UMD	88	153	SVM

Table 6. Models of arthritis classification statistically significant using a threshold of Cross-Validated Balanced Accuracy > 0.80. Only the models which will be used in the next steps of prediction are shown. The complete list can be found in Annex 1. TP: True Positive; TN: True Negatives; FP: False Positives; FN: False Negatives; Balanced ACC: Cross-Validated Balanced Accuracy; p-value: Cross-validation p-value. PRE: Precision; SNS: Sensibility; SPC: Specificity; Outliers: Outliers detection method (NA or UMD). SVM: Support Vector Machines; LR+THR: Linear Regression + Threshold; O-THR: Optimal threshold; OQ-THR: Optimal Quadratic Threshold.

Model	Feature	TP	TN	FP	FN	Balanced ACC	p-value	PRE	SNS	SPC	Outliers	Features removed	Samples removed	Base classifier
MA1	**IL-18, TNFRSF1A**	273	299	1	27	0.95	4.42E-138	100	91	100	NA	1	0	SVM
MA2	**IL-18, IL-10**	217	297	3	83	0.86	2.71E-88	99	72	99	NA	1	0	SVM
MA3	**IL-18**	203	300	0	97	0.84	2.53E-85	100	68	100	NA	0	203	LR+THR
MA4	**IL-18**	203	300	0	97	0.84	2.53E-85	100	68	100	NA	1	0	LR+THR
MA5	**IL-18**	203	300	0	97	0.84	2.53E-85	100	68	100	UMD	148	0	LR+THR
MA6	**IL-18**	203	300	0	97	0.84	2.53E-85	100	68	100	UMD	120	63	O-THR
MA7	**IL-18**	203	300	0	97	0.84	2.53E-85	100	68	100	NA	1	0	LR+THR
MA8	**IL-18**	203	300	0	97	0.84	2.53E-85	100	68	100	UMD	69	201	LR+THR
MA9	**IL-10**	203	299	1	97	0.84	3.91E-83	100	68	100	NA	1	0	O-THR
MA10	**IL-10**	203	299	1	97	0.84	3.91E-83	100	68	100	UMD	69	201	O-THR
MA11	**CXCL8**	259	229	71	41	0.81	1.03E-57	78	86	76	NA	0	203	OQ-THR
MA12	**CXCL8**	259	229	71	41	0.81	1.03E-57	78	86	76	NA	1	0	OQ-THR
MA13	**CXCL8**	259	229	71	41	0.81	1.03E-57	78	86	76	UMD	148	0	OQ-THR
MA14	**CXCL8**	259	229	71	41	0.81	1.03E-57	78	86	76	UMD	69	201	OQ-THR
MA15	**SPP1**	277	206	94	23	0.81	1.34E-58	75	92	69	NA	1	0	O-THR
MA16	**SPP1**	277	206	94	23	0.81	1.34E-58	75	92	69	NA	1	0	OQ-THR
MA…	…													

We observed that diabetes models presented a lower accuracy than arthritis', probably due to the inferior sample size of original data and the deletion of some samples compared to the arthritis models in the data cleaning process. It could also be caused by higher difficulty in classifying diabetic patients through gene expression profiles due to the multidiverse factors that cause diabetes compared to arthritis. On the other hand, arthritis classification models seemed to identify better classifiers which can correctly categorize the patients in healthy vs arthritis with superior sensitivity and specificity.

Anyway, the objective of the project is to classify diabetic patients with/without arthritis. Then, the interest here is to obtain good arthritis classifiers to later apply them to diabetic patients.

Regarding the base classifiers used, the classifier which could better model the data is apparently SVM: 95% of diabetes models and 25% of arthritis models used SVM.

In order to facilitate the interpretation of the models, graphical representations accompanied by the main quality parameters are presented here for the best classifiers obtained in diabetes and arthritis, chosen as examples.

Diabetes classifier 1 (MD1) is composed of two features: E3 ubiquitin-protein ligase CBL (CBL) and gastric inhibitory poly (GIP). The feature selection method used was random forest. Also, SVM was used as a base classifier and the cost function was balanced accuracy. A 10 K-FOLD cross-validation was used as the validation process.

Figure 2 displays (A) the sample distribution graph and (B) the ROC curve. Figure 1A represents the distribution of the samples in a 2D plot. The line that separates the background colours is the decision boundary defined by the mathematical function of the classifier and by the best discrimination threshold. Each dimension of the graph stands for a variable/feature of the classifier (in this case, CBL on the x-axis and GIP on the y-axis). The Receiver Operating Characteristic (ROC) is a graphical representation that illustrates the performance of a binary classifier system as its discrimination threshold varies. The ROC curve plots the True Positive Rate (TPR or sensitivity) against the False Positive Rate (FPR or false detections). Balanced Accuracy (ACC) is measured by the area under the ROC curve. An area of 1 represents a perfect classifier; an area of 0.5 (straight line between 0-0 and 1-1) represents a useless classifier.

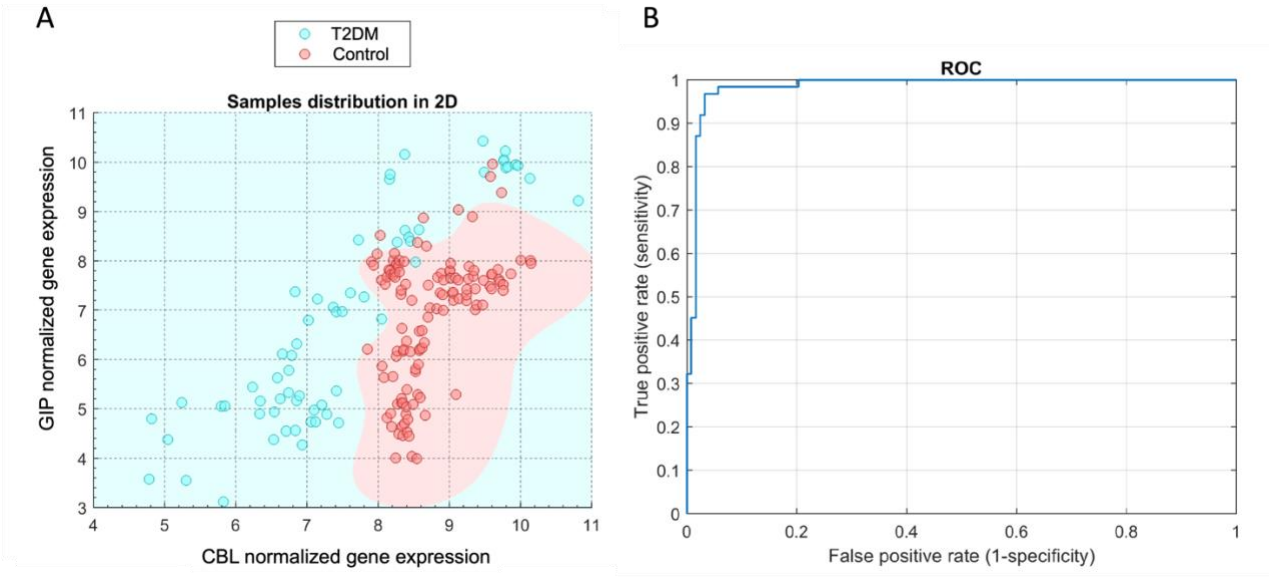

Figure 2. (A) sample distribution graph and (B) the ROC curve of the classifier MD1, composed of CBL and GIP.

Arthritis classifier 1 (MA1) also comprises two features (Figure 3): interleukin-18 (IL-18) and tumour necrosis factor receptor superfamily member 1A (TNFRSF1A). The feature selection method used was random forest. Also, SVM was used as a base classifier and

the cost function was balanced accuracy. A 10 K-FOLD cross-validation was used as the validation process.

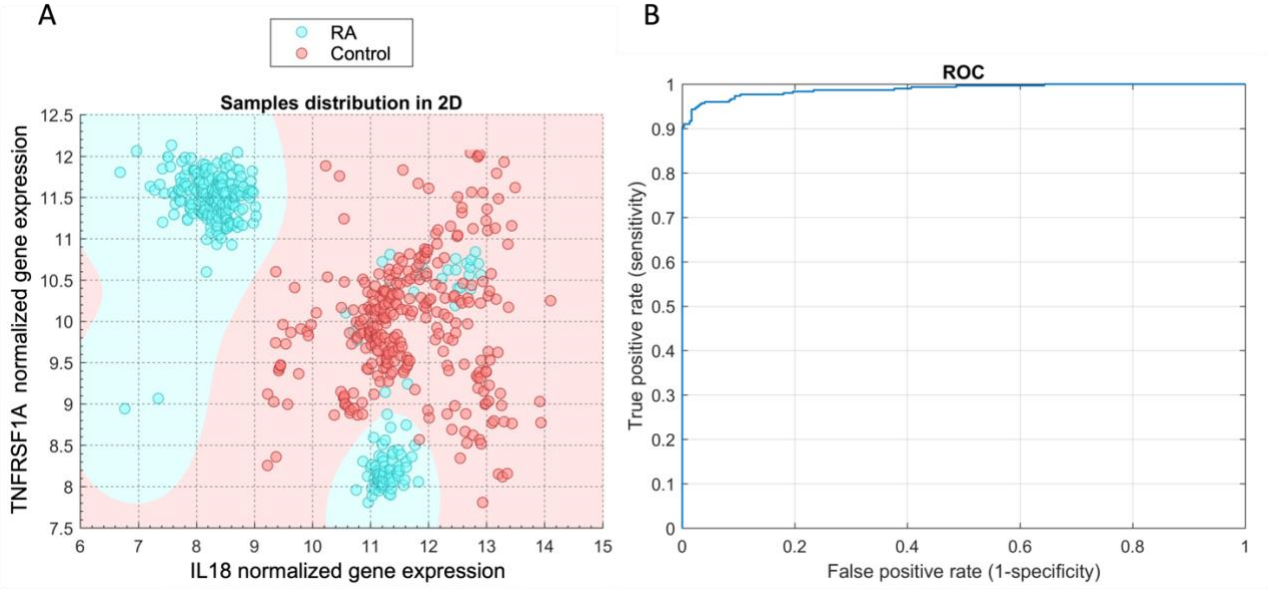

Figure 3. (A) sample distribution graph and (B) the ROC curve of the classifier MA1, composed of IL-18 and TNFRSF1A.

In the case of a one-feature classifier, as in MA9 (Figure 4), the graphical representations are modified according to their dimensions. The boxplot represents the descriptive statistical parameters of each cohort, which allows observing whether their means are separated or whether their value distributions overlap. The samples distribution graph based on one variable presents the distribution of the samples concerning the protein used by the classifier. The dotted red line shows the threshold that gives the best accuracy. The likelihood graph shows the likelihood function (as a synonym of probability function) for a prediction, displaying the value obtained for the prediction for each sample. It represents the proportion of samples of each cohort that correspond to a specific classification value. This representation allows the identification of the range of predictive values with better accuracy (range with a clear separation of or no overlap between the cohorts).

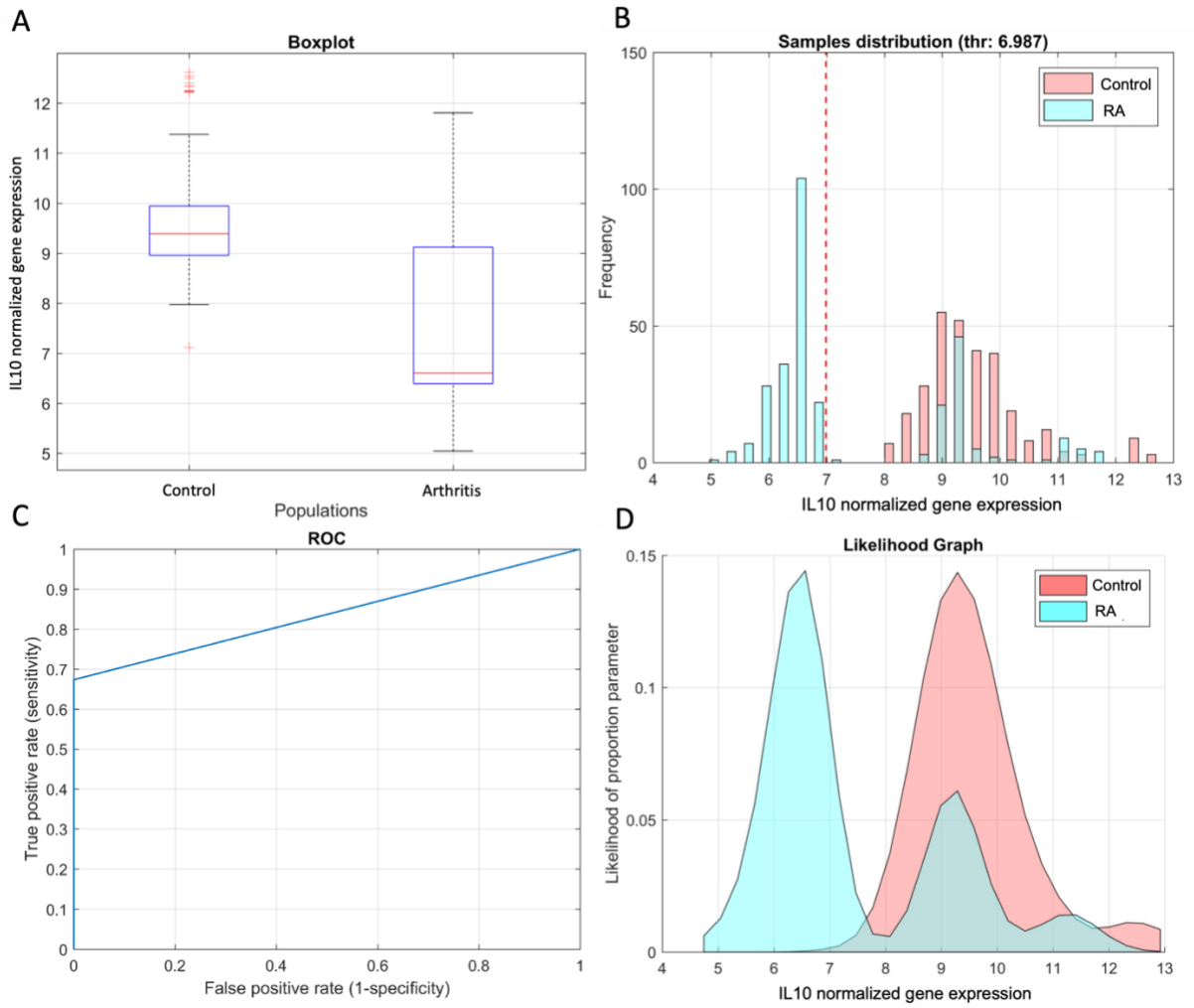

Figure 4. (A) boxplot representation of the means of normalized gene expression for both cohorts (B) sample distribution graph and (C) the ROC curve and (D) the likelihood graph of the classifier MA9, composed of IL-10.

4.3. Application of data mining algorithms to predict diseases in unlabelled patients

After the calculation of the classification models, the classifiers with accuracy > 80% were selected to predict these patterns in new patients. In this case, diabetic patients were predicted to suffer from or to be at risk of suffering from arthritis. In addition, patients with arthritis were also predicted to be diabetic.

As the features were limited to BED effector proteins, only models with classifiers that were effector proteins of the pathology suffered by the patients to be predicted could be used. In other words, a certain classifier can be used to predict only if its gene expression is recorded in the dataset of the patients that will be classified.

For the prediction, the algorithms calculated for arthritis were applied to classify the diabetic patients that were used for training the diabetes classifiers.

Under those circumstances, 16 models were able to classify diabetic patients as arthritis/no arthritis whereas only 1 model was able to classify patients with arthritis as diabetic or not. From the 16 models, 5 unique proteins could be used to classify:

osteopontin (SPP1), interleukin-8 (CXCL8), interleukin-10 (IL-10), interleukin-18 (IL-18), and tumour necrosis factor receptor superfamily member 1A (TNFRSF1A).

A novel candidate for the classification of diabetic patients into arthritis or no arthritis should accomplish some requirements. First, the genes that compose an arthritis classifier should not coincide with a diabetes classifier. In this case, the use of this marker would result in false positive results due to the intersection of both conditions. Second, we want to make sure that this protein is not being used already as a biomarker. Otherwise, the biomarker resulting from this study would not provide any new information.

The best classifier was IL-18 accompanied by TNFRSF1A, according to its high cross-validation balanced accuracy (95%). This model classified 106 diabetic patients in 22 with arthritis (21%) and 84 without arthritis (79%). The biological context based on scientific literature about this protein as an arthritis biomarker and its role in both diseases is explained in the next section, accompanied by the contextualization of the rest of the proteins.

On the other hand, the best classifiers to identify whether patients with arthritis could present T2DM are IL-10 and the transcription factor JUN showing a balanced accuracy of 85% and a significant p-value. From 300 checked patients, this classifier categorized 222 arthritis patients with diabetes (74%) and 77 without diabetes (25%).

4.4. Biological interpretation of the DM results

To discover the biological role of these proteins identified as classifiers and determine if they are novel biomarkers, we searched in the scientific literature for evidence linking them with both conditions' history.

Then, we propose 5 proteins that could potentially help to identify diabetic patients at risk of developing RA. The biological interpretation of obtaining one gene as a classifier by means of data mining procedures is that this resulting gene must be differentially expressed in most of the patients with arthritis compared to healthy patients, e.i., if the interleukin-18 expression is statistical-significantly higher in one patient compared to the others, it could indicate that this patient is more prone to suffer from arthritis than the rest.

4.4.1. Biomarkers for arthritis in diabetic patients

The proteins interleukin-8 (CXCL8), tumour necrosis factor receptor superfamily member 1A (TNFRSF1A), osteopontin (SPP1), interleukin-10 (IL-10), interleukin-18 (IL-18), and have resulted to be potential biomolecular identifiers (alone or in combinations of 2) of diabetic patients suffering from arthritis.

Interleukin-18 (IL-18):

The interleukin-18 (IL-18) is a proinflammatory cytokine highly common in inflammatory diseases. It is produced by mononuclear cells in RA and high levels of this interleukin have been associated with disease severity. Along with IL-1β, IL-12, and IL-15; IL-18 stimulate the production of IFN-γ by activated synovial T cells and promote the development of helper T cells (Th)1 response. IL-18 also acts as a direct proinflammatory cytokine in RA by promoting macrophage-driven TNF and IL-1 production. Also, it is

exacerbated via the NF-kB pathway which further supports its pivotal role in rheumatoid joint pathology. Moreover, IL-18 is also produced by FLS cells, which are the resident cells of the synovium and are responsible for the promotion of synovial hyperplasia and joint tissue destruction [74].

Autoantibodies against pro-inflammatory cytokines have been documented for TNF, IL-1 (alpha and beta), IL-2, IL-6, IL-8, IL-10, and IL-18, suggesting a potential role of these cytokines as being good biomarkers due to their easy detection in peripheral blood and its association with disease severity [82]. In addition, IL-18 is secreted by monocytes, whose peripheral blood elevates have also been suggested of acting as biomarkers of RA disease activity [74].

IL-18 is also overexpressed in diabetic patients. Its function has been reported to mediate insulin resistance [62].

Tumour necrosis factor receptor superfamily member 1A (TNFRSF1A):

Synovial fibroblastic cells in RA secrete large amounts of IL-6 when stimulated by inflammatory cytokines such as IL-1, TNF-α and IL-17. Fibroblast hyperplasia occurs due to high levels of soluble FasLigand (FasL) and TNF in the synovium; and members of the TNF receptor family, specifically TNF receptor 1 (TNFR1, or TNFRSF1A), TNF-related apoptosis-inducing ligand (TRAIL) receptors 1 and 2, and Fas have been seen to induce fibroblast apoptosis [83].

Recently, some studies revealed that expression of serum TNF-α may intensify the inflammatory activity in early RA, which indicates a strong correlation between cytokine expression and disease severity; suggesting that serum TNF-α could act as a competent biomarker for evaluation of disease activity in early RA [84, 85]. Nevertheless, no study has ever studied or obtained significant results regarding the expression of TNF-α receptors as a good biomarker.

Interleukin-10 (IL-10):

IL-10 expression in patients with arthritis has been shown to regulate endogenous proinflammatory cytokine production in synovial tissue and has been found in reduced levels, being unable to block T-cell responses to specific antigens [79]. IL-10 was analysed to be a potential biomarker of RA in Dissanayake et al. 2021, along with IL-1, TNF-α and IL-17α. Secretion of IL-10 by PBMCs negatively correlated with radiological progression of RA and swollen joint count, indicating a potential protective role of IL-10 secretion against joint swelling in RA. This could point towards a potential role of IL-10 as a biomarker but contradictory evidences might underrate the potential of this anti-inflammatory cytokine [85].

On the same line, accumulating evidence has shown that reduced serum levels of IL-10 are associated with a greater incidence of insulin resistance, which is a prominent pathophysiological process in diabetic patients. The anti-inflammatory cytokine works through the inhibition of NFκB activation mediated by TNF-α, which reduces IKK activity. Particularly, IL-10 might present insulin-sensitizing effects and, when used as treatment, it prevented insulin resistance by stopping the autophosphorylation of insulin receptors and downstream signalling mediators in the liver [86, 87]. Moreover, circulating IL-10 levels negatively correlated with the development of metabolic risk factors, indicating that IL-10 could act as a biomarker of metabolic disorders, such as diabetes [88].

Interleukin-8 (CXCL8):

The chemokine interleukin-8 (IL-8 or CXCL8) is frequently associated with inflammatory diseases, and autoantibodies against IL-8 are present in the periphery at elevated levels in RA. Indeed, high expression levels of IL-8 have been found in these patients.

During the synovial inflammation in RA, IL-8 shows an important role. This TNF-activated cytokine is produced after the activation of mesenchymal cells, recruitment of innate and adaptive immune system cells and activation of synoviocytes; leading to inflamed synovium, an increase in angiogenesis and a decrease in lymphangiogenesis.

Higher levels of cytokines that are involved in the pathogenesis of RA have been reported in patients with RA than in healthy controls and their serum levels were positively associated with disease severity [89]. Autoantibodies against pro-inflammatory cytokines have been documented for TNF, IL-1 (alpha and beta), IL-2, IL-6, IL-8, IL-10, and IL-18 [82], suggesting a potential role of these cytokines in being good biomarkers. Moreover, the levels of IL-8 have not only been significantly associated with disease severity but also with the presence of anti-citrullinated protein antibodies (ACPA), a common biomarker for RA. These findings suggest that relatively “downstream” signalling mediators such as chemokines when measured in the peripheral blood, may assist with predicting clinical outcomes for patients with RA [90].

In the diabetes context, adipose tissue produces a variety of proinflammatory cytokines, including IL-1, IL-6, IL-8 and IL-18. The first two have been extensively studied in many experiments in comparison with IL-8 and IL-18. However, the levels of these two cytokines have been suggested to be involved in metabolic disorders associated with T2DM [62].

Osteopontin (SPP1):

SPP1 is a pro-inflammatory cytokine secreted by macrophages with a critical role in immune cell recruitment, adhesion, and migration. In the pathophysiology of RA, it acts as an important mediator in the amplification and perpetuation of the disease, not only by mediating the attachment of synovial fibroblasts to cartilage, but also contributing to matrix degradation by stimulating the secretion of collagenase 1 in articular chondrocytes. In addition, SPP1 selectively induces the expression of pro-inflammatory cytokines and chemokines like IL-1 and IL-8 (another biomarker), maybe through the activation of the transcription factor NF-κB, leading to migration and the recruitment of inflammatory cells [91]. SPP1 has not been reported to be a capable biomarker of arthritis yet; however, it does have been used as a biomarker for the prediction of the remission of certain drugs [92].

In diabetes, the expression of SPP1 is elevated in obese adipose tissue and induces insulin resistance [93].

5. Discussion

Although apparently, biomarkers for diagnosis of RA are yet well-established, they do not fulfil the necessities of sufficient accuracy, leading to the detection of false positive and negative results. The most common autoantibodies (RF and ACPA) used as biomarkers were included in the new diagnostic EULAR 2010 criteria. ACPA specificity is notably high but they both lack sensitivity (<50%), especially in the early stages. In addition, since clinical distinctions between healthy and early disease states are more subtle and difficult to detect than in advanced stages; arthritis diagnosis becomes tardy, unsatisfactory and subjective based on the heterogeneous physicians' criteria of classification. In order to battle these obstacles, patients need a competent biomarker which accomplishes an accurate, objective and fast diagnosis of the early stages of the disease, characterised by both high specificity and sensitivity [82].

Here is where the results of the data mining process performed in this study take place. Most of the data mining-obtained arthritis classifiers that present a pronounced accuracy (>80%) are cytokines. Much is known regarding the involvement of cytokines in the pathophysiology of RA (particularly, in synovial inflammation); however; surprisingly few studies have investigated the reliability and validity of cytokines as predictive biomarkers in the autoinflammatory condition.

This study shows that cytokines are, in fact, promising candidate biomarkers of arthritis, presenting high sensitivity and specificity. Besides this, other meaningful advantages of using cytokines as biomarkers exist. One is the simple and accessible obtention of these predictive molecules from patients. The synovial inflammation occurs locally in the synovium but releases large amounts of cytokines into the peripheral blood circulation. Peripheral blood is the best tissue sample to detect biomarkers particularly for its good accessibility and for the lack of invasive procedures to obtain the biomarkers. The second one is that the current biomarkers commonly used, such as rheumatoid factor, can lead to non-specific binding in enzyme immunoassays; whereas the manufacturers of cytokines assays help block the interference of heterophilic antibodies, improving the accuracy of the results and permitting an effortless detection. Moreover, the clinical utility of most of the cytokines is already established in other diseases [22].

The sum of pieces of evidence such as the results from this gene expression analysis and promising studies on the availability of cytokines as biomarkers will complement the hypothesis that cytokines-based biomarkers are likely to emerge in the following years as potent predictors of disease activity and response to treatments [22, 82].

However, it is worth mentioning that, given the complexity and heterogeneous nature of RA pathogenesis, a single cytokine might not dispense sufficient discrimination to predict the outcome of interest. A more profitable approach would be to consider a combination of biomarkers, such as the combination of IL-8 and TNF receptors, to reach the most accurate prediction [22].

A notable variety of good classifiers were obtained according to the classification results, from which 5 different proteins resulted. Taking into account that a novel candidate for the classification of diabetic patients into arthritis or no arthritis should not coincide with a diabetes classifier and should not be previously reported as an arthritis biomarker, classifiers were contextualised and analysed in order to identify if they are novel

candidates. After a thorough review of the scientific literature, the cytokines were analysed if they would be potential biomarkers. TNF-alpha was recently reported as a good biomarker for RA [84]; however, no studies of its receptor have been developed yet. IL-8 and IL-18 were suggested in some studies to act as possible biomarkers although further investigation is needed to implement them in clinical usage [90]. On the other hand, IL-10 was also proposed but contradictory evidences diminished its reliability to act as a biomarker [88]. Finally, SPP1 was the only protein that was not previously reported in the literature in terms of biomarker discovery, indicating this protein could be a novel candidate biomarker for arthritis in diabetic patients.

Although SPP1 could be the most innovative biomarker, further investigations of the rest of the candidates should be conducted due to its excellent classification values and poor usage in clinical diagnosis.

6. Conclusions

Thanks to the application of artificial intelligence classification techniques to multiple-integrated datasets from the GEO database, a time-saving procedure was performed to identify useful and valid biomarkers for the classification of patients suffering from arthritis or not. Nevertheless, this was not the only objective. Besides the discovery of new arthritis biomarkers itself, this study was also useful to provide evidence that big data stored in public databases can be completely exploited for different and numerous purposes. Consequently, further considerations on the under usage of gene expression data stored in public repositories should be concluded in order to better exploit the sources of biological information that are currently available.

The main conclusion obtained is that HT data usage permits the study of other pathological conditions besides the disease labelled by the researchers in the GEO database. In the case of this work, several genes/proteins have been identified as potential biomarkers of arthritis in diabetic patients. The fact that some of the classifiers have been previously reported in the literature as good arthritis biomarkers validates and supports our findings regarding the resulting candidates that have not been examined yet, such as SPP1. We encourage clinical investigators to test the proposed cytokines in clinical practice. Additionally, these classifiers do not only own the capability of acting as biomarkers but also allow to predict diabetes and arthritis in all platforms submitted in the GEO database, even though the platform is labelled with other condition. Therefore, these classifiers could be applied to all samples and experiments in GEO, making possible the classification of all patients into arthritis and diabetes.

Another advantage of the resulting classifiers is that they permit knowing the complete biological profile of a patient based on gene expression data. The labels imposed by a physician, on the contrary, sometimes lead to biased or even false results due to a lack poor information level. For instance, the physician is missing valuable information if they cannot detect that the patient suffers from certain comorbidity, which could be causing the disease under study. Hence, the strategy here proposed can help researchers and physicians with a more accurate classification in this sense.

Inevitably, the employment of labelled data will always be better compared with the usage of classification predictors. Still, sample size can become a dominant limitation which will impede the potential growth of these methods.

In addition, the predicted labelling is based on a prediction and, thereby, is subjected to an error. The use of predicted labels to make biological conclusions must be carried out with a bigger sample size than the one used in an analysis with labels made by the investigators.

The inclusion of patients classified through these prediction-based classification techniques in the analysis of a researcher should consider a prediction of the complete clinical profile of the patient. Therefore, the researcher could include the samples that really fit the profile under study.

Finally, the cross-normalization used to scale multiple datasets is a determinant process that affects the results and therefore, could cause bias. The investigator must be extremely confident that the normalization step reaches the demanded sturdiness, independently of the employed normalization algorithm.

7. Annexes

Annexe 1. Complete list of arthritis classifiers.

Feature	TP	TN	FP	FN	Balanced ACC	p-value	PRE	SNS	SPC	Base classifier
IL18	203	300	0	97	0.84	2.53438E-85	100	68	100	LR+THR
IL18	203	300	0	97	0.84	2.53438E-85	100	68	100	LR+THR
IL18	203	300	0	97	0.84	2.53438E-85	100	68	100	LR+THR
IL18	203	300	0	97	0.84	2.53438E-85	100	68	100	LR+THR
IL18	203	300	0	97	0.84	2.53438E-85	100	68	100	LR+THR
TLR2	201	295	5	99	0.83	2.93431E-75	98	67	98	LR+THR
TLR2	201	295	5	99	0.83	2.93431E-75	98	67	98	LR+THR
TLR2	201	295	5	99	0.83	2.93431E-75	98	67	98	LR+THR
TLR2	201	295	5	99	0.83	2.93431E-75	98	67	98	LR+THR
DKK1	198	286	14	102	0.81	1.12031E-62	93	66	95	LR+THR
ENO1	202	298	2	98	0.83	1.2138E-80	99	67	99	LR+THR
MMP3	200	282	18	100	0.80	7.94696E-60	92	67	94	LR+THR
FGF2	201	280	20	99	0.80	1.59263E-58	91	67	93	LR+THR
VCAM1	201	284	16	99	0.81	2.63343E-62	93	67	95	LR+THR
F2RL1	294	194	106	6	0.81	7.80386E-70	74	98	65	LR+THR
GRB2	195	288	12	105	0.81	3.74164E-63	94	65	96	LR+THR
PTPN11	201	295	5	99	0.83	2.93431E-75	98	67	98	LR+THR
PTPN11	201	295	5	99	0.83	2.93431E-75	98	67	98	LR+THR
PTPN11	201	295	5	99	0.83	2.93431E-75	98	67	98	LR+THR
PTPN11	201	295	5	99	0.83	2.93431E-75	98	67	98	LR+THR
PTPN11	201	295	5	99	0.83	2.93431E-75	98	67	98	LR+THR
PTPN11	201	295	5	99	0.83	2.93431E-75	98	67	98	LR+THR
ENPP2	248	240	60	52	0.81	6.38152E-57	81	83	80	LR+THR
TNFRSF1A, FADD	264	272	28	36	0.89	8.58955E-94	90	88	91	MLP
F2RL1, GRB2	269	292	8	31	0.94	9.662E-122	97	90	97	MLP
IL17RA, GRB2	269	289	11	31	0.93	1.5663E-117	96	90	96	MLP
PTGES2, GRB2	262	279	21	38	0.90	4.3761E-99	93	87	93	MLP
CXCL8	259	229	71	41	0.81	1.03523E-57	78	86	76	OQ-THR
CXCL8	259	229	71	41	0.81	1.03523E-57	78	86	76	OQ-THR
CXCL8	259	229	71	41	0.81	1.03523E-57	78	86	76	OQ-THR
SPP1	277	206	94	23	0.81	1.34112E-58	75	92	69	OQ-THR
CXCL8	259	229	71	41	0.81	1.03523E-57	78	86	76	OQ-THR
TNFRSF10A	271	246	54	29	0.86	1.82279E-78	83	90	82	OQ-THR
TNFRSF10A	271	246	54	29	0.86	1.82279E-78	83	90	82	OQ-THR
TNFRSF10A	271	246	54	29	0.86	1.82279E-78	83	90	82	OQ-THR
TNFRSF10A	271	246	54	29	0.86	1.82279E-78	83	90	82	OQ-THR
TNFRSF10A	271	246	54	29	0.86	1.82279E-78	83	90	82	OQ-THR

Feature	TP	TN	FP	FN	Balanced ACC	p-value	PRE	SNS	SPC	Base classifier
RPSA	203	297	3	97	0.83	1.5739E-79	99	68	99	OQ-THR
PADI4	259	240	60	41	0.83	1.43927E-64	81	86	80	OQ-THR
PADI4	259	240	60	41	0.83	1.43927E-64	81	86	80	OQ-THR
SPP1	277	206	94	23	0.81	1.34112E-58	75	92	69	O-THR
IL10	203	299	1	97	0.84	3.9131E-83	100	68	100	O-THR
IL18	203	300	0	97	0.84	2.53438E-85	100	68	100	O-THR
IL10	203	299	1	97	0.84	3.9131E-83	100	68	100	O-THR
TLR2	201	295	5	99	0.83	2.93431E-75	98	67	98	O-THR
TLR2	201	295	5	99	0.83	2.93431E-75	98	67	98	O-THR
TLR2	201	295	5	99	0.83	2.93431E-75	98	67	98	O-THR
RPSA	203	297	3	97	0.83	1.5739E-79	99	68	99	O-THR
RPSA	203	297	3	97	0.83	1.5739E-79	99	68	99	O-THR
RPSA	203	297	3	97	0.83	1.5739E-79	99	68	99	O-THR
IL12A	202	279	21	98	0.80	3.60129E-58	91	67	93	O-THR
AIF1	204	289	11	96	0.82	2.49067E-69	95	68	96	O-THR
AIF1	204	289	11	96	0.82	2.49067E-69	95	68	96	O-THR
AIF1	204	289	11	96	0.82	2.49067E-69	95	68	96	O-THR
AIF1	204	289	11	96	0.82	2.49067E-69	95	68	96	O-THR
AIF1	204	289	11	96	0.82	2.49067E-69	95	68	96	O-THR
AIF1	204	289	11	96	0.82	2.49067E-69	95	68	96	O-THR
ANGPT1	201	289	11	99	0.82	1.31046E-67	95	67	96	O-THR
ANGPT1	201	289	11	99	0.82	1.31046E-67	95	67	96	O-THR
ANGPT1	201	289	11	99	0.82	1.31046E-67	95	67	96	O-THR
ANGPT1	201	289	11	99	0.82	1.31046E-67	95	67	96	O-THR
ANGPT1	201	289	11	99	0.82	1.31046E-67	95	67	96	O-THR
IL18, IL10	217	297	3	83	0.86	2.71768E-88	99	72	99	SVM
IL18, TNFRSF1A	273	299	1	27	0.95	4.4289E-138	100	91	100	SVM
TNFRSF10A, NFKB1	286	274	26	14	0.93	1.0003E-118	92	95	91	SVM
TNFRSF10A, RELB	266	258	42	34	0.87	2.89628E-83	86	89	86	SVM
TP53, GRB2	256	264	36	44	0.87	5.69358E-80	88	85	88	SVM
ENO1, IL32	282	292	8	18	0.96	5.5095E-136	97	94	97	SVM
ENO1, GRB2	271	252	48	29	0.87	5.81003E-83	85	90	84	SVM
ENO1, RELA	265	263	37	35	0.88	1.51609E-86	88	88	88	SVM
CTSL, TNFRSF1A	287	262	38	13	0.92	2.6265E-108	88	96	87	SVM
VIM, ENO1	271	270	30	29	0.90	2.49534E-98	90	90	90	SVM
RPSA, ENO1	271	283	17	29	0.92	5.2766E-112	94	90	94	SVM
TNFRSF1A, IL32	271	298	2	29	0.95	7.2289E-134	99	90	99	SVM
TNFRSF1A, AIF1	279	283	17	21	0.94	1.794E-120	94	93	94	SVM
TNFRSF1A, GRB2	268	281	19	32	0.92	8.865E-107	93	89	94	SVM
CTSS, F2RL1	271	283	17	29	0.92	5.2766E-112	94	90	94	SVM

Feature	TP	TN	FP	FN	Balanced ACC	p-value	PRE	SNS	SPC	Base classifier
CTSS, RIPK1	267	284	16	33	0.92	2.7511E-109	94	89	95	SVM
PTGS2, GRB2	266	292	8	34	0.93	8.4605E-119	97	89	97	SVM
JAK3, F2RL1	288	226	74	12	0.86	8.53394E-82	80	96	75	SVM
AIF1, MMP9	205	294	6	95	0.83	3.27791E-76	97	68	98	SVM
AIF1, IRF3	206	297	3	94	0.84	2.39555E-81	99	69	99	SVM
F2RL1, FADD	268	260	40	32	0.88	1.15845E-86	87	89	87	SVM
GRB2, FADD	267	280	20	33	0.91	1.0123E-104	93	89	93	SVM
NFKB2, TNFRSF10A	261	254	46	39	0.86	5.68749E-76	85	87	85	SVM
RIPK1, FADD	260	248	52	40	0.85	8.8356E-71	83	87	83	SVM
IL18, ENO1	264	263	37	36	0.88	1.10559E-85	88	88	88	SVM
ANGPT1, ENO1	278	289	11	22	0.95	5.1922E-127	96	93	96	SVM
ANGPT1, CTSS	272	283	17	28	0.93	5.2098E-113	94	91	94	SVM
IL17A, MMP3	220	289	11	80	0.85	4.53824E-79	95	73	96	SVM
HIF1A, AIF1	248	265	35	52	0.86	8.94556E-75	88	83	88	SVM
IL17RB, FLG	222	275	25	78	0.83	4.33911E-66	90	74	92	SVM

8. Bibliography

1. Pujol, A., et al., *Unveiling the role of network and systems biology in drug discovery.* Trends Pharmacol Sci, 2010. **31**(3): p. 115-23.
2. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-27.
3. Piccolo, S.R., et al., *Multiplatform single-sample estimates of transcriptional activation.* Proc Natl Acad Sci U S A, 2013. **110**(44): p. 17778-83.
4. Le Cao, K.A., et al., *YuGene: a simple approach to scale gene expression data derived from different platforms for integrated analyses.* Genomics, 2014. **103**(4): p. 239-51.
5. Meng, Q., et al., *DBNorm: normalizing high-density oligonucleotide microarray data based on distributions.* BMC Bioinformatics, 2017. **18**(1): p. 527.
6. Junet, V., et al., *CuBlock: a cross-platform normalization method for gene-expression microarrays.* Bioinformatics, 2021.
7. Boyle, J.P., et al., *Projection of diabetes burden through 2050: impact of changing demography and disease prevalence in the U.S.* Diabetes Care, 2001. **24**(11): p. 1936-40.
8. Knowler, W.C., et al., *Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin.* N Engl J Med, 2002. **346**(6): p. 393-403.
9. Artasensi, A., et al., *Type 2 Diabetes Mellitus: A Review of Multi-Target Drugs.* Molecules, 2020. **25**(8).
10. Sun, X., W. Yu, and C. Hu, *Genetics of type 2 diabetes: insights into the pathogenesis and its clinical application.* Biomed Res Int, 2014. **2014**: p. 926713.
11. Kaul, N. and S. Ali, *Genes, Genetics, and Environment in Type 2 Diabetes: Implication in Personalized Medicine.* DNA Cell Biol, 2016. **35**(1): p. 1-12.
12. Izzo, A., et al., *A Narrative Review on Sarcopenia in Type 2 Diabetes Mellitus: Prevalence and Associated Factors.* Nutrients, 2021. **13**(1).
13. Tian, Z., et al., *The relationship between rheumatoid arthritis and diabetes mellitus: a systematic review and meta-analysis.* Cardiovasc Endocrinol Metab, 2021. **10**(2): p. 125-131.
14. Rudan, I., et al., *Prevalence of rheumatoid arthritis in low- and middle-income countries: A systematic review and analysis.* J Glob Health, 2015. **5**(1): p. 010409.
15. Donath, M.Y., et al., *Inflammation in obesity and diabetes: islet dysfunction and therapeutic opportunity.* Cell Metab, 2013. **17**(6): p. 860-872.
16. Blum, A. and M. Adawi, *Rheumatoid arthritis (RA) and cardiovascular disease.* Autoimmun Rev, 2019. **18**(7): p. 679-690.
17. Chatterjee, S., K. Khunti, and M.J. Davies, *Type 2 diabetes.* Lancet, 2017. **389**(10085): p. 2239-2251.
18. Piva, S.R., et al., *Links between osteoarthritis and diabetes: implications for management from a physical activity perspective.* Clin Geriatr Med, 2015. **31**(1): p. 67-87, viii.
19. Aghaei Zarch, S.M., et al., *Molecular biomarkers in diabetes mellitus (DM).* Med J Islam Repub Iran, 2020. **34**: p. 28.

20. van der Linden, M.P., et al., *Classification of rheumatoid arthritis: comparison of the 1987 American College of Rheumatology criteria and the 2010 American College of Rheumatology/European League Against Rheumatism criteria.* Arthritis Rheum, 2011. **63**(1): p. 37-42.

21. Shapiro, S.C., *Biomarkers in Rheumatoid Arthritis.* Cureus, 2021. **13**(5): p. e15063.

22. Burska, A., M. Boissinot, and F. Ponchel, *Cytokines as biomarkers in rheumatoid arthritis.* Mediators Inflamm, 2014. **2014**: p. 545493.

23. Staiger, H., et al., *Pathomechanisms of type 2 diabetes genes.* Endocr Rev, 2009. **30**(6): p. 557-85.

24. Wu, W.T., et al., *Data mining in clinical big data: the frequently used databases, steps, and methodological models.* Mil Med Res, 2021. **8**(1): p. 44.

25. Jorba, G., et al., *In-silico simulated prototype-patients using TPMS technology to study a potential adverse effect of sacubitril and valsartan.* PLoS One, 2020. **15**(2): p. e0228926.

26. Bertran, L., et al., *Identification of the Potential Molecular Mechanisms Linking RUNX1 Activity with Nonalcoholic Fatty Liver Disease, by Means of Systems Biology.* Biomedicines, 2022. **10**(6).

27. Farres, J., et al., *Identification of the most vulnerable populations in the psychosocial sphere: a cross-sectional study conducted in Catalonia during the strict lockdown imposed against the COVID-19 pandemic.* BMJ Open, 2021. **11**(11): p. e052140.

28. Carcereny, E., et al., *Head to head evaluation of second generation ALK inhibitors brigatinib and alectinib as first-line treatment for ALK+ NSCLC using an in silico systems biology-based approach.* Oncotarget, 2021. **12**(4): p. 316-332.

29. Clough, E. and T. Barrett, *The Gene Expression Omnibus Database.* Methods Mol Biol, 2016. **1418**: p. 93-110.

30. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.

31. Wang, Y., D.J. Miller, and R. Clarke, *Approaches to working in high-dimensional data spaces: gene expression microarrays.* Br J Cancer, 2008. **98**(6): p. 1023-8.

32. Fu, G.X.X.Z.P.C.Z.Z.Y.Q.S.D., *Feature Selection based on the Bhattacharyya Distance.* 2006.

33. Tibshirani, R., *Regression Shrinkage and Selection Via the Lasso.* 1996.

34. Zou, H., *Regularization and variable selection via the elastic net (vol B 67, pg 301, 2005).*

35. Kira K, R.L., *Feature selection problem: traditional methods and a new algorithm.* . Proceedings Tenth National Conference on Artificial Intelligence. 1992.

36. Pedregosa F, V.G., Gramfort A, Michel V, Thirion B, Grisel O, et al., *Scikit-learn: Machine learning in Python.* J Mach Learn Res. 2011;12: 2825–2830. .

37. Haury, A.C., P. Gestraud, and J.P. Vert, *The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures.* PLoS One, 2011. **6**(12): p. e28210.

38. Christin, C., et al., *A critical assessment of feature selection methods for biomarker discovery in clinical proteomics.* Mol Cell Proteomics, 2013. **12**(1): p. 263-76.

39. Anand, D., B. Pandey, and D.K. Pandey, *A Novel Hybrid Feature Selection Model for Classification of Neuromuscular Dystrophies Using Bhattacharyya Coefficient, Genetic Algorithm and Radial Basis Function Based Support Vector Machine.* Interdiscip Sci, 2018. **10**(2): p. 244-250.

40. TK., H., *Random decision forests. .* Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. 1995. pp. 278–282. doi:10.1109/ICDAR.1995.598994.

41. *C. Cortes and V. Vapnik, Support-Vector Networks, Machine Learning, 20(3):273-297, September 1995. [Vladimir,Vapnik] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.*

42. *Haykin, Simon (1998). Neural Networks: A Comprehensive Foundation (2 ed.). Prentice Hall. ISBN 0-13-273350-1.*

43. *Keinosuke Fukunaga, Introduction to statistical pattern recognition (2nd ed.) Academic Press Professional, Inc. San Diego, CA, USA 1990 ISBN:0-12-269851-7.*

44. *Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. 2 (12): 1137–1143. CiteSeerX: 10.1.1.48.529.*

45. Galicia-Garcia, U., et al., *Pathophysiology of Type 2 Diabetes Mellitus.* Int J Mol Sci, 2020. **21**(17).

46. Veronese, N., et al., *Type 2 diabetes mellitus and osteoarthritis.* Semin Arthritis Rheum, 2019. **49**(1): p. 9-19.

47. Lin, Y. and Z. Sun, *Current views on type 2 diabetes.* J Endocrinol, 2010. **204**(1): p. 1-11.

48. Kasuga, M., *Insulin resistance and pancreatic beta cell failure.* J Clin Invest, 2006. **116**(7): p. 1756-60.

49. Chang-Chen, K.J., R. Mullur, and E. Bernal-Mizrachi, *Beta-cell failure as a complication of diabetes.* Rev Endocr Metab Disord, 2008. **9**(4): p. 329-43.

50. Stoffers, D.A., *The development of beta-cell mass: recent progress and potential role of GLP-1.* Horm Metab Res, 2004. **36**(11-12): p. 811-21.

51. Cersosimo, E., et al., *Assessment of pancreatic beta-cell function: review of methods and clinical applications.* Curr Diabetes Rev, 2014. **10**(1): p. 2-42.

52. Glamoclija, U. and A. Jevric-Causevic, *Genetic polymorphisms in diabetes: influence on therapy with oral antidiabetics.* Acta Pharm, 2010. **60**(4): p. 387-406.

53. Wang, H., et al., *Dominant-negative suppression of HNF-1alpha function results in defective insulin gene transcription and impaired metabolism-secretion coupling in a pancreatic beta-cell line.* EMBO J, 1998. **17**(22): p. 6701-13.

54. Kitamura, T., *The role of FOXO1 in beta-cell failure and type 2 diabetes mellitus.* Nat Rev Endocrinol, 2013. **9**(10): p. 615-23.

55. Donath, M.Y., *Targeting inflammation in the treatment of type 2 diabetes.* Diabetes Obes Metab, 2013. **15 Suppl 3**: p. 193-6.

56. Banerjee, M. and M. Saxena, *Interleukin-1 (IL-1) family of cytokines: role in type 2 diabetes.* Clin Chim Acta, 2012. **413**(15-16): p. 1163-70.

57. Frojdo, S., H. Vidal, and L. Pirola, *Alterations of insulin signaling in type 2 diabetes: a review of the current evidence from humans.* Biochim Biophys Acta, 2009. **1792**(2): p. 83-92.

58. Besse-Patin, A. and J.L. Estall, *An Intimate Relationship between ROS and Insulin Signalling: Implications for Antioxidant Treatment of Fatty Liver Disease.* Int J Cell Biol, 2014. **2014**: p. 519153.

59. Abbas, A., P.J. Grant, and M.T. Kearney, *Role of IGF-1 in glucose regulation and cardiovascular disease.* Expert Rev Cardiovasc Ther, 2008. **6**(8): p. 1135-49.

60. Li, P., et al., *Interferon gamma (IFN-gamma) disrupts energy expenditure and metabolic homeostasis by suppressing SIRT1 transcription.* Nucleic Acids Res, 2012. **40**(4): p. 1609-20.

61. Zhang, J., et al., *The diabetes gene Hhex maintains delta-cell differentiation and islet function.* Genes Dev, 2014. **28**(8): p. 829-34.

62. Feve, B. and J.P. Bastard, *The role of interleukins in insulin resistance and type 2 diabetes mellitus.* Nat Rev Endocrinol, 2009. **5**(6): p. 305-11.

63. McInnes, I.B. and G. Schett, *Pathogenetic insights from the treatment of rheumatoid arthritis.* Lancet, 2017. **389**(10086): p. 2328-2337.

64. Smolen, J.S., D. Aletaha, and I.B. McInnes, *Rheumatoid arthritis.* Lancet, 2016. **388**(10055): p. 2023-2038.

65. Song, X. and Q. Lin, *Genomics, transcriptomics and proteomics to elucidate the pathogenesis of rheumatoid arthritis.* Rheumatol Int, 2017. **37**(8): p. 1257-1265.

66. Deane, K.D., et al., *Genetic and environmental risk factors for rheumatoid arthritis.* Best Pract Res Clin Rheumatol, 2017. **31**(1): p. 3-18.

67. Chen, X.M., et al., *Role of Micro RNAs in the Pathogenesis of Rheumatoid Arthritis: Novel Perspectives Based on Review of the Literature.* Medicine (Baltimore), 2015. **94**(31): p. e1326.

68. Guo, Q., et al., *Rheumatoid arthritis: pathological mechanisms and modern pharmacologic therapies.* Bone Res, 2018. **6**: p. 15.

69. Koushik, S., et al., *PAD4: pathophysiology, current therapeutics and future perspective in rheumatoid arthritis.* Expert Opin Ther Targets, 2017. **21**(4): p. 433-447.

70. Schutyser, E., S. Struyf, and J. Van Damme, *The CC chemokine CCL20 and its receptor CCR6.* Cytokine Growth Factor Rev, 2003. **14**(5): p. 409-26.

71. Withrow, J., et al., *Extracellular vesicles in the pathogenesis of rheumatoid arthritis and osteoarthritis.* Arthritis Res Ther, 2016. **18**(1): p. 286.

72. Sarmiento Salinas, F.L., et al., *NF-kappaB1/IKKepsilon Gene Expression and Clinical Activity in Patients With Rheumatoid Arthritis.* Lab Med, 2017. **49**(1): p. 11-17.

73. Cuda, C.M., R.M. Pope, and H. Perlman, *The inflammatory role of phagocyte apoptotic pathways in rheumatic diseases.* Nat Rev Rheumatol, 2016. **12**(9): p. 543-58.

74. Rana, A.K., et al., *Monocytes in rheumatoid arthritis: Circulating precursors of macrophages and osteoclasts and, their heterogeneity and plasticity role in RA pathogenesis.* Int Immunopharmacol, 2018. **65**: p. 348-359.

75. Itoh, Y., *Metalloproteinases in Rheumatoid Arthritis: Potential Therapeutic Targets to Improve Current Therapies.* Prog Mol Biol Transl Sci, 2017. **148**: p. 327-338.

76. Hashizume, M., et al., *Tocilizumab, a humanized anti-IL-6R antibody, as an emerging therapeutic option for rheumatoid arthritis: molecular and cellular mechanistic insights.* Int Rev Immunol, 2015. **34**(3): p. 265-79.

77. Weitoft, T., et al., *Cathepsin S and cathepsin L in serum and synovial fluid in rheumatoid arthritis with and without autoantibodies.* Rheumatology (Oxford), 2015. **54**(10): p. 1923-8.

78. Yamashita, T., et al., *Effect of a cathepsin K inhibitor on arthritis and bone mineral density in ovariectomized rats with collagen-induced arthritis.* Bone Rep, 2018. **9**: p. 1-10.

79. Panagopoulos, P.K. and G.I. Lambrou, *Bone erosions in rheumatoid arthritis: recent developments in pathogenesis and therapeutic implications.* J Musculoskelet Neuronal Interact, 2018. **18**(3): p. 304-319.

80. *C. M. Bishop, Neural Networks for Pattern Recognition. USA: Oxford University Press, Inc.*

81. *BIPM. Guides in metrology, Guide to the Expression of Uncertainty in Measurement (GUM).*

82. Peichl, P., et al., *Anti-IL-8 autoantibodies and complexes in rheumatoid arthritis: polyclonal activation in chronic synovial tissue inflammation.* Rheumatol Int, 1999. **18**(4): p. 141-5.

83. Turner, J.D. and A. Filer, *The role of the synovial fibroblast in rheumatoid arthritis pathogenesis.* Curr Opin Rheumatol, 2015. **27**(2): p. 175-82.

84. Inam Illahi, M., et al., *Serum Tumor Necrosis Factor-Alpha as a Competent Biomarker for Evaluation of Disease Activity in Early Rheumatoid Arthritis.* Cureus, 2021. **13**(5): p. e15314.

85. Dissanayake, K., et al., *Potential applicability of cytokines as biomarkers of disease activity in rheumatoid arthritis: Enzyme-linked immunosorbent spot assay-based evaluation of TNF-alpha, IL-1beta, IL-10 and IL-17A.* PLoS One, 2021. **16**(1): p. e0246111.

86. de Luca, C. and J.M. Olefsky, *Inflammation and insulin resistance.* FEBS Lett, 2008. **582**(1): p. 97-105.

87. Shoelson, S.E., L. Herrero, and A. Naaz, *Obesity, inflammation, and insulin resistance.* Gastroenterology, 2007. **132**(6): p. 2169-80.

88. Kulshrestha, H., et al., *Interleukin-10 as a novel biomarker of metabolic risk factors.* Diabetes Metab Syndr, 2018. **12**(4): p. 543-547.

89. Kaneko, Y. and T. Takeuchi, *Targeted antibody therapy and relevant novel biomarkers for precision medicine for rheumatoid arthritis.* Int Immunol, 2017. **29**(11): p. 511-517.

90. Hitchon, C.A., et al., *A distinct multicytokine profile is associated with anti-cyclical citrullinated peptide antibodies in patients with early untreated inflammatory arthritis.* J Rheumatol, 2004. **31**(12): p. 2336-46.

91. Zhang, F., et al., *Role of osteopontin in rheumatoid arthritis.* Rheumatol Int, 2015. **35**(4): p. 589-95.

92. Liu, L.N., et al., *Circulating Levels of Osteoprotegerin, Osteocalcin and Osteopontin in Patients with Rheumatoid Arthritis: A Systematic Review and Meta-Analysis.* Immunol Invest, 2019. **48**(2): p. 107-120.

93. Schenk, S., M. Saberi, and J.M. Olefsky, *Insulin sensitivity: modulation by nutrients and inflammation.* J Clin Invest, 2008. **118**(9): p. 2992-3002.