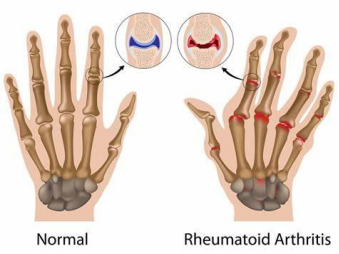
Shayla Ghose

Thursday July 10, 2025



1) What is the description of the disease explored? Where does it affect the body?

   -The disease is rheumatoid arthritis. Rheumatoid arthritis is an autoimmune disorder associated with inflamed joints and often accompanied by systemic symptoms. It causes pain, swelling, and stiffness in the joints. It occurs when the body's tissues, specifically the lining of the joints, are mistakenly attacked by its own immune system (autoimmune condition). It also eats away at the bone under the joints. This leads to inflammation and possibly joint damage, sometimes causing joints to be out of place. Currently, there is no cure, but it can be managed through medication. RA affects the joints in your feet, hands, wrists, knees, ankles, fingers, and toes, but can also damage other parts of the body including the skin, eyes, lungs, heart, and blood vessels. RA can increase the risk of heart disease and can cause symptoms of tiredness, fever, and not wanting to eat. Smaller joints are affected first, and the disease can continue to spread from there, most often on both sides of the body.

2) How many samples does this dataset have in total?

   -This dataset contains 275 samples.

3) What platform was used for this dataset?

   -The platform used in this dataset was the Affymetrix Human Genome U133 Plus 2.0 Array (microarray analysis), the experiment type was profiling by array.

4) What was the sample used?

   -The sample used was homo sapiens, comprising of both male and females. There were two groups, one healthy, and one diseased (diagnosed with RA).

5) How many samples were male, and how many were female?

-32 samples were male, and 243 samples were female. (could be equalized by using SMOTE)

6) How many females are healthy, and how many are diseased?

   -39 females were healthy, and 204 were diseased (RA).

7) How many males were healthy, and how many were diseased?

   -5 males were healthy, and 27 were diseased (RA).

8) Is the data already normalized?

   -The data was not normalized directly from the dataset, a log2 function in the code was used to normalize the various data ranges.

9) How was the data stored in the NCBI GEO database?

   -All the data was stored as one large data frame in the GEO database.

10) What is whole blood?

    -Whole blood is blood in its natural state, containing all components including red blood cells, white blood cells, platelets, and plasma.

11) What is the difference between whole blood and peripheral blood?

    -Whole blood is collected directly from a vein or artery without any processing or separation and is a complete mixture of all blood components. Peripheral blood refers to blood circulating throughout the circulatory system and can be used to isolate specific components.

SMOTE (Synthetic Minority Over-Sampling Technique):

Because imbalanced data sets impact the performance of machine learning models, SMOTE is used to address this imbalance problem by generating synthetic samples for the minority class.

Working Procedure of SMOTE:

- Identifies Minority Class – SMOTE operates on datasets where one or more groups are underrepresented.
- Nearest Neighbor Selection – SMOTE randomly chooses one of its nearest neighbors.
- Synthetic Sample Generation – SMOTE generates synthetic samples along the line segment by joining minority class instance and the selected nearest neighbor.
- Create Balanced Dataset – After generating synthetic samples for the minority class, the resulting dataset becomes more balanced with an equitable distribution amongst classes.

Affymetrix Human Genome Platform (GPL570):

- Affymetrix is a brand of DNA microarray products.
- Complete coverage of the Human Genome U133 set including 6,500 additional genes for analysis of over 47,000 transcripts.
- The sequences from which the probe sets were derived were selected from GenBank, dbEST, and RefSeq.
- Sequence clusters created from UniGene database.
- In addition, there are also 9,921 probe sets representing approximately 6,500 new genes.