A REPORT

ON

DATA VISUALIZATION & MACHINE LEARNING IN AYURGENOMICS

UNDERTAKEN AT

CSIR – INSTITUTE OF GENOMICS & INTEGRATIVE BIOLOGY

A PRACTICE SCHOOL-I STATION OF

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

JULY, 2019

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI (RAJASTHAN)

Practice School Division

Station(s): New Delhi	**Centre:** CSIR-IGIB, Mathura Road
Duration From: 21 May 2019	**To:** 14 July 2019

Date of submission: 14 July 2019

Title of the Project: To develop machine learning algorithm for visualizing heterogeneous multidimensional phenomics and genomics data

Name of Student	BITS ID	Discipline
Rohit Jain	2017A7PS0122P	Computer Science
Nishchit Soni	2017B3A71035P	Economics + Computer Science
Ishita Mediratta	2017A7PS1013G	Computer Science
Anmol Agarwal	2017B3A70489G	Economics + Computer Science
Kartik Bhatia	2017A7PS0051G	Computer Science
Syed Ahsan Abbas	2017B3A70507P	Economics + Computer Science

Name(s) of the expert	Designation
Rintu Kutum	PHD Scholar
Dr Mitali Mukerji	Senior Principal Scientist
Dr Bhavana Prasher	Principal Scientist

Name of the PS Faculty: Dr Deepak Chitkara

Key Words: Data Visualization, Machine Learning, Ayurgenomics, Genomics, Prakriti, Phenotypes, Divergence, Quantiles, Entropy

Project Area(s): Data Visualization, Machine Learning, Ayurgenomics

ACKNOWLEDGEMENT

We wish to express our gratitude towards Mr Rintu Kutum, Teaching Assistant of our project group, for his enthusiastic support, cooperation and help. We wish to thank him for facilitating this programme which gave us an opportunity to have academic exposure. We sincerely thank the staff of IGIB, especially Dr Mitali, Sr. Principal Scientist (Genomics and Molecular Medicine – Ayurgenomics) and Dr Bhavana, Principal Scientist (Ayurgenomics: Translational Research on Ayurveda) for their time, support, guidance and encouragement. We would also like to thank Dr Anurag Agarwal, Director, IGIB and Dr Jyoti, Head PME, IGIB for providing us an opportunity to work and learn amongst the best scientists of the country.

We express our gratitude towards Dr Deepak Chitkara, our PS-I faculty for his guidance and support throughout the training here. We also wish to thank the Practice School Division of BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI for giving us the opportunity to train and learn at CSIR-IGIB.

ABSTRACT

In the Ayurveda system of medicine, individuals are classified into seven constitution types, "Prakriti", for assessing disease susceptibility and drug responsiveness, which are assigned to an individual based on a set of phenotypes. Phenotypes can be categorised into 4 main types, 1) Anatomy 2) Physical 3) Physiological and 4) Psychological. In total, more than 100 phenotypes are required to assess *Prakriti types* of *an* individual by an Ayurvedic physician. Existing visual representation exist such as **somatotype**, which represents persons' physical and psychological traits based on quantitative measures. But *Prakriti* phenotyping is through only qualitative measurements. Various machine learning algorithms exist such as PCA, MDS plot, tSNE, etc, for the visual representation of large scale datasets. Here, we aim to develop machine learning algorithms for visualizing this heterogeneous multi-dimensional phenomics and genomics data. Firstly, we used qualitative phenotypes to develop and design algorithms for visual representation at individual-level. Secondly, we intend to use existing algorithms and design algorithms to find an association between quantitative phenotype with molecular phenotypes such as gene expression profiles.

OBJECTIVES

- To develop ML algorithm to visualize individual-level signature based on multiple phenotypes
    - Develop algorithm to capture phenotype to phenotype relationship
    - Develop algorithm to capture within Prakriti signatures
    - Develop algorithm to capture between Prakriti signatures
    - Design visual representation of individual-level signature based on above algorithms
- To develop ML algorithm to associate phenotype with gene expression
    - Use existing methodologies to find above associations
    - Develop novel ML algorithms to find multi-phenotype association with molecular cues.

KNOWING THE ORGANIZATION

IGIB

INSTITUTE OF GENOMICS
& INTEGRATIVE BIOLOGY

Genomics Knowledge Partner

CSIR-Institute of Genomics & Integrative Biology (IGIB) is a premier Institute of Council of Scientific and Industrial Research (CSIR), engaged in research of national importance in the areas of genomics, molecular medicine, bioinformatics and proteomics.

MISSION-

"To translate concepts developed in basic biological research to commercially viable technologies for health care"

IGIB was established in 1977 as the Centre for Biochemical Technology (CBT). The Functional Genomics Unit was established in 1998 with the focus shifting from chemical to genomics research. The institute was renamed "Institute of Genomics and Integrative Biology" in 2002.

RESEARCH AREAS

I. Genomics and Molecular Medicine

Genomics and Molecular Medicine is the major research focus of IGIB. From large collaborative projects like the Indian Genome Variation Consortium project to exploring the genetics of complex disorders using a candidate gene approach several groups at IGIB are involved in studying the molecular basis of human diseases.

They focus on:

* Neuropsychiatric disorders like Schizophrenia
* Diabetes and other complex disorders.

II.Cardio-Respiratory Disease Biology

A significant number of IGIB scientists focus on respiratory diseases using clinical, genetic, molecular and drug development approaches to tackle this challenging area.

The diseases of interest here are:

* Tuberculosis
* Asthma and Allergy
* Chronic Obstructive Pulmonary Disorder (COPD)

III. Chemical and Systems Biology

Chemical approaches are essential in the understanding of many biological phenomena. Several research groups at IGIB have come together to utilize their varied expertise in different disciplines of chemistry and biology to address contemporary research problems that require interdisciplinary cross-talk. Research carried out at IGIB in this area involves:

* Chemical biology and systems biology of M. tuberculosis and skin pigmentation
* chemically modified oligonucleotides for biological applications
* Nano biotechnology
* Novel immunoassay procedures
* New molecules

IV. Informatics and Big Data

IGIB has over the years built up expertise in high-throughput data analysis and genome annotation. It is participating in international efforts like the Gen2Phen Consortium for unifying genetic variation databases. Genome Informatics also forms an integral part of most other research areas at IGIB and contributes to the development of tools and hypotheses. The areas where informatics has contributed to genome analysis include:

* Indian Genome Variation: analysis of genome variation data
* Next-gen sequencing, assembly and annotation
* Unfolded proteins and adhesins
* Prediction of microRNA-target interaction
* Structural regulatory motifs in the genome

V. Integrative and Functional Genomics

Functional & Integrative Genomics is devoted to large-scale studies of genomes and their functions, including systems analyses of biological processes.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

It is well recognized that genetic variations in genes, protein to gene interactions and gene to gene interactions can lead to system-wide level changes, which might confer differential response among individuals to a disease condition. Currently, various -omics studies suggest better management or diagnosis of diseases such as, cancer immunotherapy, through stratification of patients based on endophenotypes (mRNA, miRNA, protein profiles). Ayurveda is a traditional Indian system of medicine which provides a methodology to stratify healthy individuals into categories on the basis of phenotypic attributes. The individuals in each category can have specific susceptibility towards different diseases, and drug responsiveness as mentioned in Ayurveda. The underlying philosophy of Ayurveda emphasizes on the use of natural means to eliminate the root cause of the disease and maintain homeostasis with the environment. The system proposes the concept of 'Tridoshas' namely Vata, Pitta, and Kapha. Different proportions of these 'Tridoshas' yields into seven Prakriti types, namely Vata(V), Pitta(P), Kapha(K), VP, PK, VK and VPK; thus prescribing a distinctive treatment plan based on their unique constitution. Prakriti evaluation involves clinical examination including questions about anatomical, physiological, physical and psychological traits. This approach while providing an ideal framework for the stratification of the diseased and the healthy population also takes into account the individual variations thus improving the drug efficacy. The traits, when identified with the corresponding Prakriti type, can be used as the basis for the prescription of the medicine thus working towards the idea of Predictive, Preventive, Personalized and Participatory (P4) medicine.

1.1 ABOUT GENOMICS

Genomics or functional genomics aims to characterize the function of every genomic element of an organism by using genome-scale high throughput assays such as genome sequencing, transcriptome profiling, epigenomics, metabolomics, and proteomics (2). Genomics can give insights about plausible associations between genotype and phenotype (3), discovering biomarkers for patient stratification, predicting functions of genes etc.(4)

After the publication by James D. Watson and Francis Crick confirmed the structure of DNA in 1953 (6), nucleic acid sequencing became a major point of interest for early molecular biologists, leading to the discovery of "codons" in the DNA by Marshall Nirenberg and Har Gobind Khorana led research team in 1961 and the first nucleic acid sequence in 1964 by Robert Holley and his colleagues (7).

In 1977, Frederick Sanger developed a sequencing technique for DNA to sequence the first complete genome, called phiX174 virus, which opened the doorway to the possibility in the field of genomics.

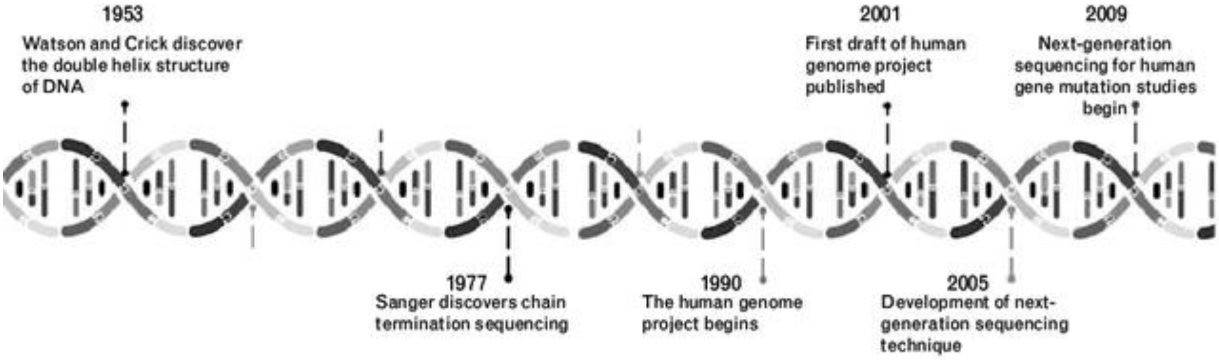

1953
Watson and Crick discover the double helix structure of DNA
2001
First draft of human genome project published
2009
Next-generation sequencing for human gene mutation studies begin
1977
Sanger discovers chain termination sequencing
1990
The human genome project begins
2005
Development of next-generation sequencing technique

Fig 1: Genesis of Genomics

The Human Genome Project was launched in 1990 with the aim to sequence all 3 billion letters of the human genome. Chromosome 22 was the first chromosome to be sequenced as a part of this project in 1999. The project was completed in 2003 and confirmed that humans have 20,000-25,000 genes (8). In 2007, there was a breakthrough in the technology used to sequence DNA, which led to a 70-fold increase in the output of DNA sequencing in one year. This led to the launch of the 1000 Genes Project in 2008, which aimed to sequence the genomes of a large population group of 2500.

1.2. INDIAN GENOME VARIATION PROJECT [25]

The Indian Genome variation project was initiated in 2003 by six laboratories of Council of Scientific and Industrial Research (CSIR) based on a network program which focussed primarily on repeats and single nucleotide polymorphism[25]. The objective was to collect 15000 individual from different subpopulation considering the ethnic diversity in the country and identify 1000 genes related to common diseases and drug responses and identify a minimum of five to ten informative markers per the Indian subpopulation gene. Being the first large scale comprehensive study of the structure of Indian population, the project aims to meet the ultimate goal of creating a DNA variation database of the Indian population to study the human biology with respect to disease predisposition and adverse drug reaction.

IGIB
INSTITUTE OF GENOMICS & INTEGRATIVE BIOLOGY
Genomics Knowledge Partner
CSIR
iitr
CCMB
CSIR
CDRI
CSIR-IMTECH

Fig 2: Institutes involved in Indian Genome Variation Project

1.3. 1000 GENOMES PROJECT

The project was launched in January 2008 with an international research effort to develop the most detailed catalogue of human genetic variation. Secondary goals of the project will include the support of better SNP and probe selection for genotyping platforms in future studies and the improvement of the human reference sequence. It has become a valuable tool for all fields of biological science, especially in the disciplines of genetics, medicine, pharmacology, biochemistry, and bioinformatics also leading to the need for novel and efficient data compression algorithms

1.4. MACHINE LEARNING IN GENOMICS

There has been an exponential increase in the number of multi-dimensional, highly-complex datasets available, that have been generated through years of research in the last 20 years. The immense amount of genomics data generated by genomic researchers provide a huge opportunity for the development of statistical machine learning algorithms to uncover patterns from it. These statistical machine learning methods can be used in identifying different types of genomic elements, recognizing patterns in DNA sequences, and developing models that can take other genetic and genomic information as input to build systems to help understand the biological mechanisms of underlying genes etc (9)

Machine learning is an emerging field in computer science wherein algorithms are developed and programmed to infer patterns and gain novel insights from the data, using a plethora of mathematical concepts. The learned model can then be used to predict any range of outputs, such as binary responses, categorical labels, or continuous values (13).

A few applications of statistical machine learning in the field of genomics is mentioned in **Table 01.**

Table 1: Selected applications of machine learning in genomics

Input data	**Task**
DNA sequence	Identify transcription start sites, splice sites, exons, etc
DNA sequence	Identify TF binding sets
DNA sequence	Identify genes
Gene Expression	Predict regulatory relationship
Gene Expression data	Identify biomarkers for a disease
Histone and TF ChIP-seq data	Partition and label the genome with chromatin state annotation
DNA sequence + gene expression +...	Predict gene function
DNA sequence + histone mods + ...	Predict gene expression
DNA sequence + histone mods + ...	Predict variant deleteriousness
Sequence variants + gene expression +...	Predict disease phenotype or prognosis

1.4.1. ENCODE PROJECT

ENCODE Project has applied machine learning approaches to enable integration and exploration of large and diverse data. Example - Prediction of TF binding sites (ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge - using ChIP-Seq data from the ENCODE project, to map sequences against these data in order to identify all TFBS sites for a giving TF)

1.4.2. DEEP GENOMICS

Companies like Deep Genomics, use machine learning to help researchers interpret genetic variation. Specifically, algorithms are designed based on patterns identified in large genetic data sets which are then translated to computer models to help clients interpret how genetic variation affects crucial cellular processes.

1.4.3. THE CANCER GENOME ATLAS (TCGA)

It was started in 2005 to catalogue genetic mutations responsible for cancer using high-throughput genome analysis techniques. Initially, the focus was on characterization of glioblastoma multiforme, lung, and ovarian cancer. In phase II, the project completed the genomic characterization and sequence analysis of 33 cancer types including 10 rare cancers. The focus shifted to creating a large, statistically significant data set for further discovery. Machine learning analysis of gene expression data can be used in identifying biomarkers associated with cancer. For example - **Predict cancer survival models.** *(Zhang X, Li Y, Akinyemiju T, et al. Pathway-Structured Predictive Model for Cancer Survival Prediction: A Two-Stage Approach).*

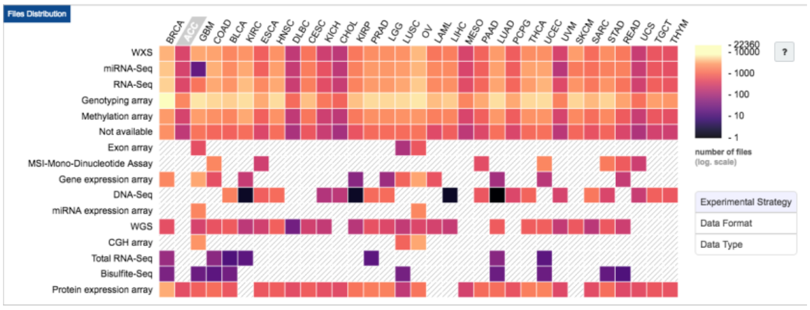


Fig. 3: Molecular Data in TCGA

2. AYURVEDA

The Ayurvedic system is a traditional Indian system of medicine which is nearly 3,500 years old and prescribes remedies for a range of problems. The Sanskrit classics Atharvaveda, Charak Samhita and Sushrut Samhita together describe more than 700 medicinal herbs, cataloguing everything from their taste, appearance and digestive effects to safety, efficacy, dosage and benefits.

The underlying philosophy of Ayurveda emphasizes on the use of natural means to eliminate the root cause of the disease and maintain homeostasis with the environment. The system proposes the concept of ‘Tridoshas’ namely Vata, Pitta, and Kapha. Different proportions of these ‘Tridoshas’ yields into seven Prakriti types, namely Vata(V), Pitta(P), Kapha(K), VP, PK, VK and VPK. This concept of Tridosha can be used as a common organizing principle in health and disease thus working towards the concept of personalized medicine.

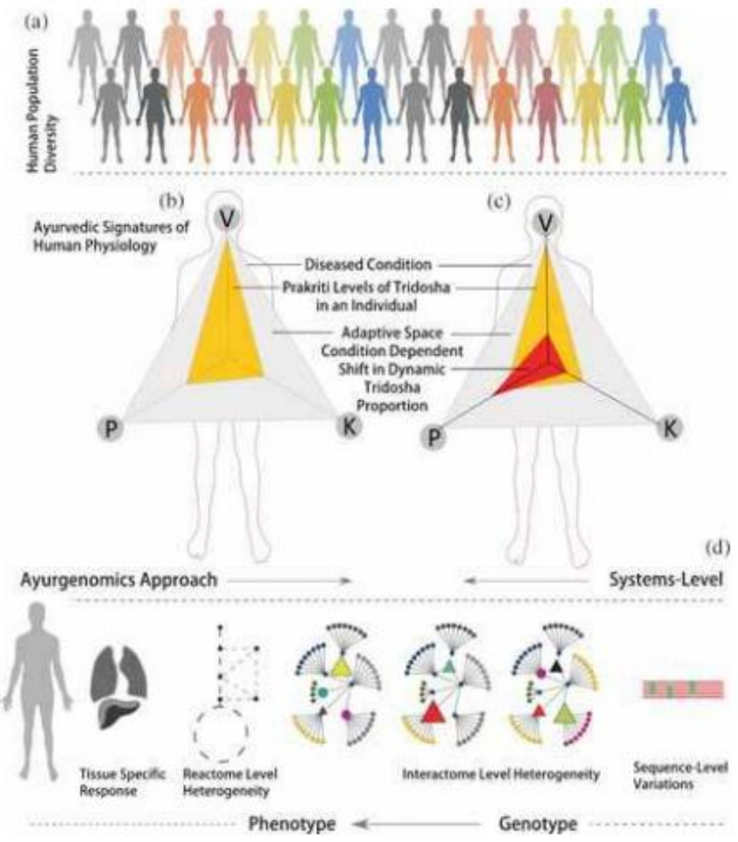

(a)
Human Population Diversity
(b)
Ayurvedic Signatures of Human Physiology
V
P
K
(c)
V
P
K
Diseased Condition
Prakriti Levels of Tridosha in an Individual
Adaptive Space
Condition Dependent Shift in Dynamic Tridosha Proportion
(d)
Ayurgenomics Approach
Systems-Level
Tissue Specific Response
Reactome Level Heterogeneity
Interactome Level Heterogeneity
Sequence-Level Variations
Phenotype
Genotype

Fig 4: Concept of Tridoshas in Ayurveda

Figure (a) represents system’s understanding of physiological networks in a population. Figures (b) and (c) represent the tridosha concept in human physiology. The three vertices of the triangle represent vata, pitta and kapha axes. Figure (b) represents the static state (prakriti) and figure (c) represents the dynamic (vaikarik) represented by yellow and red triangles respectively. Figure (d) depicts all levels of organization where prakriti tridosha proportions can bring about variability. Thus clearly illustrating the Tridoshas as an organizing principle

2.1. PRAKRITI- THE STATIC COMPONENT

Prakriti evaluation involves clinical examination including questions about anatomical, physiological, physical and psychological traits. This approach while providing an ideal framework for the stratification of the diseased and the healthy population also takes into account the individual variations thus improving the drug efficacy. The traits, when identified with the corresponding Prakriti type, can be used as the basis for the prescription of the medicine thus working towards the idea of Predictive, Preventive, Personalized and Participatory (P4) medicine.

Fig 5 on the next page illustrates how different traits can be used to identify the constitution type- Prakriti. For example, body weight change if is difficult to gain is a Vata property while if it changes easily it’s a Kapha property.

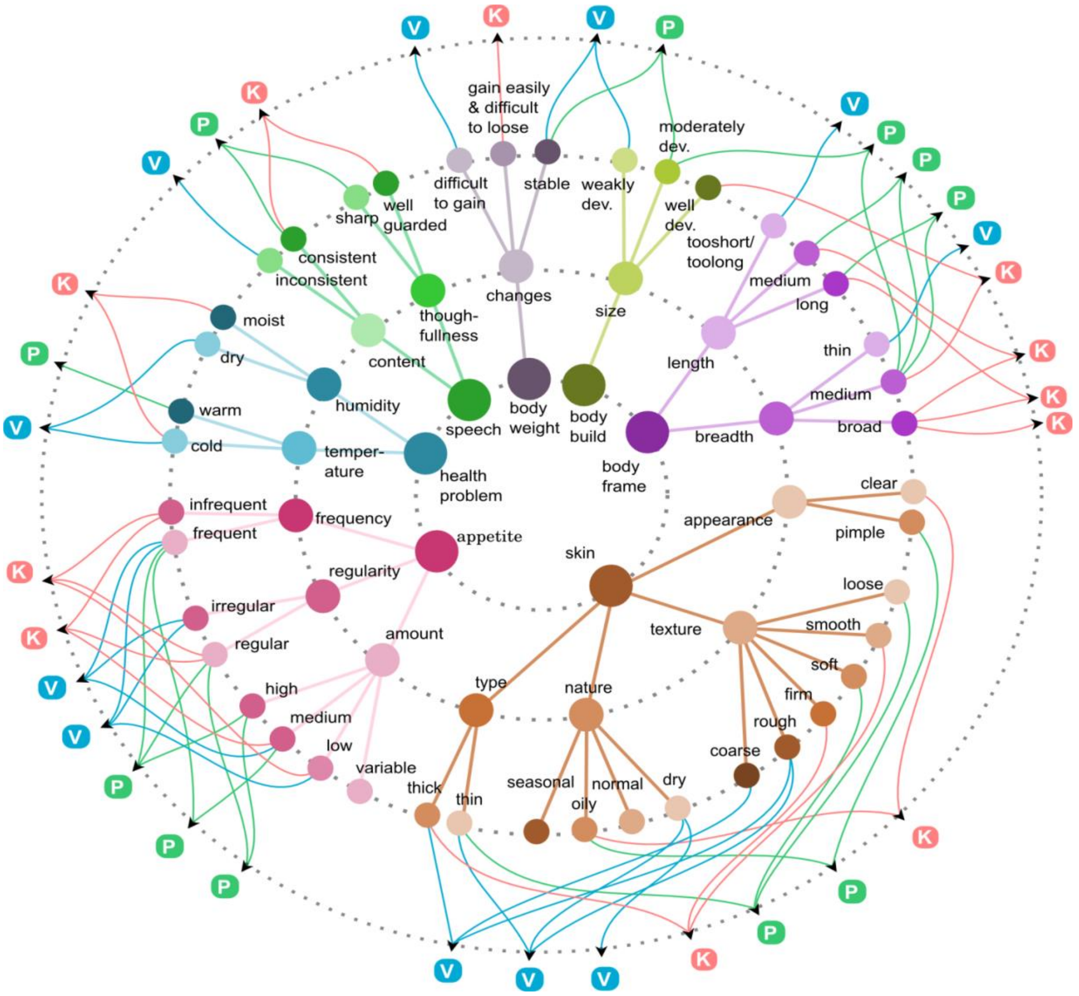


Fig 5: Map of Prakriti: The Static Component

2.2. VAIKARIK - THE DYNAMIC COMPONENT

The dynamic component can fluctuate along any of the three axes in response to intrinsic and extrinsic stimuli-

1. Age of the individual
2. Climate/ Season
3. Time of the day

Health state is depicted as the adaptive space is constrained within the limit depicted by the grey triangle beyond which, is the diseased state. Thus, an individual should maintain his lifestyle so as to balance the three axes into the health state. Disease in an individual onsets when the dynamic component crosses the threshold towards any of the vertices.

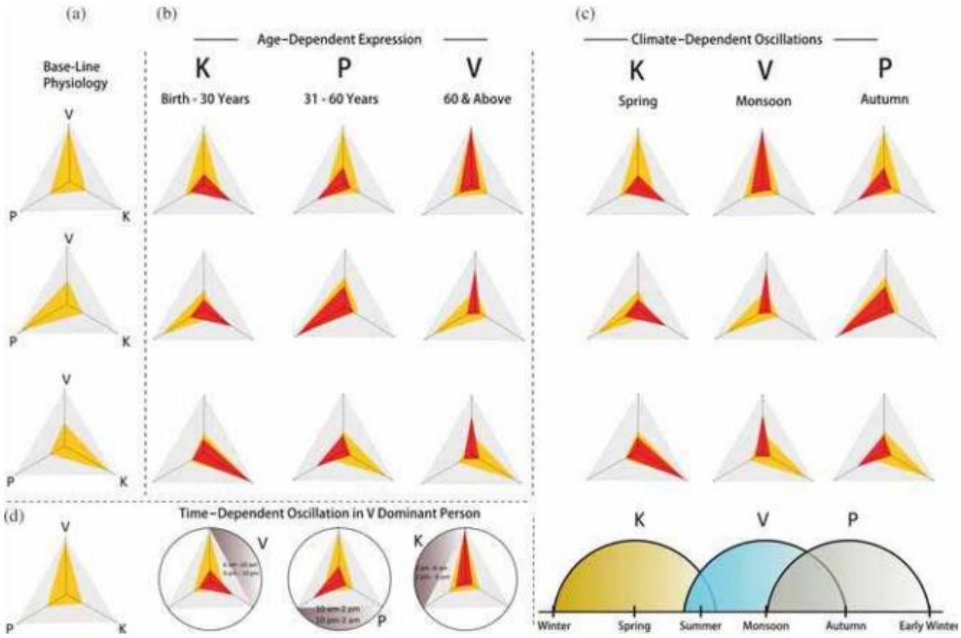


Fig 6: Vaikarik: The Dynamic Component of the Constitution

Figure 6 shows how different internal and external stimuli decide the health state. For example, it may be seen that K is dominant in first 30 years of an individual does making a K individual more prone to a disease. However, it changes in the later cycle of life with P more prone to diseases in the 31-60 years and V after 60 years clearly depicting the dynamic nature of a constitution.

3. AYURGENOMICS

Ayurgenomics is an emerging field that integrates high-throughput genomics experiments to investigate Ayurvedic stratification methodology (*Prakriti*)(18). Prasher B. *et.al.*, (24) showed molecular signatures and biochemical correlates among extreme Prakriti types (K, P & V) based on 96 healthy individuals. One of the genes identified in this study was *EGLN1(27)*, which was further studied in detail to understand the association of genotypes(SNP) with high-altitude adaptation (disease condition, High Altitude Pulmonary Edema).

Ayurgenomics is an emerging field that integrates high-throughput genomics experiments to investigate Ayurvedic stratification methodology i.e. the concept of Prakriti.

3.1. EGLN1 - CORRELATING WITH HIGH ALTITUDE ADAPTATION

EGLN1 is an oxygen-sensor gene that negatively regulates HIF protein. At high altitudes, hypoxic conditions inactivate the gene, thus increasing the amount of HIF, which leads to the adaptation. Therefore, Aggarwal et al postulated that the genetic variations because of differential expression correlating with Prakriti phenotypes could provide leads for understanding adaptation to external environment and susceptibility to diseases such as High Altitude Pulmonary Edema, and observed 5 of these differentially expressed genes to differ significantly, one of which was the EGLN1.

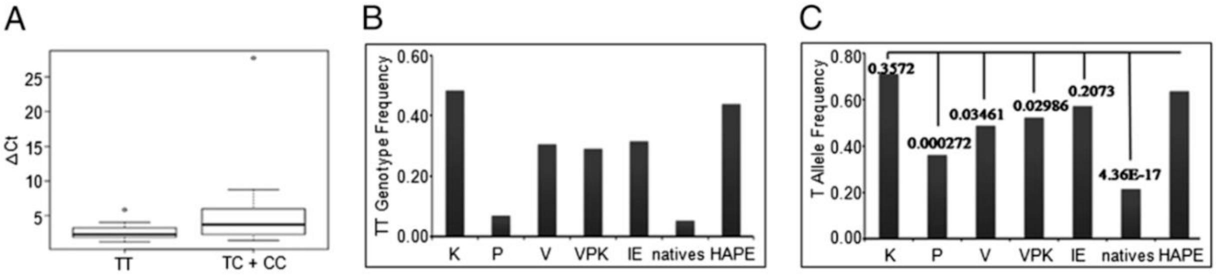


Fig. 7: Genotype frequencies w.r.t. Various subpopulations and cohorts

In the paper, allele frequencies of two common variants (rs479200 and rs480902) of the gene were studied. Box Plot A represent the TT, TC and CC genotypes of variant rs479200, where TT genotype is linked to more expression of the EGLN1 gene. Because higher expression of EGLN1 is inversely correlated to HIF activity, it was hypothesized that individuals with genotypes associated with high EGLN1 expression may not be able to perform well under hypoxic conditions. As shown in the Box

Plot B, individuals with Kapha prakriti have more frequency of the TT genotype, which is also higher in patients suffering from HAPE. In addition, the frequency of C allele of rs480902 and the T allele of rs479200 (0.63 and 0.64, respectively) in HAPE patients was similar to K type. The alleles associated with the K constitution were significantly underrepresented in P constitution (0.36 and 0.36) as well as the natives (0.28 and 0.21) of high altitude.

Therefore, TT genotype, corresponding to higher gene expression of EGLN1, is overrepresented in K and rare in natives and P, which raises the possibility that K may have a higher risk of HAPE and P Prakriti could be more protected. Thus, stratification of individuals based on their prakritis can provide some insightful gains, as the averaged out VPK population (without stratification) is nearly similar to the overall IE population, as shown in Box Plots B & C.

4. MACHINE LEARNING IN AYURGENOMICS

Tiwari P, Kutum R, Sethi TP *et.al.*[1], developed machine learning models to classify extreme Prakriti types based on 154 questionnaire data. They used unsupervised clustering approaches to uncover structure within the questionnaire data of 147 individuals. Based on supervised learning, minimum feature sets were identified in the classification of extreme Prakriti types. More than 85% mean accuracy was observed in LASSO, Elastic net and Random forest models, which were validated in another cohort (North cohort). Additionally, models were built to identify extreme Prakriti types from non-extreme Prakriti types.

Today, there has been an increase in emphasis on endophenotyping along with omics profiling for building a framework for stratification of individuals into groups in accordance with their disease susceptibility and drug responsiveness.

It has been well established that there are molecular differences between different 'Prakriti' types[1], the basic constitution of an individual as mentioned in Ayurveda. This approach can form a basis for the framework to stratify the healthy and diseased while taking into account the individual differences by analysing the Prakriti profiles.

4.1. THE STUDY

The study was conducted with two main objectives-

1. Clustering Analysis of the Data collected from the questionnaire to identify and hence classify the data into three clusters corresponding to each 'Tridosha'. Random forests algorithms were used for classification into clusters and MDS plot for visualizing the clusters.
2. Supervised models to train the model for 'Prakriti' classification model to use different machine learning algorithms to train the model for identification of extreme 'Prakritis' and use the same model for distinguishing between extreme and non-extreme 'Prakriti' types. Three models were used to ensure robustness:
   - Lasso Regression
   - Elastic Net Model
   - Random Forests Algorithms

Figure 8 below illustrates the whole study that was carried out by the authors of the paper.

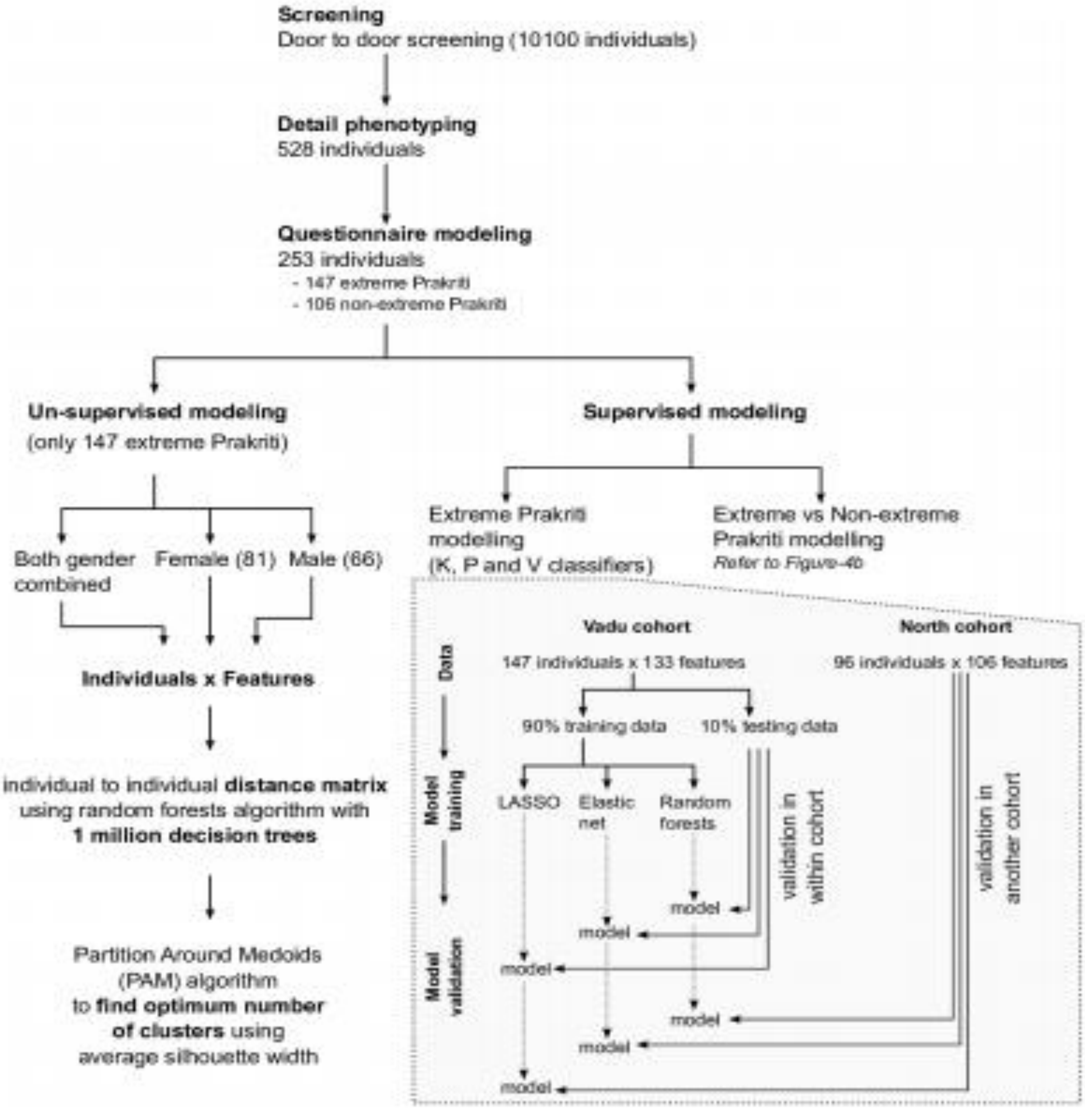

Screening
Door to door screening (10100 individuals)
Detail phenotyping
528 individuals
Questionnaire modeling
253 individuals
- 147 extreme Prakriti
- 106 non-extreme Prakriti
Un-supervised modeling
(only 147 extreme Prakriti)
Supervised modeling
Both gender combined
Female (81)
Male (66)
Individuals x Features
individual to individual distance matrix using random forests algorithm with 1 million decision trees
Partition Around Medoids (PAM) algorithm to find optimum number of clusters using average silhouette width
Extreme Prakriti modelling (K, P and V classifiers)
Extreme vs Non-extreme Prakriti modelling
Refer to Figure-4b
Data
Model training
Model validation
Vadu cohort
North cohort
147 individuals x 133 features
96 individuals x 106 features
90% training data
10% testing data
LASSO
Elastic net
Random forests
validation in within cohort
validation in another cohort
model
model
model
model
model
model

Fig. 8: Flow chart depicting all the steps employed in the manuscript

4.2. UNSUPERVISED CLUSTERING USING RANDOM FORESTS

Unsupervised algorithms such as Random Forests were used to identify clusters which led to the emergence of three natural clusters corresponding to three extreme Prakriti classes. Figure 9 on the next page shows the results for MDS plotting. Three clusters can easily be identified from the lot corresponding to each Tridosha.

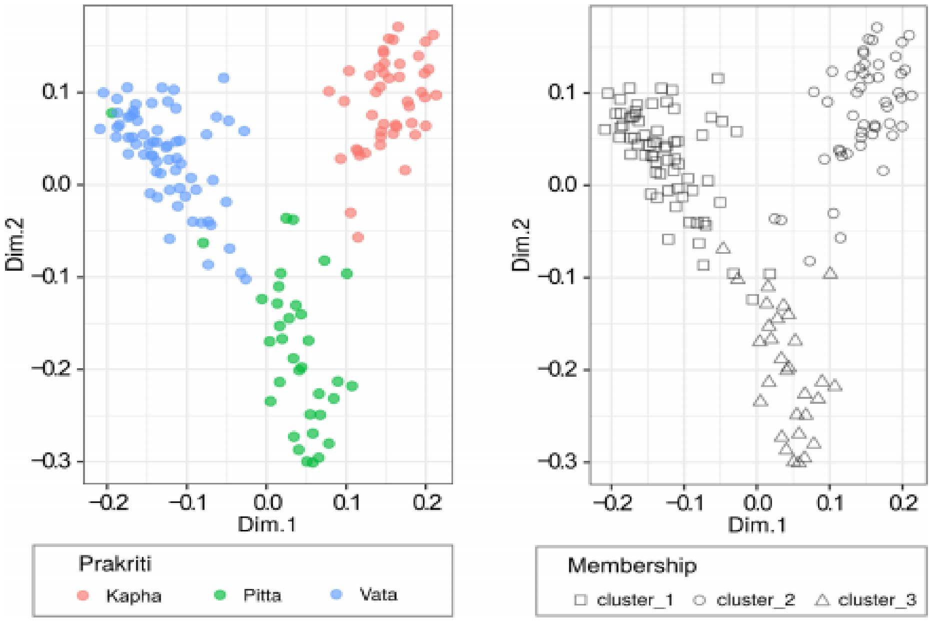


Fig. 9: MDS Plots for Unsupervised Clustering of Individuals

4.3. SUPERVISED LEARNING FOR PREDICTION OF PRAKRITI TYPE

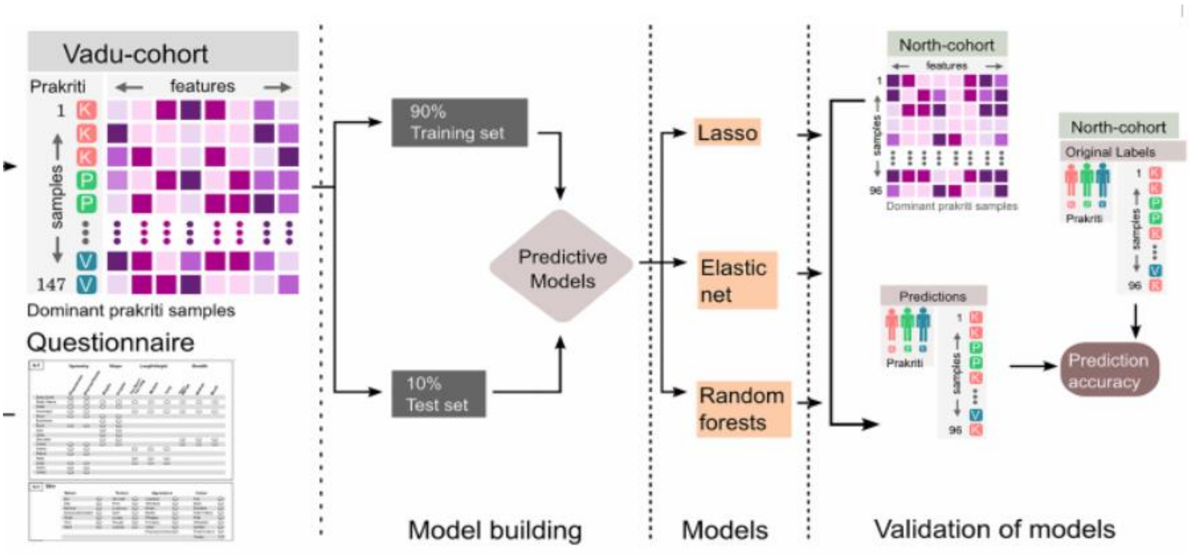


Fig. 10: Mind Map for Supervised Learning

The authors used three regression models and tested on 10% test data from the Vadu Cohort plus whole of the dataset from the North Indian cohort. They also tested it out in the opposite way with

90% training data and 10% testing from North Indian population and 100% testing data from Vadu Cohort The specificity from all three models were more than 90% .

4.3.1. THE PROBLEM OF MULTICOLLINEARITY

When predictor variables are correlated to each other and to the response variable it is called multicollinearity.

To picture this let’s say we’re doing a study that looks at a response variable—patient weight, and our predictor variables would be height, weight, and age. The problem here is that height and weight are also correlated and can inflate the standard error of their coefficients which may make them seem statistically insignificant.

To produce a more accurate model of complex data we can add a penalty term to the OLS equation. A penalty adds a bias towards certain values. These are known as L1 regularization (Lasso regression) and L2 regularization (ridge regression).

4.3.2. LASSO REGRESSION

The regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is useful in producing simpler models. However, when the data have highly correlated predictors, it tends to select only one variable and removes the rest of those.

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

4.3.3. ELASTIC NET MODEL

The model linearly combines the L1 and L2 penalties of the lasso and ridge methods. It is a result critique on lasso, whose variable selection can be too dependent on data and thus unstable. Effectively this will shrink some coefficients and set some to 0 for sparse selection thus, correlated variables are shrunk or removed at once. It is useful for Non-extreme Prakriti stratifying, as we need some correlated variables to be left.

$$\boxed{L_{enet}(\hat{\beta}) = \frac{\sum_{i-1}^{m}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda(\frac{1-\alpha}{2}\sum_{j-1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j-1}^{m}|\hat{\beta}_j|),}$$

where α is the mixing parameter between ridge (α = 0) and lasso (α = 1).

4.3.4. RANDOM FOREST ALGORITHM

Random forest model is nothing but ensemble of decision trees where 64% of the total data set is randomly selected as training set and the rest is used as test set. A decision tree consists of the root- the first node, intermediate nodes and the leaves that contain purity in the classified group.

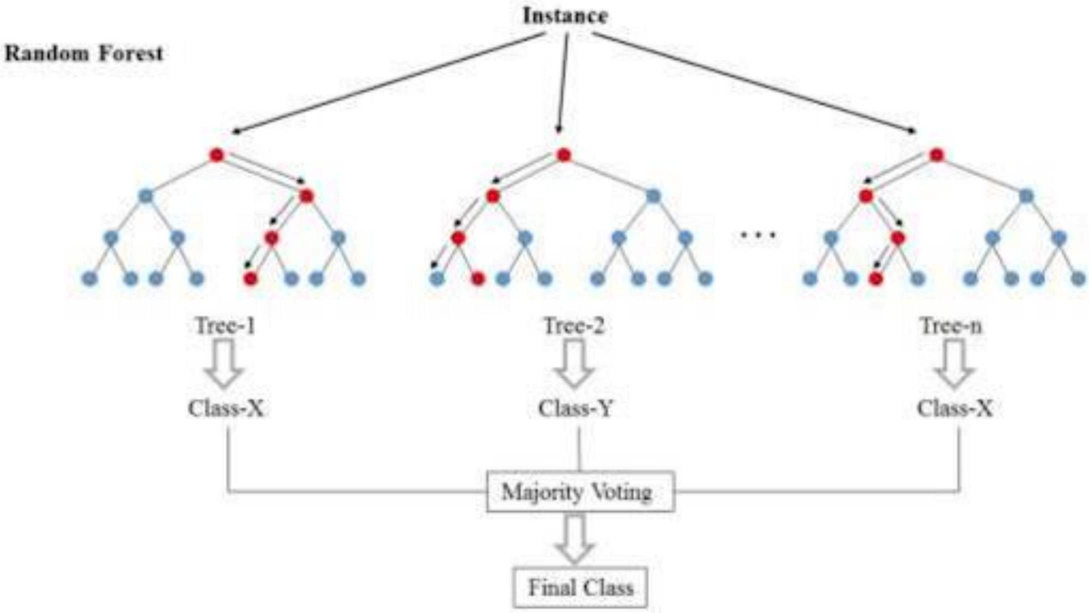


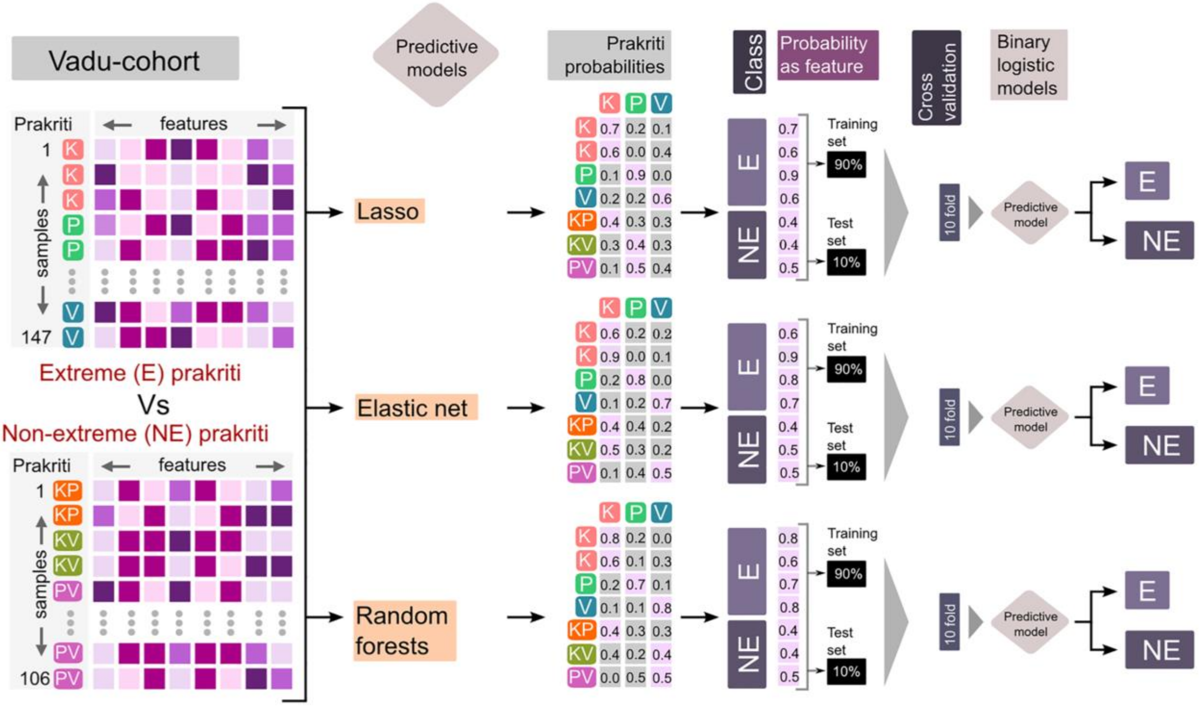
Fig. 11: Random Forest Algorithm

If we randomly pick data points from our dataset and label them randomly then the probability that we labeled them incorrectly is called Gini impurity. The aim is to make the trees de-correlated and prune the trees by setting a stopping criteria for node splits - Gini Impurity.

From the total features we take greatest integer of the square root of total number of features. Now that number of features are randomly selected from the total pool and their purity is found through gini impurity the most pure feature is then used to classify the training set. If we get the node as pure then we stop and get a leaf there else we again randomly select features and the most pure from them is again used to classify the intermediate node.

4.4. CLASSIFYING EXTREMES VS NON-EXTREMES

Using the same model, an attempt was made to classify extremes vs non extremes in an hetregenous population. The Non-extreme type did not show any preferential probability (<0.5) to the three extreme class types. This could then be used in a Binary Logistic Regression to categorize the two types by finding an optimum threshold. Figure 12 below illustrates the methodology that was used.



	K	P	V
K	0.7	0.2	0.1
K	0.6	0.0	0.4
P	0.1	0.9	0.0
V	0.2	0.2	0.6
KP	0.4	0.3	0.3
KV	0.3	0.4	0.3
PV	0.1	0.5	0.4

	K	P	V
K	0.6	0.2	0.2
K	0.9	0.0	0.1
P	0.2	0.8	0.0
V	0.1	0.2	0.7
KP	0.4	0.4	0.2
KV	0.5	0.3	0.2
PV	0.1	0.4	0.5

	K	P	V
K	0.8	0.2	0.0
K	0.6	0.1	0.3
P	0.2	0.7	0.1
V	0.1	0.1	0.8
KP	0.4	0.3	0.3
KV	0.4	0.2	0.4
PV	0.0	0.5	0.5

Class	Probability as feature
E	0.7
E	0.6
E	0.9
E	0.6
NE	0.4
NE	0.4
NE	0.5

Class	Probability as feature
E	0.6
E	0.9
E	0.8
E	0.7
NE	0.4
NE	0.5
NE	0.5

Class	Probability as feature
E	0.8
E	0.6
E	0.7
E	0.8
NE	0.4
NE	0.4
NE	0.5

Fig. 12: Classification of Extreme and Non-extreme prakriti types using supervised learning.

5. THE DIVERGENCE CLASSIFICATION MODEL (23)

Technological advances have enabled global profiling of genetic variants, RNA species, epigenetic marks, proteins, metabolites, and other previously unknown molecular features, enabling the characterization of complex biological systems over distinct molecular domains. These high-dimensional measurements have been made on thousands of samples and are collectively referred to as omics data.

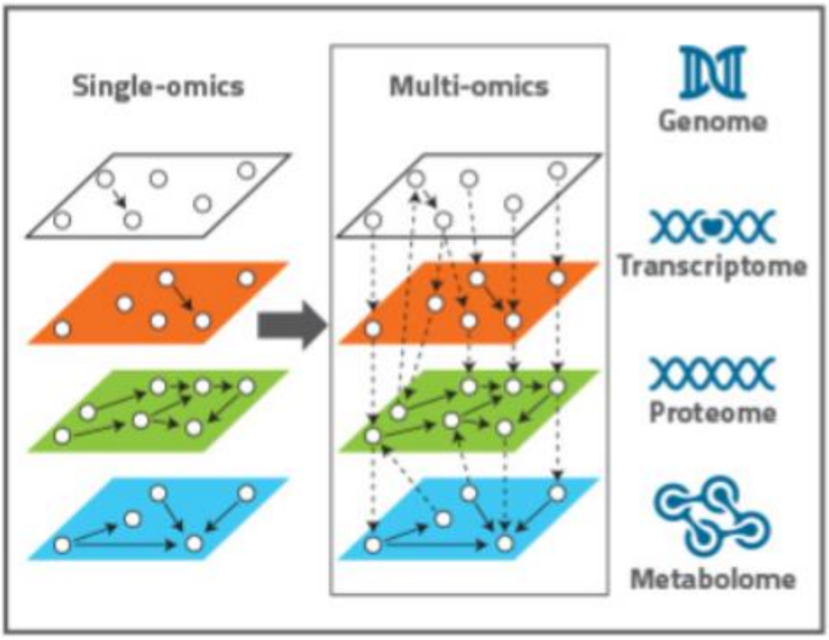


Fig. 13: Representation of omics data

5.1. MOTIVATION FOR THE MODEL

Insufficient genome-wide characterization of normal variation is impeding progress towards determining whether a single omics profile reflects molecular dysregulation that could indicative of a complex disease. There is a need for a highly simplified and personalized data representation using techniques to determine where an omics profile falls relative to a baseline, may be essential to fully realize the potential of high-dimensional omics assays.

The following datasets were used by the author for the classification model:

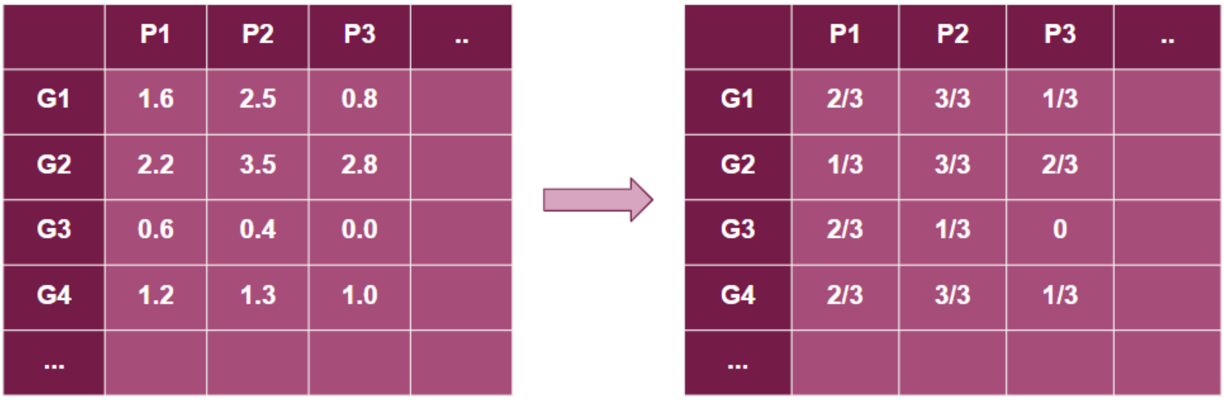
1. RNA-seq data from **TCGA** project (Breast & Prostate Cohorts) and the **UCSC** Cancer Browser project (Pan Cancer Dataset - many tissues)

2. RNA-seq expression data from the Genotype-Tissue Expression (**GTEx**) project (53 tissue subtypes)

3. Microarray-based data from the Gene Expression Omnibus (**GEO**) database (Breast Cancer profiles)

4. CpG level DNA methylation analysis - β values - from TCGA

5.2. QUANTILE NORMALIZATION



	P1	P2	P3	..
G1	1.6	2.5	0.8	
G2	2.2	3.5	2.8	
G3	0.6	0.4	0.0	
G4	1.2	1.3	1.0	
...				

	P1	P2	P3	..
G1	2/3	3/3	1/3	
G2	1/3	3/3	2/3	
G3	2/3	1/3	0	
G4	2/3	3/3	1/3	
...				

Fig. 14: Quantile Normalization on three samples and four genes

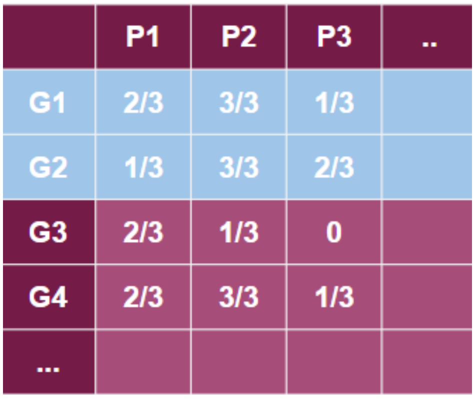
Each individual profile is transformed into quantile space that is each element of the profile is replaced by its normalized rank with respect to the other elements in the same profile. Sensitivity to platform, data modalities and preprocessing are important barriers. Divergence model begins with the initial conversion of intra-sample raw feature values to sample ranks to minimize preprocessing and batch effects. Figure 14 above explains the quantile normalization of three samples of four genes.

5.3. NON PARAMETRIC ESTIMATION

It makes no assumptions about the probability distributions of the variables being assessed. Non-parametric approaches are appropriate for exploratory purposes, and should be used if the data does not follow a simple parametric form eg. ranked data. The shape of the density function cannot easily be determined algebraically but visualization methodology can assist in this task, e.g using histograms and scatter plots. (28)

5.4. SUPPORT ESTIMATION

From all the omics profiles, let's suppose a subset of four genes (G1, G2, G3 and G4) is selected and N people (P1...PN) are used to form the baseline distribution, which form the support space as shown in the figure below.



	P1	P2	P3	..
G1	2/3	3/3	1/3	
G2	1/3	3/3	2/3	
G3	2/3	1/3	0	
G4	2/3	3/3	1/3	
...				

Fig. 15: Selection of subset of genes for support estimation

This space acts as a "covering" to find the divergent set for the test features. The space is formed by the union of the ball-spaces centred at these selected baselines points with radius determined by a tuned-parameter (gamma) and the number of baseline samples selected (n) that is-

$$\hat{\mathcal{U}}_0^S = [0, 1]^m \cap \bigcup_{k=1}^{n_0} B(U^S(k), r_k),$$

where-

- $B(U^S(k), r_k)$ denotes the closed ball of center $U^S(k)$ and radius r_k
- $\hat{U}^S$ is the estimated support
- i.i.d. samples $U^S(1), \ldots, U^S(n_0)$ are now n_0 points in $[0, 1]^m$.

As shown in figure 16, all the points lying outside the estimated support area are said to be 'divergent' or 'dysregulated'. These points make the divergent set which reflects the degree of dysregulation in the population tested.

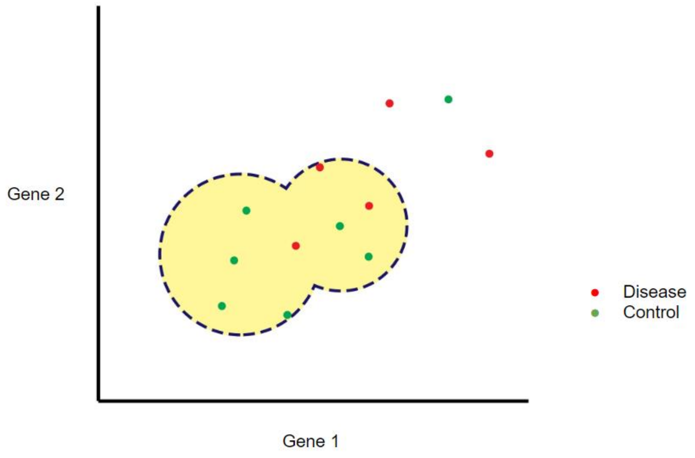


Fig. 16: Testing a disease population on a support estimated on Control population

5.5. APPLICATIONS OF THE DIVERGENCE MODEL

5.5.1. DIVERGENCE OF TUMOR PROFILES FROM A NORMAL BASELINE

50 normal samples (blue points) were used to compute the area of support (shown by the gray shade). 50 Luminal A samples are tested and if falling outside the support, were declared divergent. The difference in the level of divergence between normal and tumor samples was highly significant. (P-value < 10^{6}). *(Refer Figure 17)*

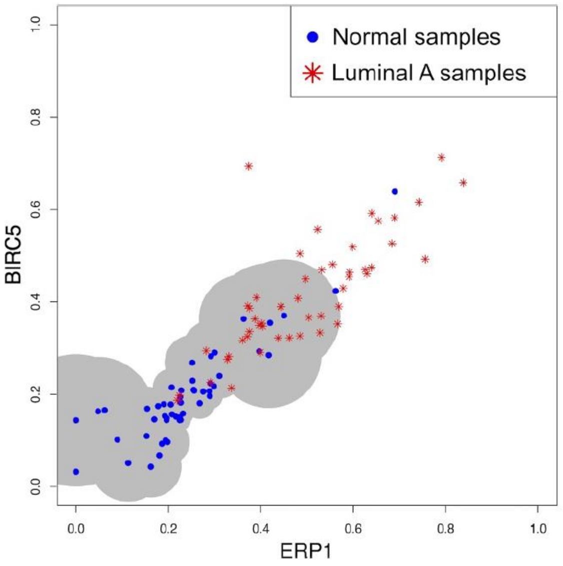


Fig. 17: Divergence of Tumor Profiles

5.5.2. COMPARISON BETWEEN NORMAL TISSUE TYPES

The relative degrees of divergence depend on how related the tissue types are, e.g. samples from other tissues that share common cell types with the baseline have fewer divergent features (e.g., breast and adipose tissue in A) than those that do not (e.g. skin and brain in B). *(see figure below)*

This suggests that tissue-specific gene expression baselines could be used to predict the tissue type of samples of unknown type.

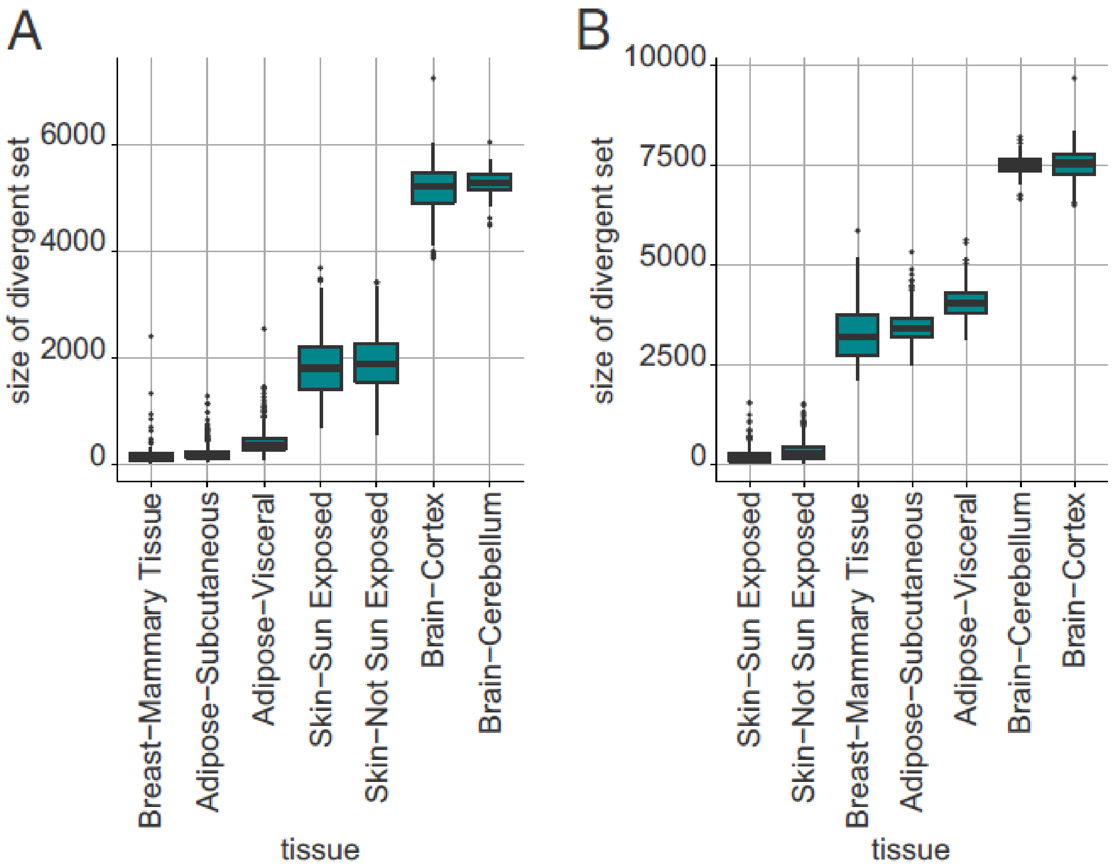


Fig. 18: Divergent profile comparisons between normal tissue types.

5.5.3. DISEASE-PATHWAY ASSOCIATIONS

Since diseases involve entire groups of genes, changes of expression in groups of pre-defined gene sets (Hallmark FGS) is used. Divergence probabilities of 50 Hallmark Functional Gene sets were used to understand divergence among Breast cancer phenotypes. Large differences in divergent probabilities was observed among subtypes for certain pathways. The figure below shows a heat map of the results.

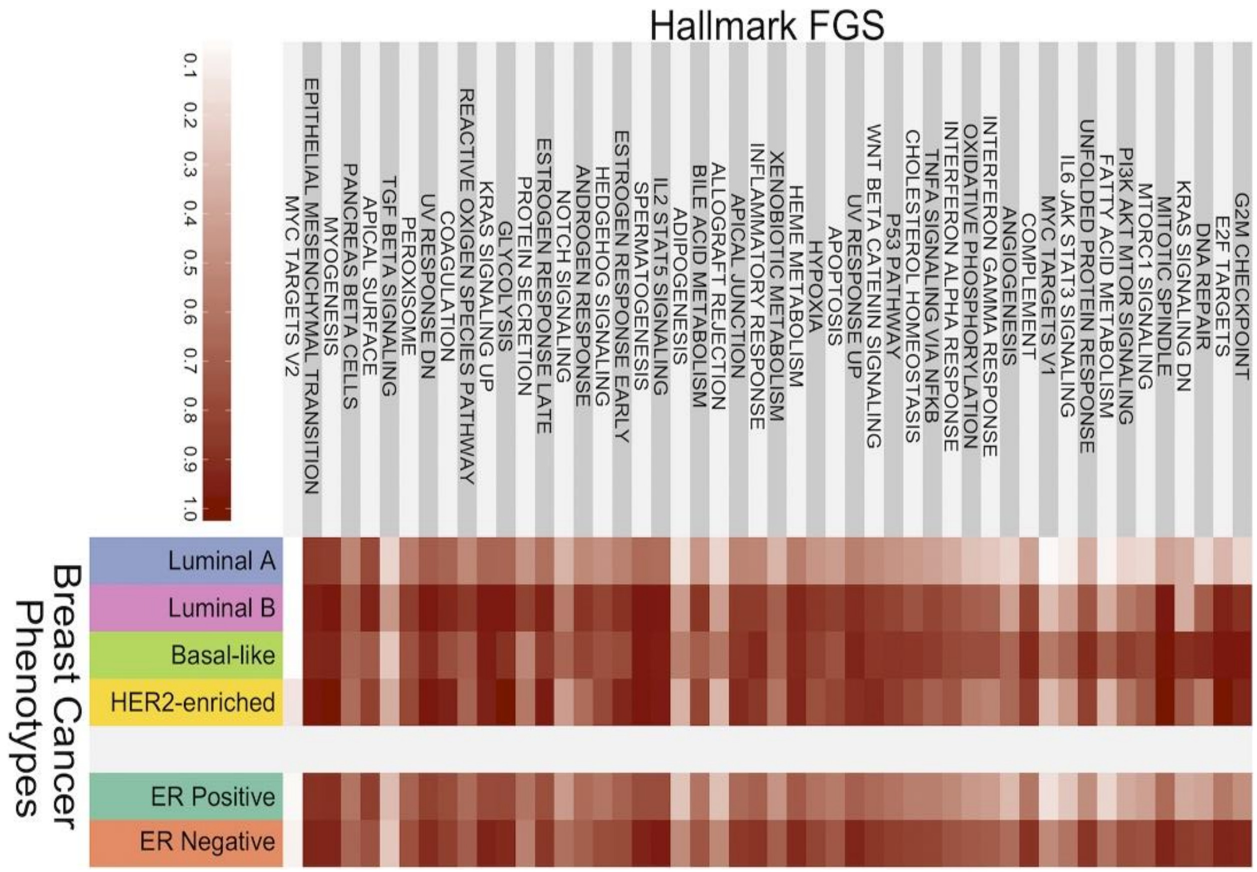


Fig. 19: Heat Map for disease pathway associations.

6. ENTROPY BASED DISTANCE METRIC (29)

The data can be classified into two major classes the numeric type and the categorical type. The latter one is further bifurcated into Nominal and Ordinal type. Data of nominal type is of qualitative nature, and hence no arithmetic operations can be performed on them, on the other hand, the ordinal data type has similar attributes as that of nominal data type except it inherits a natural order in it. The following example will clarify more: - consider a set of eye color {red, blue, green, grey, black, brown} and another ordinal dataset of eyebrow size {large, medium, small} as we can see the ordinal dataset contains qualitative data along with a natural order in it.

In data analysis many times categorical data contain both subtypes and hence treating ordinal data as nominal leads to loss of the constitutional order it has. For example, given choices between categories {very-good, good, neutral, bad, and very-bad} comparing any two choices, let them be “very good” and “bad” the intermediate ones the “good” and the “neutral” have to be considered on deciding the final choice and they can’t be skipped. Hence there is a need to have a distance metric which preserves the information regarding the order.

The idea of entropy-based distance metric (EBDM) is based on the Shannon information entropy theory. The theory suggests that the amount of entropy of an element can be associated with the information contained within it or in other words, the thinking cost associated with it. To mine, the information stored in the inherent constitutional order of the ordinal data type the distance calculated between two categories can be thought as the “thinking cost” for deciding between these two categories.

This uniform definition for both the nominal and ordinal data types postulate that the distance between two choices is the **cumulative thinking cost** of all the choices that are considered. Hence the distance between two ordinal *categories* (choices) is measured by calculating the entropy of all the categories ordered between them including themselves (See figure 20).

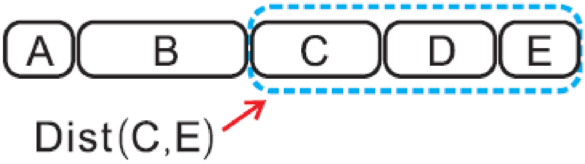


Fig. 20: The distance between two categories

The EBDM allows us to quantify categorical data containing both ordinal and nominal *attribute* (data or data associated with it) and hence allows us to find distance between two *data-objects* (in our case they are individuals, and generally means a single set that is characterized by numerous features).

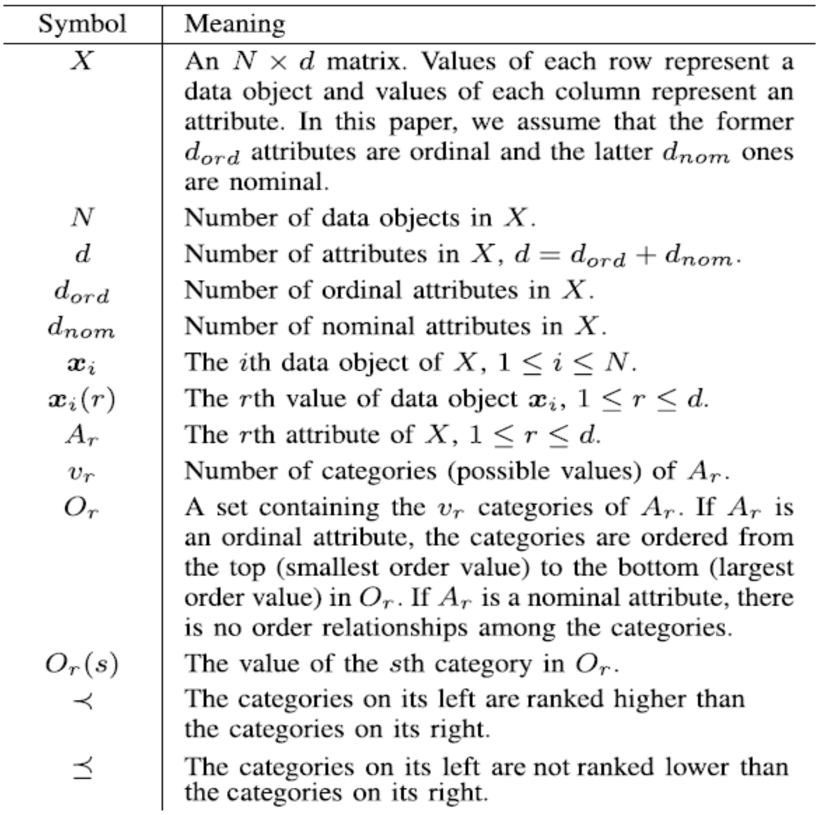
There is also a feature or attribute weighting scheme that allocates the weight to each feature according to the contribution made by each in calculating distance measurement and is also universal.

6.1. THE APPROACH

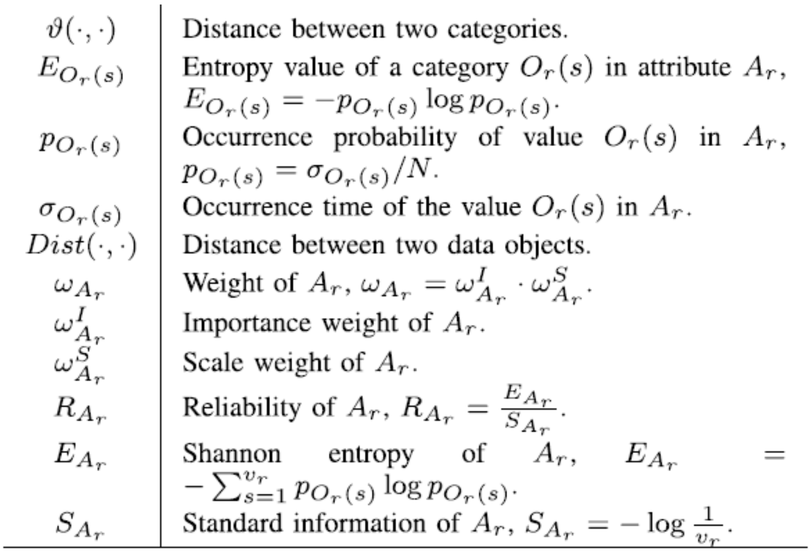
let the data set be {x1,x2,……..xN} having N elements with every data point being characterized by *d* features and for simplicity let us assume that the first d_{orb} be ordinal features and latter d_{nom} be nominal ones, then each feature let A_r can be represented by category set *Or* = {*Or (*1*), Or (*2*)*, . . . , *Or (vr)*} having v_r number of categories. Moreover, in the ordinal features the categories are represented as *Or (*1*)* ≺ *Or (*2*)* ≺ ... ≺ *Or* (v_r *)* where the symbol "≺" means categories in left are ranked higher. Hence a data object let say $\mathbf{x}_i$ can be represented as ***xi*** = {*O*1*(i*1*), O*2*(i*2*)* ..., O_{dord} (i_{dord}),O_{dord}+1(i_{dord}+1), O_{dord}+2(i_{dord}+2),.., *Od* (i_d *)*}. For the ordinal part of ***xi*** , the sequence numbers *i*1, *i*2, . . . , *id*ord indicate that the categories ranked *i*1th, *i*2th,…, i_{dord}th in category sets *O*1, *O*2, . . . , O_{dord} have been taken by the 1st, 2nd,…,d_{ord}th values of object ***xi*** , respectively. For the nominal part of ***xi*** , the sequence numbers *id*ord+1, *id*ord+2, . . . , *id* indicate that the *id*ord+1th, *id*ord+2th,…, i_d th categories in category sets O_{dord+1}, O_{dord+2}, . . . , *Od* have been taken by the d_{ord} + 1th, d_{ord} + 2th,…, *d*th values of object ***xi*** , respectively.

The commonly used symbols are shown in Table 2.

Table 2: Frequently Used Symbols



Symbol	Meaning
X	An $N \times d$ matrix. Values of each row represent a data object and values of each column represent an attribute. In this paper, we assume that the former d_{ord} attributes are ordinal and the latter d_{nom} ones are nominal.
N	Number of data objects in X.
d	Number of attributes in X, $d = d_{ord} + d_{nom}$.
d_{ord}	Number of ordinal attributes in X.
d_{nom}	Number of nominal attributes in X.
$\boldsymbol{x}_i$	The ith data object of X, $1 \leq i \leq N$.
$\boldsymbol{x}_i(r)$	The rth value of data object $\boldsymbol{x}_i$, $1 \leq r \leq d$.
A_r	The rth attribute of X, $1 \leq r \leq d$.
v_r	Number of categories (possible values) of A_r.
O_r	A set containing the v_r categories of A_r. If A_r is an ordinal attribute, the categories are ordered from the top (smallest order value) to the bottom (largest order value) in O_r. If A_r is a nominal attribute, there is no order relationships among the categories.
$O_r(s)$	The value of the sth category in O_r.
$\prec$	The categories on its left are ranked higher than the categories on its right.
$\preceq$	The categories on its left are not ranked lower than the categories on its right.

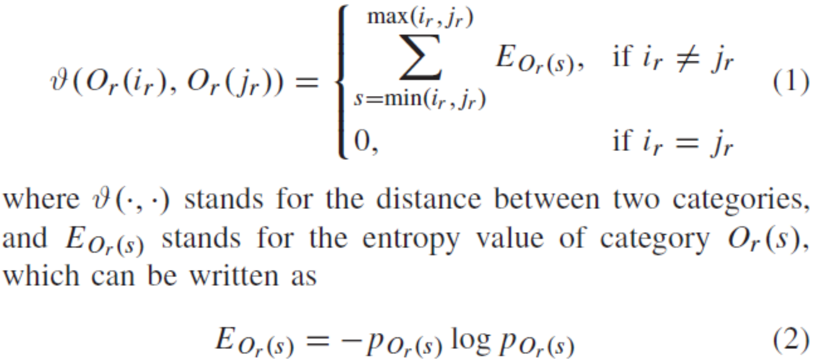


$\vartheta(\cdot,\cdot)$	Distance between two categories.
$E_{O_r(s)}$	Entropy value of a category $O_r(s)$ in attribute A_r, $E_{O_r(s)} = -p_{O_r(s)} \log p_{O_r(s)}$.
$p_{O_r(s)}$	Occurrence probability of value $O_r(s)$ in A_r, $p_{O_r(s)} = \sigma_{O_r(s)}/N$.
$\sigma_{O_r(s)}$	Occurrence time of the value $O_r(s)$ in A_r.
$Dist(\cdot,\cdot)$	Distance between two data objects.
ω_{A_r}	Weight of A_r, $\omega_{A_r} = \omega^I_{A_r} \cdot \omega^S_{A_r}$.
$\omega^I_{A_r}$	Importance weight of A_r.
$\omega^S_{A_r}$	Scale weight of A_r.
R_{A_r}	Reliability of A_r, $R_{A_r} = \frac{E_{A_r}}{S_{A_r}}$.
E_{A_r}	Shannon entropy of A_r, $E_{A_r} = -\sum_{s=1}^{v_r} p_{O_r(s)} \log p_{O_r(s)}$.
S_{A_r}	Standard information of A_r, $S_{A_r} = -\log \frac{1}{v_r}$.

6.2. DISTANCE MEASUREMENTS USING EBDM

In figure 20 (page 23) we can clearly see to measure distance between two categories C and E all the categories between them have to considered including themselves and hence the thinking cost is also related to D if the thinking cost for D is more, then the decision making becomes more rigorous. To be specific the distance between categories *Or (ir)* and *Or (jr)* with *jr* > *ir* can be measured by estimating the cost (distance) contributions of the *jr* − *ir* + 1 categories, i.e., *Or (ir)*, *Or (ir* + 1*)*,…, *Or (jr)* and the sum of the entropy values of these categories represent the measure for above-mentioned distance.

Therefore, the distance between the *r* th value of two objects, ***xi*** and ***xj*** , from an ordinal data set *X* with *N* objects represented by d_{ord} ordinal attributes, is defined as



$$\vartheta(O_r(i_r), O_r(j_r)) = \begin{cases} \sum_{s=\min(i_r,j_r)}^{\max(i_r,j_r)} E_{O_r(s)}, & \text{if } i_r \neq j_r \\ 0, & \text{if } i_r = j_r \end{cases} \quad (1)$$

where $\vartheta(\cdot,\cdot)$ stands for the distance between two categories, and $E_{O_r(s)}$ stands for the entropy value of category $O_r(s)$, which can be written as

$$E_{O_r(s)} = -p_{O_r(s)} \log p_{O_r(s)} \quad (2)$$

One of the important conclusions that can be drawn from formula 2 is that higher the occurrence of a category lower is its entropy i.e. more occurring is “more common” and does not characterize the data set.

The measure of occurrence probability of a category can be obtained through Table 2. and the distance between two data points xi and xj is defined as

$$Dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{r=1}^{d_{\text{ord}}} \vartheta(O_r(i_r), O_r(j_r))^2.}$$

6.3. ATTRIBUTE WEIGHING

The attribute or feature weighing is important as there are two issues that have to be answered. Firstly, as we know from the information entropy theory that higher entropy means that it contains more information and hence the attribute containing more information contributes more in calculation of distance between two data points. Hence the weight for allocating importance to attributes can be defined as :-

$$\omega_{A_r}^{I} = \frac{E_{A_r}}{\sum_{s=1}^{d_{\text{ord}}} E_{A_s}}$$

Where E_{Ar} is Shannon’s entropy (Table 2). Secondly, with more categories the feature may produce larger distances and hence contribute more than needed therefore it has to be weighted properly.

$$\omega_{A_r}^{S} = \frac{\frac{1}{S_{A_r}}}{\sum_{s=1}^{d_{\text{ord}}} \frac{1}{S_{A_s}}}$$

Whereas S_{As} is the maximum entropy of an attribute (Table 2) that is the case when the occurrence of every category is equally likely. The combined weighting using the above two weights can be denoted by ω_{Ar} *the integrated weight* (Table 2).

We can explain the integrated weight through another approach of reliability (Table 2) which indicate the convincing power of the distances measured w.r.t. that attribute and its weight can be rewritten as

$$\omega_{A_r} = \frac{R_{A_r}}{\sum_{s=1}^{d_{\text{ord}}} R_{A_s}}.$$

6.4. GENERALIZING FOR CATEGORICAL DATA

When generalizing the EBDM for categorical data we only consider entropy of two categories we are considering hence the distance is measured by the sum of entropy of those two categories only. Hence the modified distance measure between categories for categorical data becomes -

$$\vartheta(O_r(i_r), O_r(j_r)) \\ = \begin{cases} \omega_{A_r} \cdot \displaystyle\sum_{s=\min(i_r, j_r)}^{\max(i_r, j_r)} E_{O_r(s)}, & \text{if } i_r \neq j_r, 0 < r \leq d_{\text{ord}} \\ \omega_{A_r} \cdot \displaystyle\sum_{s=i_r, j_r} E_{O_r(s)}, & \text{if } i_r \neq j_r, d_{\text{ord}} < r \leq d \\ 0, & \text{if } i_r = j_r \end{cases}$$

and the weight is also generalized as

$$\omega_{A_r} = \frac{R_{A_r}}{\sum_{s=1}^{d} R_{A_s}}.$$

6.5. WORK FLOW OF EBDM

We can calculate the distance matrices for distances between categories for each attribute using the given flow diagram(See figure 21) Given a data set with let N number of data points or data objects we follow following procedure.(see figure 21)

1. Calculation of v_r occurrence probability and hence entropy for each category of every attribute.
2. Calculation of v_r x v_r distance matrix for every attribute.

3. Calculation of corresponding weights for the attribute
4. Calculation of distance between two individuals.

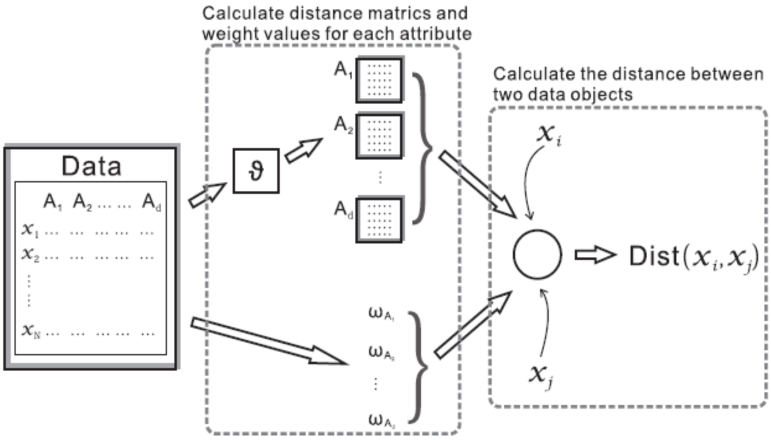

Calculate distance matrics and weight values for each attribute
Calculate the distance between two data objects
Data
A_1 A_2 …… A_d
x_1
x_2
x_N
ϑ
A_1
A_2
A_d
ω_{A_1}
ω_{A_2}
ω_{A_d}
x_i
x_j
$\mathrm{Dist}(x_i, x_j)$

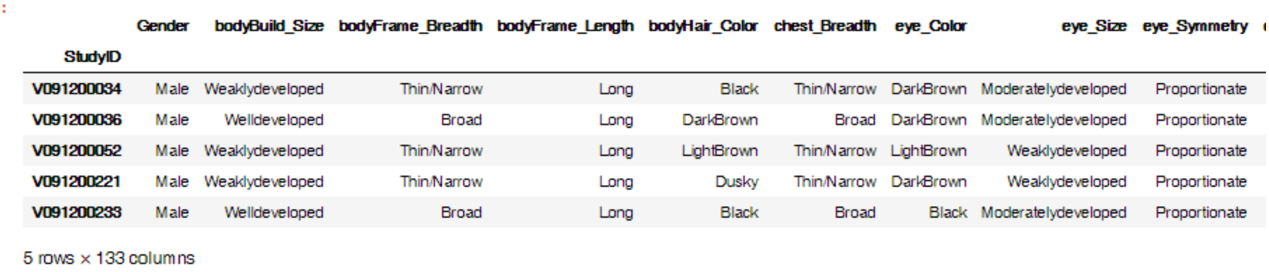
Fig. 21: Work flow of EBDM

7. METHODOLOGY

The unlabeled dataset of 147 individuals collected through a questionnaire was used across 133 attributes or phenotypes. The phenotypes were mainly categorized on four types:

1. Anatomical Features (body build size, eye size)
2. Physical Features (walking, working)
3. Physiological Features (appetite, bladder)
4. Psychological Features (memorizing speed, anger speed)

Table 3: Panda DataFrame of five individuals from Vadu Cohort



	Gender	bodyBuild_Size	bodyFrame_Breadth	bodyFrame_Length	bodyHair_Color	chest_Breadth	eye_Color	eye_Size	eye_Symmetry
StudyID									
V091200034	Male	Weaklydeveloped	Thin/Narrow	Long	Black	Thin/Narrow	DarkBrown	Moderatelydeveloped	Proportionate
V091200036	Male	Welldeveloped	Broad	Long	DarkBrown	Broad	DarkBrown	Moderatelydeveloped	Proportionate
V091200052	Male	Weaklydeveloped	Thin/Narrow	Long	LightBrown	Thin/Narrow	LightBrown	Weaklydeveloped	Proportionate
V091200221	Male	Weaklydeveloped	Thin/Narrow	Long	Dusky	Thin/Narrow	DarkBrown	Weaklydeveloped	Proportionate
V091200233	Male	Welldeveloped	Broad	Long	Black	Broad	Black	Moderatelydeveloped	Proportionate

5 rows × 133 columns

7.1. DATA CLEANING

To maintain the integrity and readability of the dataset, data cleaning techniques were used:

1. Converting all the dataset to lower case.
2. Converting the dichotomic phenotypes to Yes/ No questions instead of like/ dislike or suit/ donotsuit.
3. Identifying the individuals with unique Study ID and phenotypes by their names.

7.2. DATA PREPROCESSING

A new file was created to categorize the phenotypes into two kinds of categories:

1. Nominal Categories
2. Ordinal Categories

CategoricalDType from pandas library was used to use less memory for data frames. The idea behind the data compression algorithm used is to reduce the string data to integers. In case a dataset stores same strings a large number of times (usually a case in questionnaire data), the dataset could then be stored as integers inherently and a dictionary/ mapping from strings to integers. For example, instead of storing Poor, Good and Excellent every time for a particular attribute 0,1 and 2 could be stored inherently with a mapping {Poor: 0, Good: 1, Excellent: 2} also providing us an extra benefit of maintaining the ordering of categories within an attribute.

To maintain the readability and reusability of the code, dictionaries were used everywhere.

7.3. USING ENTROPY BASED DISTANCE METRIC

1. The pre-processed nominal and ordinal categories files were read and a mapping was made between the attribute and it's categories. For example,

{Attribute : Category1, Category2, Category3 }

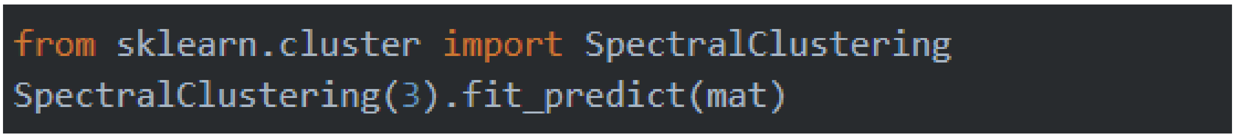
where attribute can be bodyBuild_size and it's categories be

['weaklydeveloped', 'moderatelydeveloped', 'welldeveloped']

2. For each category of an attribute its occurrence count in the dataset was calculated and the corresponding probabilities were calculated.
3. The next step was calculating category wise entropies, and storing them, for both ordinal and nominal data.
4. Reliabilities of all attributes and total reliability were calculated for assigning weights to the attributes.
5. Using these weights and the entropies, distance between categories of each attribute was calculated according to the methodology prescribed.
6. Distances between all individuals in the dataset were calculated with each other to obtain a distance matrix which served as a distance metric for further analysis and clustering objectives.

Using this distance matrix, we performed a wide-array of clustering techniques to verify if there are 3 clusters corresponding to the three extreme Prakriti types.

First of these was the Spectral Clustering, where it performs dimensionality reduction for clustering the data using the eigenvalues of the similarity matrix. In Python, this is using a library, “sklearn”.



```
from sklearn.cluster import SpectralClustering
SpectralClustering(3).fit_predict(mat)
```

Using this on our dataset, we were able to get an output assigning IDs (0,1,2) to the input data points, corresponding to the three cluster labels. However, it was in the format of an array and therefore, to further visualise those clusterings, we used the below-mentioned techniques of clustering.

7.4. MULTI-DIMENSIONAL SCALING

In multidimensional scaling (MDS), the goal is to get a visual representation of the pattern of distances (or similarities) among a set of data points. Objects that are very similar to each other (and therefore, have similar entropies in our case) are placed near each other on the plotted graph and those that are perceived to be very different from each other are placed far away from each other. Therefore, MDS can be useful in identifying underlying dimensions and plotting the multi-dimensional data using a lesser number of dimensions.
The range of dimensions starts from 2-D and extends upto the actual dimension of the data set. However, since the human mind is capable of understanding only upto 3-D, MDS is usually done to give a 2-D perspective of the data.

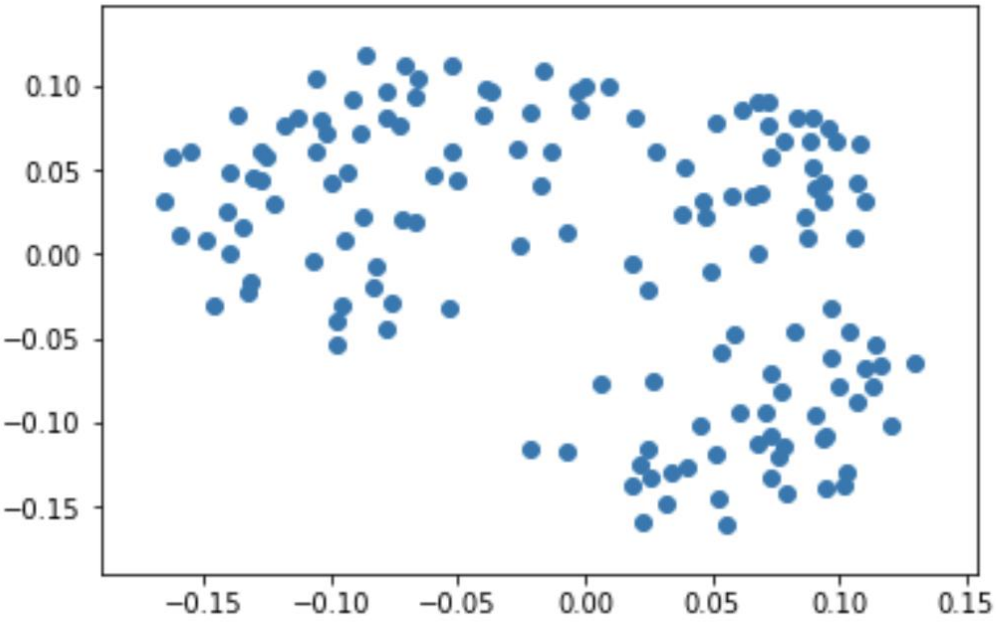
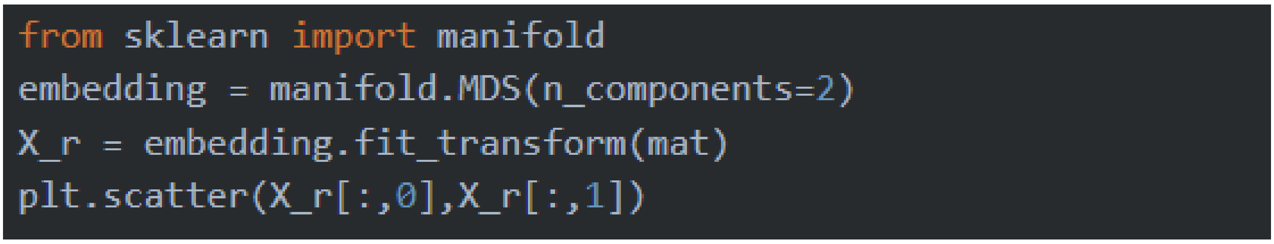


Fig. 22: MDS applied on the overall population dataset which consists only of extreme Prakriti individuals. 3 clusters can be seen in the plot, which is consistent with our hypothesis.

In Python, MDS is performed using a third-party library called sklearn. Below is the code that was used to generate the plot-



```
from sklearn import manifold
embedding = manifold.MDS(n_components=2)
X_r = embedding.fit_transform(mat)
plt.scatter(X_r[:,0],X_r[:,1])
```

7.5. AGGLOMERATIVE CLUSTERING

It is a type of hierarchical clustering where, initially each data point is considered as an individual cluster and after each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed.

Clusters are formed using a proximity (similarity) matrix. This clustering is visualised using a special tree-like diagram called a Dendogram.

In the dendrogram displayed above, each leaf corresponds to one object. As we move up the tree, objects that are similar to each other are combined into branches, which are themselves fused at a higher height. The higher the height of the fusion, the less similar the objects are.

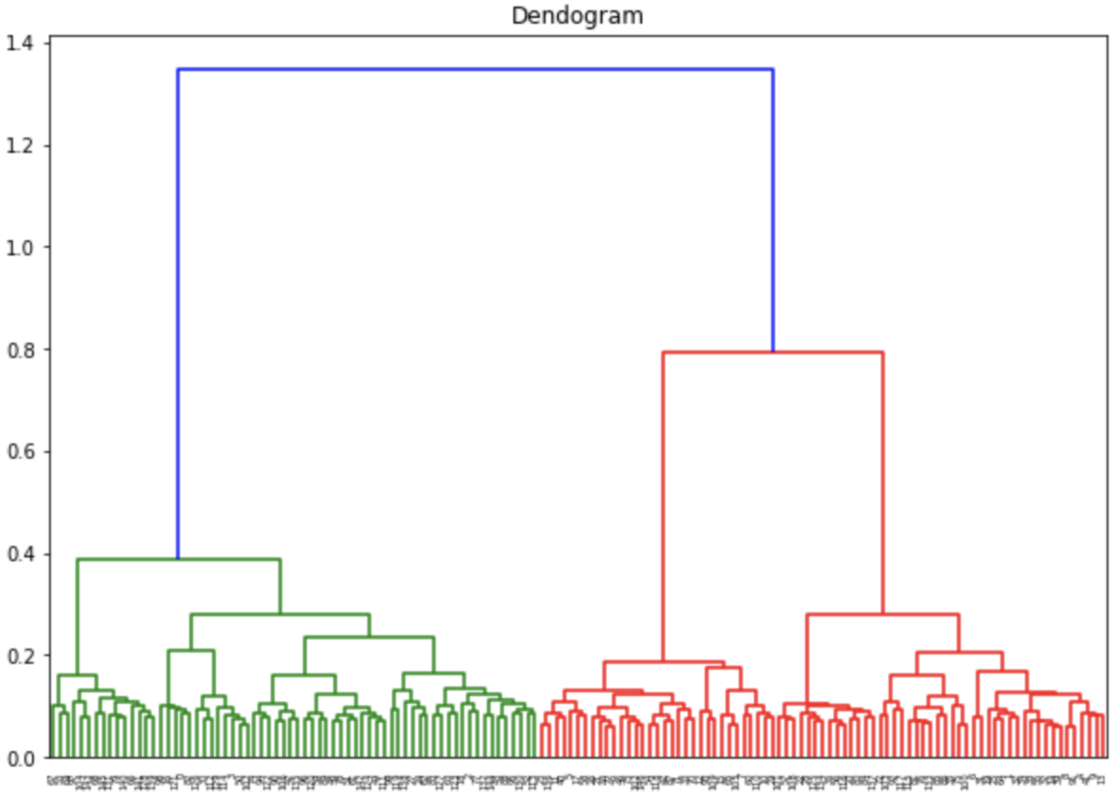
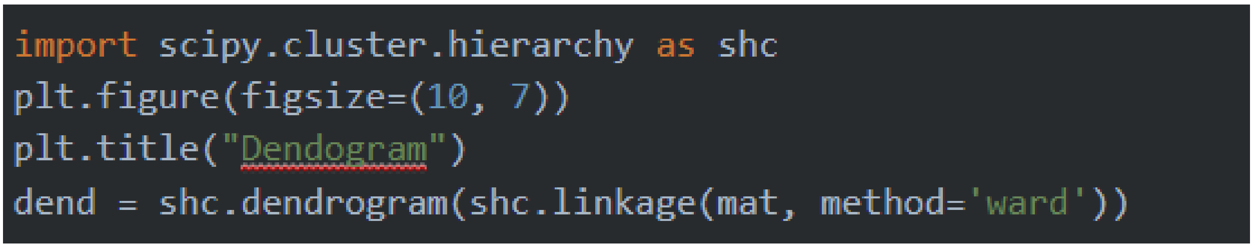


Fig. 23: Dendogram obtained from Agglomerative Clustering on the overall population.

This was performed using the “scipy” library in Python.



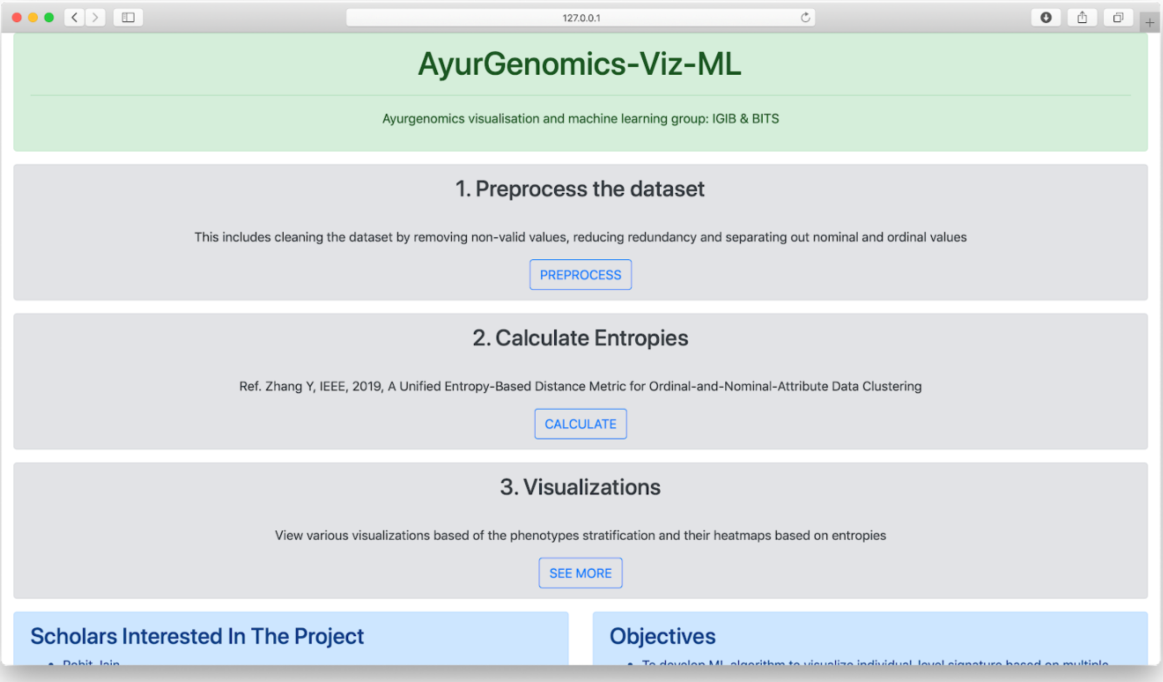
```
import scipy.cluster.hierarchy as shc
plt.figure(figsize=(10, 7))
plt.title("Dendogram")
dend = shc.dendrogram(shc.linkage(mat, method='ward'))
```

8. THE INTERFACE

The next objective of our project was to create an online, web-interface where a user can upload his/her custom dataset, and can get immediate results in the form of graphical visualizations.

The interface was created using the Django programming language, and is scalable to work on any device, mobile or computer. The UI was designed keeping in mind that even a person with no knowledge about the domain can easily work-around the website.

The first step is where the user can upload a dataset which will be preprocessed and cleaned in an easy-to-read format. (See figure 24)



AyurGenomics-Viz-ML

Ayurgenomics visualisation and machine learning group: IGIB & BITS

1. Preprocess the dataset

This includes cleaning the dataset by removing non-valid values, reducing redundancy and separating out nominal and ordinal values

2. Calculate Entropies

Ref. Zhang Y, IEEE, 2019, A Unified Entropy-Based Distance Metric for Ordinal-and-Nominal-Attribute Data Clustering

3. Visualizations

View various visualizations based of the phenotypes stratification and their heatmaps based on entropies

Scholars Interested In The Project

- Rohit Jain

Objectives

- To develop ML algorithm to visualize individual level signature based on multiple

Fig. 24: The User Interface

The interface allows us to:

1. Preprocess the data

2. Calculate Entropies

3. Visualizations

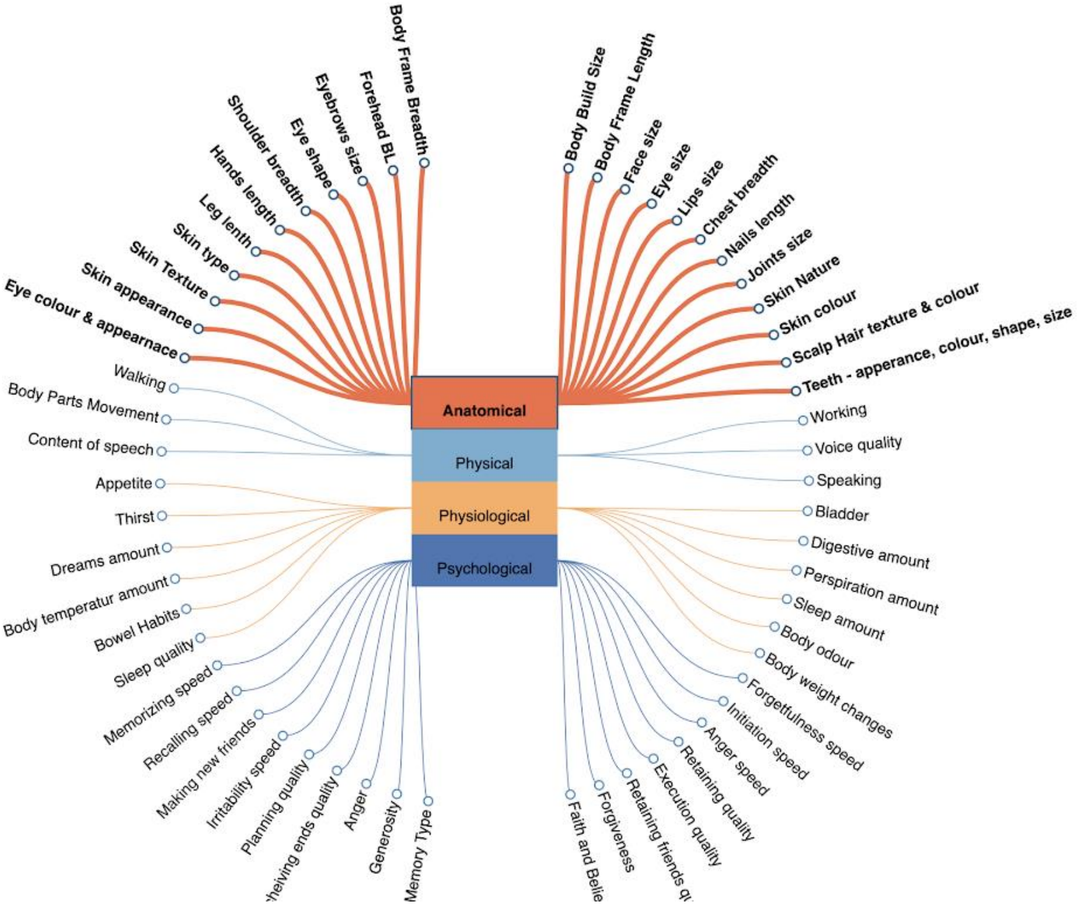


Fig. 25: Selecting a subset of Phenotypes

In the second step, all the metric calculations will be performed and the user can give queries on the entropy matrix. (See figure 25)

The entropy matrix can be visualized as a matrix with:

1. Columns c as phenotypes/ attributes of an individual.

2. Rows r as individual of the population.

3. Cell (r,c) representing the entropy of the individual corresponding to that particular attribute.

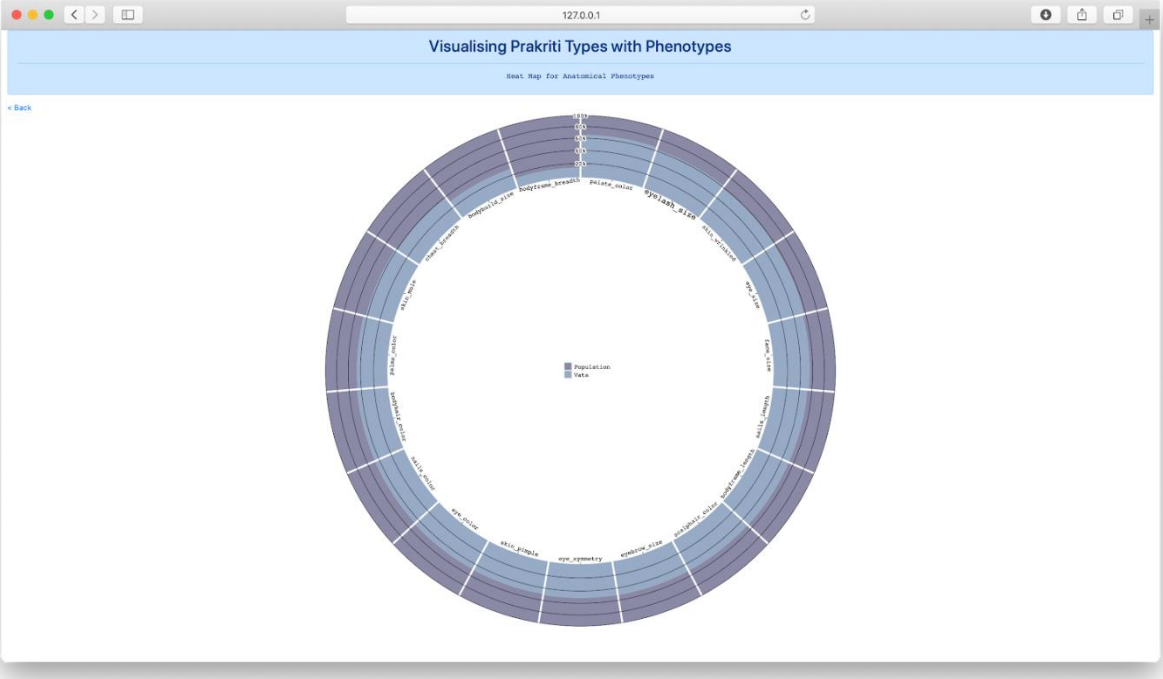
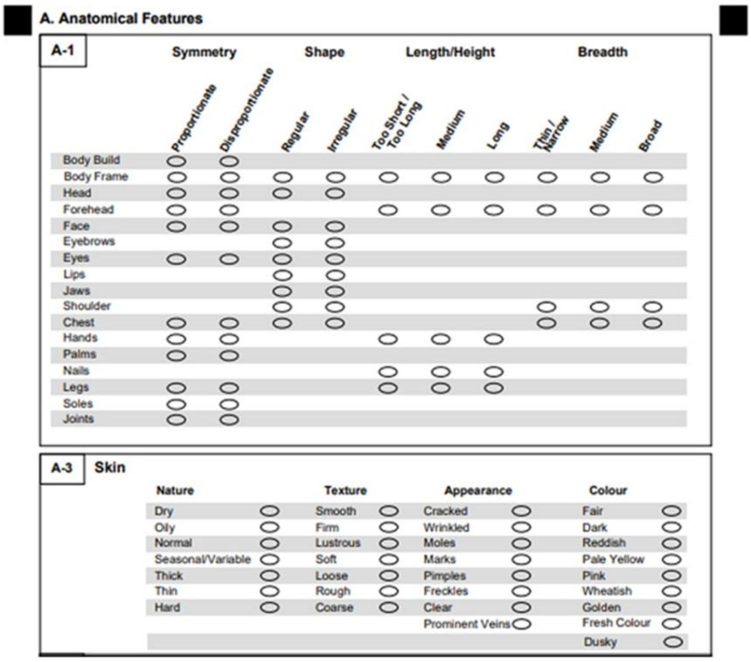


Fig. 26: Comparing Vata entropies with Population mean entropies on the Anatomical Phenotypes

In the last step, visualizations can be explored where various phenotypes can be compared with the Prakriti types and their respective entropies for those attributes.

9. INFOGRAPHICS FOR THE METHODOLOGY

Questionnaire



A. Anatomical Features

A-1	Symmetry		Shape		Length/Height			Breadth		
	Proportionate	Disproportionate	Regular	Irregular	Too Short / Too Long	Medium	Long	Thin / Narrow	Medium	Broad
Body Build	○	○								
Body Frame	○	○	○	○	○	○	○	○	○	○
Head	○	○	○	○						
Forehead	○	○			○	○	○	○	○	○
Face	○	○	○	○						
Eyebrows			○	○						
Eyes	○	○	○	○						
Lips			○	○						
Jaws			○	○						
Shoulder			○	○				○	○	○
Chest	○	○	○	○				○	○	○
Hands	○	○			○	○	○			
Palms	○	○								
Nails					○	○	○			
Legs	○	○			○	○	○			
Soles	○	○								
Joints	○	○								

A-3 Skin

Nature		Texture		Appearance		Colour	
Dry	○	Smooth	○	Cracked	○	Fair	○
Oily	○	Firm	○	Wrinkled	○	Dark	○
Normal	○	Lustrous	○	Moles	○	Reddish	○
Seasonal/Variable	○	Soft	○	Marks	○	Pale Yellow	○
Thick	○	Loose	○	Pimples	○	Pink	○
Thin	○	Rough	○	Freckles	○	Wheatish	○
Hard	○	Coarse	○	Clear	○	Golden	○
				Prominent Veins	○	Fresh Colour	○
						Dusky	○

Fig. 27: The Questionnaire

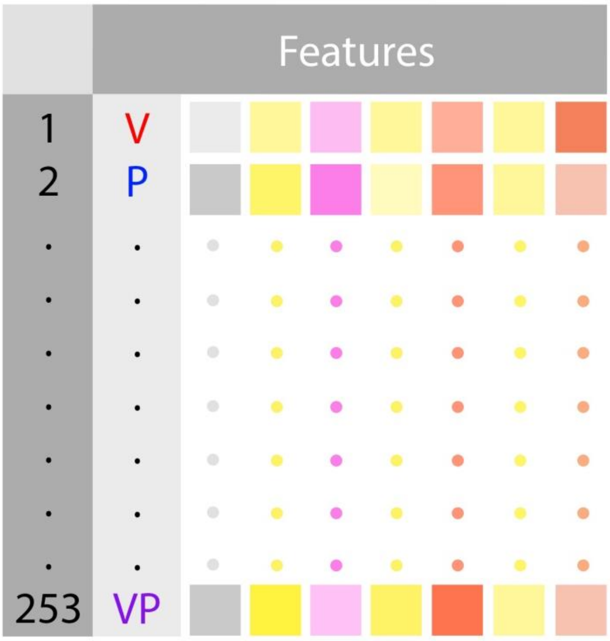
The data was collected via a survey conducted through a questionnaire on the individuals from Vadu Cohort based on their phenotypic attributes:

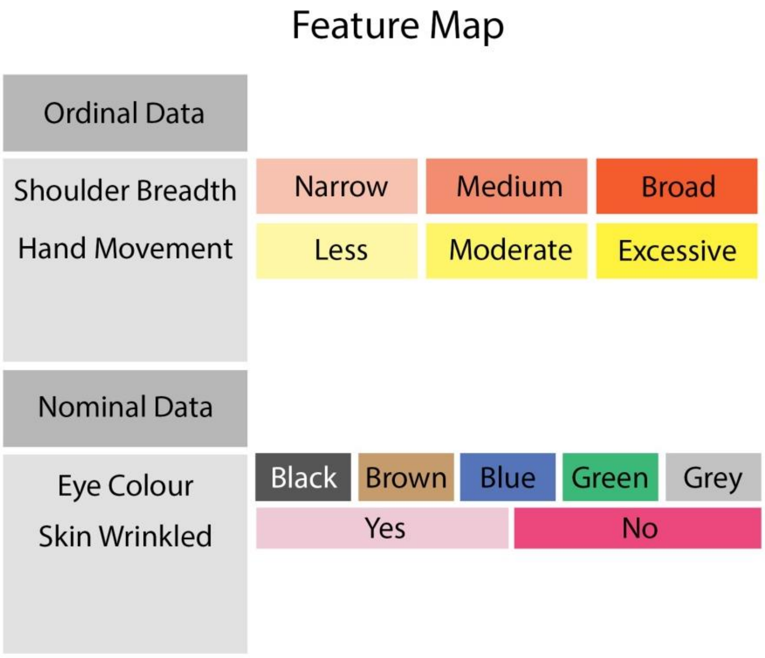
1. Anatomical
2. Physical
3. Physiological
4. Psychological

Fig. 28: The Dataset

The data collected was stored as n*m matix where n = number of rows represting the individuals and m = number of columns with column 1 as the Prakriti labels and the next (m-1) columns as the phenotypic attributes. Individuals are identified using their unique study ID while the columns were identified using the phenotype name. A cell represents the response of an individual corresponding to a particular attribute.

Dataset





Feature Map

Ordinal Data			
Shoulder Breadth	Narrow	Medium	Broad
Hand Movement	Less	Moderate	Excessive

Nominal Data					
Eye Colour	Black	Brown	Blue	Green	Grey
Skin Wrinkled	Yes	No			

Fig. 29: Feature Map

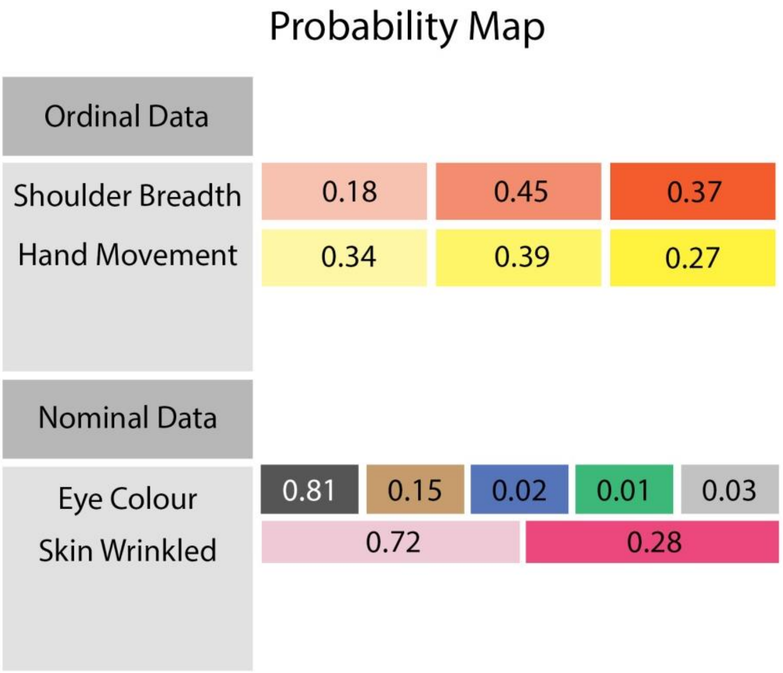
The phenotypes were categorized into two types:

1. Ordinal Data
2. Nominal Data

A feature map (a mapping) was created from the categories to a number system as to preserve the ordering property in the ordinal data. It was extended to the nominal data as well so as to maintain the uniformity in the data since any ordering can be considered in case of nominal data.

Fig. 30: Probability Map

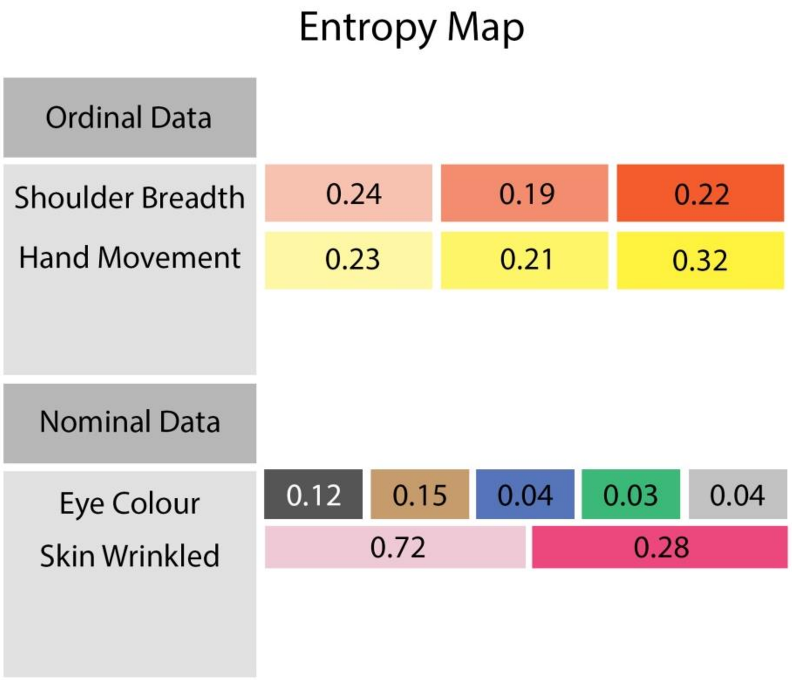
The occurrence probabilities were calculated for each of the choices of an attribute to identify the popular choices for each of the attributes within a population. A mapping similar to the feature map was created as shown in the figure.



Probability Map

Ordinal Data			
Shoulder Breadth	0.18	0.45	0.37
Hand Movement	0.34	0.39	0.27

Nominal Data					
Eye Colour	0.81	0.15	0.02	0.01	0.03
Skin Wrinkled	0.72	0.28			



Entropy Map

Ordinal Data					
Shoulder Breadth	0.24	0.19	0.22		
Hand Movement	0.23	0.21	0.32		
Nominal Data					
Eye Colour	0.12	0.15	0.04	0.03	0.04
Skin Wrinkled	0.72	0.28			

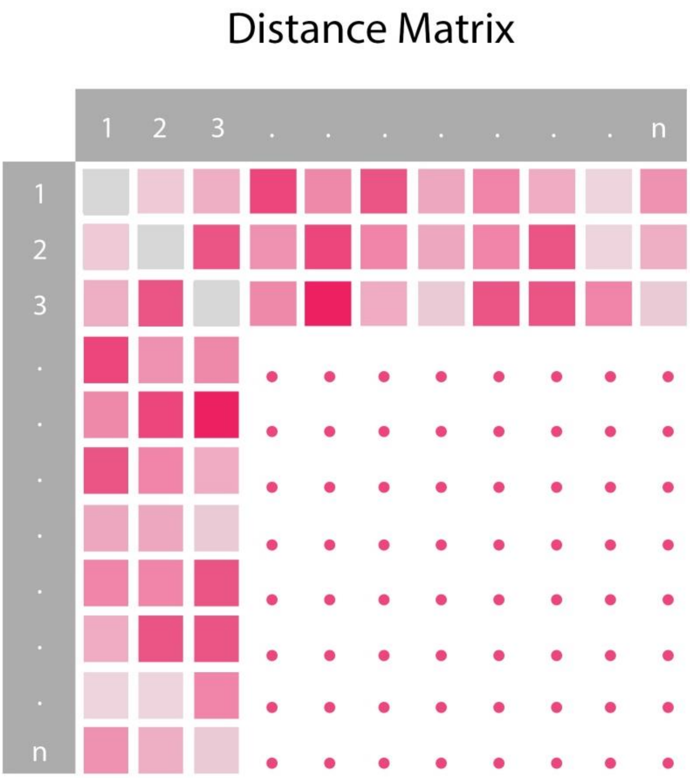
Fig. 31: Entropy Map

The mapping was further extended to calculate the entropies for the choices of each attribute. The higher the probability of a choice, the lesser is the information stored in the choice and thus, a lesser entropy value.

Fig. 32: Distance Matrix

The distance matrix is a n x n matrix where n = number of individuals in the population. The matrix was created using the entropies of the choices selected by the individuals. (Refer section 6 for the entropy concept.)

The matrix obtained was a symmetric matrix with a diagonal elements zero since the an individual is identi



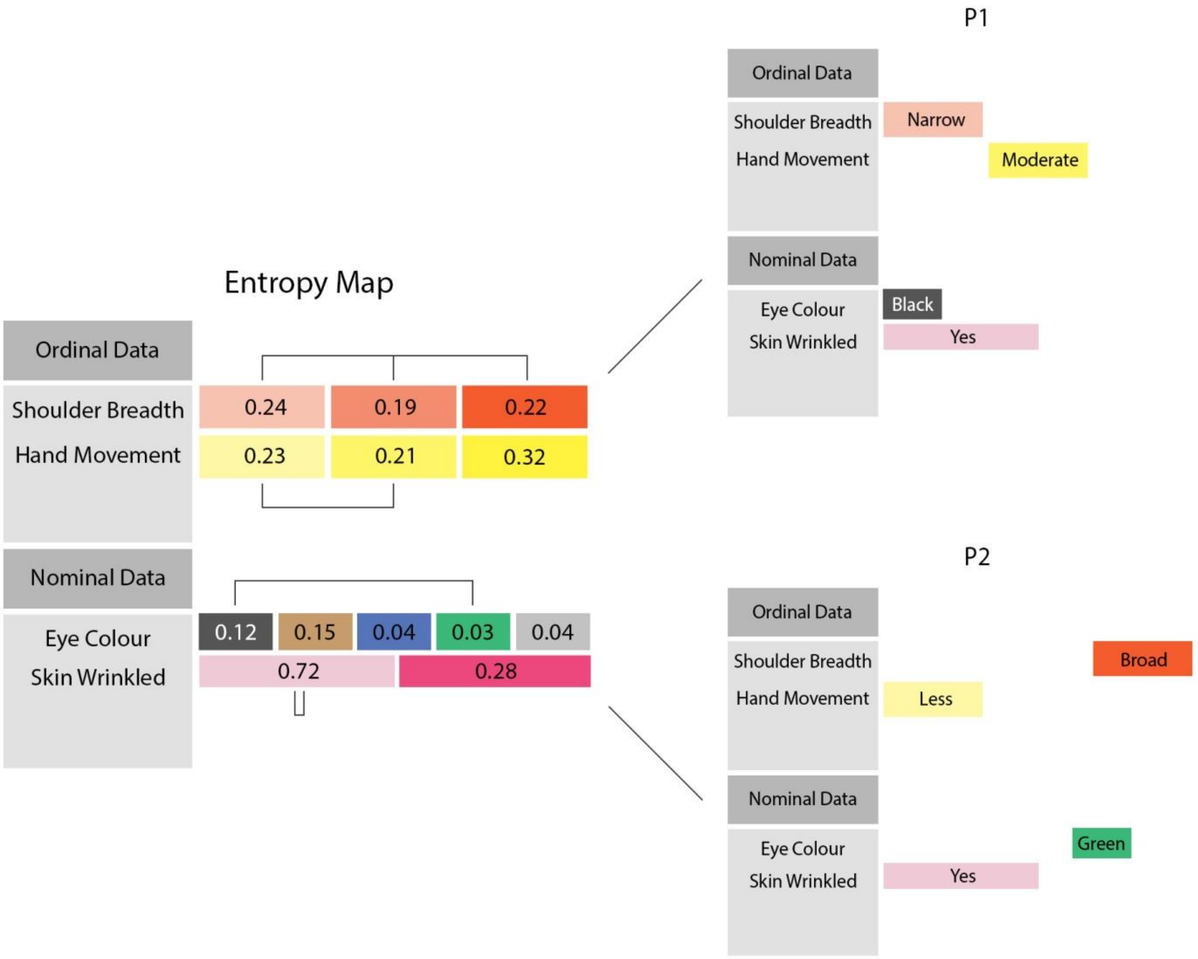


Fig. 33: Visualizing the mappings

The mappings can be understood using the above info graphic. The figure shows an entropy map for different attributes. Let’s say we take person 1 whose response was Narrow for Shoulder Breadth. Thus, from the entropy map we could identify the entropy for this choice as follows:

Ordinal Data -> Shoulder Breadth -> 0.24

Similarly it can be done for all the choices of each attribute for all the individuals to get a distance metric that may be used for further exploration.

10. RESULTS

1. We used the concepts of entropy based distance metric and information theory in order to visualize data. The first attempt was to **integrate ordinal and nominal data** on a common metric.

2. Using the distance matrix and different clustering techniques (MDS plotting and agglomerative clustering) we **got three distinct natural clusters** corresponding to each of the Tridosha.

3. The entropy method could be used to identify the **phenotypic attributes and/or choices** contributing towards a particular Prakriti by comparing the entropy values in a Prakriti vs Population graph.

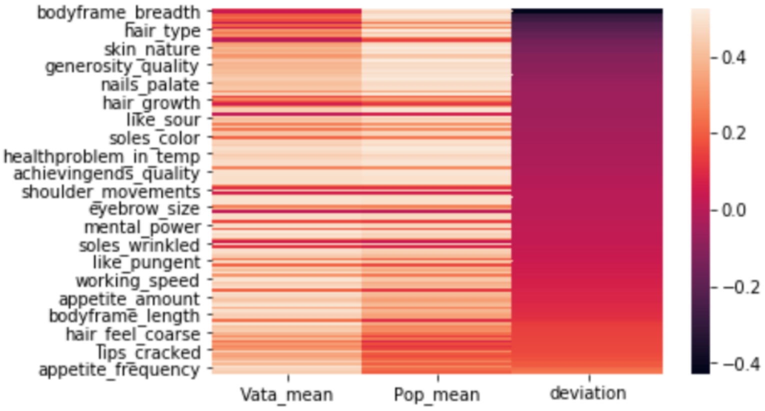


Fig. 34: Heatmap to identify the Vata attributes

Figure 34 shows the entropy deviation of Vata w.r.t. The overall population. Negative deviation indicates that Vata has lesser entropy for that attribute as compared to the overall population, and hence that attribute carries lesser importance in Vata. For positive deviation, it's the vice-versa

4. Rare phenotype choices could be identified using the concept. The rarer the choice, the lower the occurrence probability and thus higher the entropy value.

5. An interface was made to make different kind of visualizations to interpret the data.

11. CONCLUSION

The entropy based distance metric provides a good metric to analyze a questionnaire data. The metric could be used to integrate the different types of categorical data for further analysis. An attempt was made to identify different attributes contributing towards each of the Prakriti type using the fundamental concepts of information theory and mathematics. We intend to extend the idea of divergence classification on our model to create a prediction model to identify the Prakriti type of a new individual and hence stratify individuals into different groups. The stratification model could then be used to work towards precision medicine and help towards drug efficacy rates.

12. REFERENCES:

1. Systems Biology and P4 Medicine: Past, Present, and Future PMID: 23908862
2. Hieter, P. & Boguski, M. Functional genomics: it's all how you read it. Science 278, 601–602 (1997). PMID: 9381168
3. Ozaki, K. et al. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat. Genet. 32, 650–654 (2002). PMID: 12426569
4. Golub, T. R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
5. Oliver, S. Guilt-by-association goes global. Nature 403, 601–603 (2000) PMID: 10688178
6. Watson JD, Crick FH. 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 171: 737–38
7. Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. PNAS 74: 560–64 PMID: 265521
8. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860–921 PMID: 11237011
9. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321–332. doi:10.1038/nrg3920 PMID: 25948244
10. Camacho et al., Next-Generation Machine Learning for Biological Networks, Cell (2018), PMID: 29887378
11. E. S. Lander et al., ''Initial sequencing and analysis of the human genome,'' PMID: 15829235
12. Deep learning for genomics, https://doi.org/10.1038/s41588-018-0328-0
13. Karr et al., "A whole-cell computational model predicts phenotype from genotype. Cell 150, 389-401. PMID: 22817898

14. Chen, B., and Butte, A.J. (2016). Leveraging big data to transform target selection and drug discovery. Clin. Pharmacol. Ther. 99, 285–297. PMID: 26659699

15. di Bernado et al. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nat. Biotechnol. 23, 377–383.

16. L. Cong et al., ‘‘Multiplex genome engineering using CRISPR/Cas systems,’’ Science, vol. 339, no. 6121, pp. 819–823, 2013. PMID: 23287718

17. September 2016, IAS, blogadmin, Ayurgenomics – a new player in biomedical sciences

18. Prasher B., Gibson G. and Mukerji M. 2016 Genomic insights into ayurvedic and western approaches to personalized medicine. PMID: 27019453

19. Parasuraman S, Thing GS, Dhanaraj SA. Polyherbal formulation: Concept of Ayurveda. Pharmacogn Rev. PMID: 25125878

20. 2017, PLoS One, Recapitulation of Ayurveda constitution types by machine learning of phenotypic traits PMID: 28981546

21. 2011, ACS Chemical Biology, Ayurgenomics: A New Way of Threading Molecular Variability for Stratified Medicine

22. 2016, Journal of Genetics, Genomic insights into ayurvedic and western approaches to personalized medicine

23. 2018, PNAS, Digitizing omics profiles by divergence from a baseline

24. 2008, Journal of Translational Medicine, Whole genome expression and biochemical correlates of extreme constitutional types defined in Ayurveda

25. 2005, Springer link, the Indian genome variation database: a project overview.[pmid :16133172]

26. 2017, Journal of Ethnopharmacology, Ayurgenomics for stratified medicine: TRISUTRA consortium initiative across ethnically and geographically diverse Indian populations[pmid:28981546]

27. 2010, PNAS, EGLN1 involvement in high altitude adaptation revealed through genetic analysis of extreme constitution datatypes defined in Ayurveda.[pmid: 20956315]

28. Devroye L, Wise GL (1980) Detection of abnormal behaviour via nonparametric estimation of the support. SIAM J Appl Math 38:480–488.

29. Zhang Yiqun, Cheung Yiu-ming.2019. A Unified Entropy-Based Distance Metric for Ordinal-and-Nominal-Attribute Data Clustering. [PMID:30908240]