Classification of *E.coli, Lactobacillus* and *Bacillus subtilis* using computational intelligence approach

Thesis submitted in fulfillment of the requirements for the award of degree of

Master of Science

in

Biochemistry

Submitted by

Chirag Sharma
(Reg no: 301507002)

Under Supervision of

Dr. Prashant Singh Rana

Thapar University
Patiala-147004, Punjab, India
June 2017

Candidate's Declaration

I, hereby declare that the work presented in the thesis entitled "**Classification of *E.coli,* *Lactobacillus* and *Bacillus subtilis* using computational intelligence approach**" in the fulfillment of the requirement for the award of the Degree of Masters of Science in Biochemistry, School of Chemistry and Biochemistry, Thapar University, Patiala, India, is an authentic record of my own and carried out under the supervision of Dr. Prashant Singh Rana , Assistant Professor, Department of Computer Science and Engineering, Thapar University, Patiala, India. The matter embodied in this thesis has not been submitted in any part or full to any other university or institute for the award of any degree in India or abroad.

Date: 17th July, 2017 Chirag Sharma

This is to certify that the above statement made by student concerned is correct and true to the best of my knowledge.

Dr. Prashant Singh Rana
Assistant Professor
Department of Computer Science and Engineering
Thapar University, Patiala-147004
Punjab (India)

Abstract

Classification of bacterial species is an important thing in the biochemical sciences. The right distinguishing proof of microorganisms is of essential significance to microbial systematists and in addition to researchers required in numerous different zones of connected research and industry. Practically, it can be time consuming and costly as well.

The work in this thesis mainly focuses on prediction of bacterial classes using machine learning methods. Its objective is to find out the optimum parameters for the bacterial classification from the measurable features such as concentration, absorbance and pH values of the given solution.

We collected the practically performed data and arranged it in order for simulating it in R. Four different bacterial species were taken and their ordered data was simulated in R. Four machine learning models were used i.e. Random forest, Decision tree, SVM and Linear model with eight different parameters.

Comparison of the performances of each of the applied machine learning model was done to know about the most accurate model. And at last, k fold cross validation was done in order to investigate the robustness of the best fit model.

Keywords: Machine learning models, Simulation, R, Classification of bacteria.

Acknowledgements

Firstly, my sincere thanks go to Thapar University, Patiala for providing me with this opportunity.

I also thank Computer Science and Engineering Department and School of Chemistry and Biochemistry for providing me with all the necessary facilities for the research.

I express my gratitude to Dr. Prashant Singh Rana for his encouragement and aspiring guidance.

My heartfelt thanks go to Miss Vanshita Goel and Miss Jagmeet Kaur for their supervision, friendly advice and invaluable contribution in every aspect of research.

I am extremely thankful to Ms. Daisy Sharma, Ms. Pawan, Ms. Priya Singla, Ms. Palak Middha, Ms. Shweta Sharma, Mr. Abhishek Kapoor, Mr. Arshpreet Singh and Mr. Vivek Raturi for their timely guidance.

I would also like to acknowledge Dr. Diptiman Choudhury, Assistant Professor, Thapar University, Patiala.

Table of Contents

List of figures

List of tables

..

List of abbreviations

AUC	Area Under Curve
RF	Random Forest
SVM	Support Vector Machines
PPV	Positive Predictive Value
NPV	Negative Predictive Value
LM	Linear Model
GLM	Generalized Linear Models
TP	True Positive
FN	False Negative
FP	False Positive
TN	True Negative
Abs.	Absorbance
AI	Artificial Intelligence
SERS	Surface Enhanced Raman Spectroscopy
FAME	Fatty Acid Methyl Ester
DNA	Deoxyribo-Nucleic Acid
e.coli	*Escherichia coli*

Chapter 1

Introduction

Study of microorganisms is an important part of research in the biochemical field. Every discovery starts from base and microorganisms are the base material. Their study is known as Microbiology. They are microscopic organisms and can be single-celled or multi-cellular. Microorganisms are extremely different and incorporate all microscopic organisms, archaea and generally protozoa. This gathering additionally contains a few parasites, green growth, and some miniaturized scale creatures. Viruses are characterized as microorganisms and also some label them as non-living.

These organisms are found in different parts of biosphere such as rocks, soil, hot water springs, in the sea and as well as air. Microorganisms, under certain test conditions, have been seen to flourish in the vacuum of space. Prokaryotic organisms such as archaea and microbes weigh around 0.8 trillion tones of carbon out of the overall biomass of around 4 trillion tones [1].

Micro-organisms are significant in supplement reusing in biological systems as they decompose organic material. Few micro-organisms can settle nitrogen and are an essential piece of nitrogen cycle; moreover late examinations demonstrate that airborne micro-organisms might assume the part of precipitation as well as climate. Micro-organisms are additionally misused in the field of biotechnology i.e., in customary nourishment and refreshment readiness, and also in current innovations in light of hereditary designing . A little extent of micro-organisms is pathogenic, which causes malady and moreover demise in creatures and plants.

We have focused on bacteria.

Microscopic organisms constitute an expansive space of prokaryotic microorganisms. Regularly a couple of micrometers long, microscopic organisms have various shapes, extending from circles to poles and spirals. Microbes were among the primary living things to show up on Earth, and are available in the greater part of its living spaces. Microbes possess water, soil and radioactive waste, and the profound segments of the outside layer of earth.

Micro-organisms likewise live on parasitic and cooperative associations with living creatures. Most of microbes are not described, and just a portion of the species of bacterial phyla that are developed in research facility. Branch of microbiology known to be bacteriology refers to the investigation of microscopic organisms. Normally there are approximately 40 million bacterial cells in one gram of soil and around million bacterial cells in one milliliter of fresh water. Around 5×10^{30} microscopic organisms survive on the Earth, framing the biomass that surpasses that of all the creatures and plants[2].

Microbes are indispensable in lots of phases of supplement cycle process by reusing supplements, for example, the nitrogen's obsession from environment. The supplement cycle incorporates deterioration of dead matter and microscopic organisms are in charge of the rot arrange in the procedure. In 2013, information announced by analysts in 2012, was distributed. They proposed that the microscopic organisms flourish in Mariana Trench with profundity of nearly 11 kilometers and are the most profound piece of the seas that are known[2].

The biggest number exists in gut greenery, also a vast of them on skin. Most by far of the microscopic organisms in our body are protected safe by defensive impacts of insusceptible framework, however some are advantageous especially in gut vegetation. Moreover a few types of microbes cause irresistible infections and are pathogenic, including syphilis, cholera, Bacillus anthracis, bubonic torment and other diseases.

The widely recognized lethal maladies are the respiratory diseases in which alone tuberculosis slaughters around two million individuals every year, generally in sub-Saharan Africa[3]. In created nations, anti-microbials are utilized to treat bacterial diseases and are likewise utilized as a part of cultivating, making anti-infection resistance a developing issue.

In industries, microorganisms plays a vital in the breakdown of oil slicks and sewage treatment, the generation of yogurt and cheddar through aging, and recuperation of palladium, gold and copper and different metals in mining area and in addition in biotechnology, and in the fabrication of anti-toxins and different chemicals.

In spite of the fact that the term microbes customarily incorporated all of the prokaryotes, logical order changed after revelation in 1990s that prokaryotes comprise of two altogether distinguishing gatherings of creatures which are developed from an antiquated normal precursor. These transformative areas are called Archaea and Bacteria[4].

Microscopic organisms show wide differing qualities of sizes and shapes which is called morphology. Cells of bacterium are around one-tenth of the extent of the eukaryotic cells which are normally 0.5–5.0 micrometers long. Among the littlest microbes are individuals from the sort Mycoplasma, which measure just 0.3 micrometers, as little as the biggest infections[5].

Most bacterial species are round, called cocci or pole molded, called bacilli. Some microscopic organisms, called vibrio, are formed like somewhat bended poles; others are winding molded, known as spirilla, or firmly snaked, known as spirochaetes. Few species even have tetrahedral or cuboidal shapes[6]. Many species exist just as single cells whike others relate in the trademark designs: Staphylococcus aggregate together in "bundle of grapes" groups, Streptococcus shape chains and Neisseria frame diploids (sets). Microscopic organisms are likewise lengthened to frame fibers, for example Actinobacteria.

Filamentous microbes are frequently encompassed by a sheath which contains numerous cells that are individual. Some sorts, for example, types of family Nocardia, frame mind boggling, extended fibers, comparable in appearance to parasitic mycelia[7].

In our work, we have classified *E.coli, Bacillus subtilis* and *Lactobacillus bacteria*.

1.1 *Escherichia coli*

It is a facultatively anaerobic, gram negative, bar molded bacterium of the sort Escherichia and is usually found in lower digestive tract of creatures that are warm blooded. Generally its strains are innocuous, however some of the serotypes causes genuine nourishment harming in hosts and are sporadically in charge of item reviews because of sustenance sullying. *E. coli* is ousted into nature inside fecal issue.

The bacterium develops hugely in crisp fecal issue under vigorous conditions. The bacterium can be developed and refined effortlessly and modestly in a research facility setting. It is considered a chemoheterotroph and its synthetically characterized medium incorporates a wellspring of energy and carbon. It is generally the most examined prokaryote living being, also a vital species in microbiology and biotechnology, where it fills as a host life form for a greater part with recombinant DNA[8].

It duplicated in approximately twenty minutes under ideal surroundings.

It stains Gram-negative due to the reason that its cell wall is made of a layer that is a thin peptidoglycan and also an external film. Amid this recoloring procedure, it stains pink and get a shade of counterstain safranin.

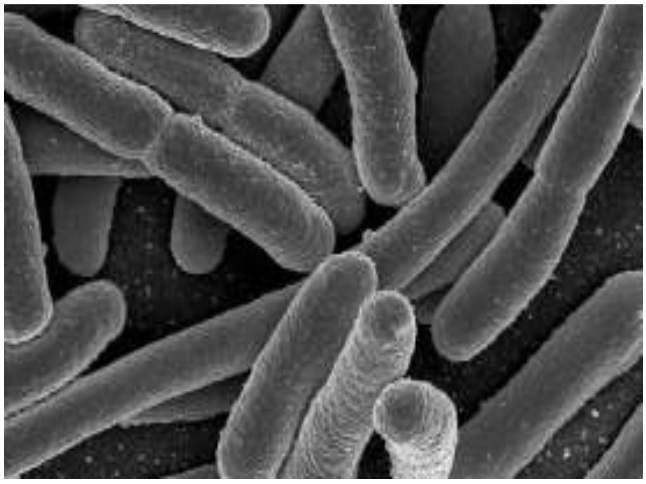

Fig. 1 Appearance of *e.coli*

Ref. http://columbiariverkeeper.org

It can live on wide assortment of substrates and also uses blended corrosive maturation in anaerobic surroundings, delivering ethanol, succinate, lactate, carbon dioxide and acetic acid derivation. Ideal development of E. coli happens at 37 °C, yet some lab strains can increase at temperatures up to 49 °C.

It keeps on growing without oxygen utilizing maturation or anaerobic breath. E. coli and related microbes have the capacity to exchange DNA by means of transduction or bacterial conjugation, enabling hereditary material to develop evenly by a current populace[9].

Generally *E. coli* strains don't cause illness, yet harmful strains can cause gastroenteritis, urinary tract contaminations, neonatal meningitis, hemorrhagic colitis, and Crohn's infection.

1.2 *Bacillus*

It is a variety of gram-positive, pole formed microorganisms and an individual from the Firmicutes phylum.

These can be oxygen dependent, or anaerobes that are facultative.
They test positive for the chemical catalase when there is the utilization of oxygen.
Bacillus incorporates both parasitic pathogenic species and non-parasitic as well.
In the upsetting natural surroundings, microscopic organisms create endospores that are oval and are not the spores that can be called as a genuine one, but rather to which microorganisms can decrease themselves and stay in the torpid state for a long stretch of time.
Numerous types of *Bacillus* can deliver overflowing measures of chemicals which are utilized as a part of various ventures.
Following species of *bacillus* are being considered restoratively critical: *B. anthracis* that causes Bacillus anthracis, and *B. cereus* that causes nourishment harming like that caused by Staphylococcus. The cell mass of *Bacillus* is a structure outwardly of cell and structures the boundary between the earth and bacterium, in meantime keeps up a pole shape and also withstands weight produced by cell's turgor pressure[10].

Bacillus subtilis has a rod shape and is a gram positive bacteria. It is mainly found in the soil. Its original name was "*Vibrio subtilis*". This bacteria forms endospore which helps it to tackle dry environments and extremely high temperatures. It is not toxic or pathogenic and does not cause any disease. It can handle high temperatures at the time of cooking as well.
It can work in the anaerobic environment also. It is readily found in plant compost soil and air. It occurs in spore and inactive form. It also helps in the production of many enzymes and those enzymes can degrade the plants as well. It is also found in the intestinal tract or on the skin of human body. It also produces subtilisin as a toxin when in association with enzymes which can cause allergy and this toxin can also be used as a detergent in laundry.
Amylases and proteases enzymes can be created by bacillus subtilis. It was earlier used as an anti-biotic also. It can be easily manipulated genetically and hence used as a model agent in research laboratories.
B. subtilis has demonstrated a significant model for look into. Different types of *Bacillus* are critical pathogens, causing Bacillus anthracis and nourishment harming. It is one of the best comprehended bacteria, as far as atomic and cell science. Its wonderful hereditary manageability and moderately substantial size have given the effective apparatuses required to explore a bacterium from every single conceivable angle.

1.3 *Lactobacillus*

It is a family of Gram-positive, facultative anaerobic or microaerophilic, bar molded, non-spore-shaping microscopic organisms. They are a noteworthy piece of the lactic acid microscopic organisms gathering (i.e. they change over sugars to lactic acid). In people, they constitute a noteworthy part of the microbiota at various body destinations.

Numerous lactobacilli work utilizing homofermentative digestion (they create just lactic acid from sugars), and a few species utilize heterofermentative digestion (they can deliver either liquor or lactic acid from sugars).

Numerous types of this variety don't require press for development and have a to a great degree high hydrogen peroxide resistance. The class *Lactobacillus* as of now contains more than 180 species and envelops a wide assortment of life forms. *Lactobacillus* species deliver hydrogen peroxide which restrains the development and destructiveness of the contagious pathogen Candida albicans in vitro and in vivo.

Some *Lactobacillus* species are utilized as starter societies in industry for controlled maturation in the creation of yogurt, cheddar, sauerkraut, pickles, brew, juice, kimchi, cocoa, kefir, and other aged nourishments, and additionally creature encourages. The antibacterial and antifungal action of *Lactobacillus* species depend on creation of bacteriocins and low atomic weight aggravates that hinders these microorganisms[11].

In numerous customary pickling forms, vegetables are submerged in saline solution, and salt-tolerant Lactobacillus species feast upon common sugars found in the vegetables. The subsequent blend of salt and lactic acid is a threatening situation for different organisms, for example, growths, and the vegetables are in this way protected.

Lactobacillus casei and *Lactobacillus sporogene*s are the two species of lactobacillus that were classified in our work.

1.4 Tulsi silver nano-particles

They were incorporated against different bacterial species under different concentrations of nano-particles of tulsi and at different pH conditions in order to study its anti-bacterial activity.

Silver nano-particles are of intrigue as a result of the one of a kind property such as attractive, optical and electrical properties that can be merged into biosensor materials, antimicrobial applications, cryogenic superconducting materials, composite filaments, electronic parts and corrective items.

A few physical techniques are utilized for the integration as well as settlement of silver nano-particles. One of the substances utilized as a part of nano-formulation is silver (nano-silver). Because of its antimicrobial properties of silver, it consolidates in channels for filtering drinking water and cleaning the water used in swimming pools.

To create nano-silver, metallic silver is moulded to ultra-fine particles through a few techniques; incorporate start releasing, electrochemical diminishment, arrangement light and cryo-substance blend.

Due to the non-lethal, safe inorganic antibacterial operator of silver nano-particles being utilized for quite a long time and is fit for killing around 650 microorganisms that reason ailments. Silver has been depicted as being 'oligo-dynamic', that is, its particles are fit for causing a bacteriostatic (development restraint) or even a bactericidal (antibacterial) affect. Thus, it can apply a bactericidal impact at minute fixation.

Plants give a superior stage to nano-particle blend as they are free from dangerous chemicals and also give normal topping operators. In addition, utilization of plant extricates likewise lessens the cost of miniaturized scale creature's separation and culture media upgrading the cost focused practicality over nano-particles blend by microorganisms[12].

Solutions of different pH were made for each bacterial species against varied concentrations of tulsi silver nano-particles and there absorbance was recorded using spectrophotometer.

This data was incorporated against the machine learning models for the prediction of the bacterial species.

1.5 Research Motivation

There are number of bacterial species present in the ecosystem. They have distinct properties, advantages and disadvantages. So, there classification is an important process in order to study their effect on the ecosystem. Knowing about the details of bacterial species can help us to utilize them in several ways such as to tackle against their pathogenic activity. Different species cause different types of diseases and at some places they can be harmful or can be helpful, so classification is must.

Classification is already done by scientists practically and new discoveries are on the verge.

It is time consuming and also there is use of the chemical species which may lead to toxic chemical waste. So, machine learning can be used to predict the bacterial class, however there is no on rack informational index accessible for foreseeing the class of bacteria.

1.6 Thesis Organization

This thesis is divided into 7 chapters and a shorthand review of each is given below:

- Chapter 1: This chapter discusses the basic information about the microorganisms and gives information about what bacteria is. It also defines the classes of bacteria that we have used in our thesis project. Also, it gives some information about the tulsi silver nano-particles and also the motivation behind the research.

- Chapter 2: This chapter shows some review of literature of available work that has been in the field of classification of bacteria. It also discusses the approaches that have been done in this field. Numerous machine learning approaches have also been discussed in this chapter.

- Chapter 3: The definition of the research problem is given in this chapter. Moreover the current gaps in this field of research are also being discussed and then the objectives of our thesis are also mentioned.

- Chapter 4: The datasets being used for simulation are being discussed in this chapter. It also discusses about the features of the dataset and also gives a sample dataset to get an overall idea of the simulation. It also shows the importance of each feature. The general information about features is also covered in this chapter.

- Chapter 5: In this chapter, methodology and models are discussed. Detailed information about the used methodology and general information about the four models that we used for the simulation are given in this chapter.

- Chapter 6: Results of the research are discussed in this chapter. All the four models are tested against the different parameters used in the methodology and are then compared against each other in order to show which one is the best model.

- Chapter 7: This section outlines the key discoveries and principle commitments of the postulation and records the conceivable future research bearings.

Chapter 2

Literature Survey

Hayes et al. [13] addressed the issue of precise DNA statistical modeling for bacterial species whose vast part was sequenced but not yet described tentatively. Accessibility of these models is basic for fruitful arrangement of the genome annotation task undertaking by statistical strategies for gene finding.

They presented a method known as GeneMark-Genesis in which the parameters of Markov models of non-coding and coding regions of protein from unspecified genomic sequence of bacteria. The diversity of composition of oligonucleotide is reflected by the diversity of models of protein coding.

Two gene models are described i.e. typical and atypical. They showed that the genes that escape the identification by typical model are predicted by the atypical method.

B. Slabbinck et al. [14] showed identification of bacterial species using large scale FAME-based machine learning methods. FAME stands for Fatty Acid Methyl Ester. It is a bacterial profiling and is used as a first-line identification method.

FAME profiles of the strains of *paenibacillus, pseudomonas and bacillus* were selected from the database. Many computational models are built for species and genus identification that are the applications of machine learning techniques.

They used three techniques i.e. SVM, Random Forest and Neural networks. Random Forest model was the best tested model with sensitivity values of 0.847, 0.901 and 0.708 for *bacillus, paenibacillus and pseudomonas* respectively.

R.M. Jarvis et al. [15] identified and characterized bacteria by using SERS.

Raman Spectroscopy is regarded as a vital fingerprinting technique that is utilized to identify, discriminate and characterize microorganisms and know how they react to biotic or abiotic stress. For the rapid bacterial analysis, the sensitivity of the Raman spectroscopy needs to be enhanced. They did Raman spectroscopy and also included SERS that provide rapid analysis of biological samples without the requirement of cell culture.

K De Bruyne et al. [16] demonstrated the identification of bacterial species using MALDI-TOF mass spectra through machine learning and data analysis. A basic protocol was built to create MALDI-TOF mass spectra that were obtained from a bunch of reference strains of *fructobacillus, leuconostoc* and *lactococcus*.

Bacterial cell were grown for a day. The set of binary character derived from the spectra and the spectra as well were used for the identification of the species. 84% of MALDI-TOF mass spectra was rightly predicted for *fructobacillus* and *leuconostoc* strains at the species level. Similarly, 94% of MALDI-TOF mass spectra was rightly predicted for *lactococcus* in which species and subspecies levels were considered. Random Forest and SVM gave accuracy between 94% and 98% for *leuconostoc* and *fructobacillus* respectively.

M.Trincavelli et al. [17] introduced a technique that can directly identify bacteria in the samples of blood culture by using electronic nose. Features are used in this method that capture the dynamic and static properties of signals from gas sensor array and propose a method to ensemble the outcome from consecutive samples. This mechanism is based on posterior probability that was extracted from SVM classifier. Ten distinguishing cultures of bacteria were used to validate the model. The proposed algorithm accounted for significant reliability for accuracy of classification.

2.1 Machine Learning Approach

Learning is only deducing information from the data we assembled in past. In people taking in begins from the moment we are conceived and this procedure of learning proceeds till the finish of life and in this term people attempt to assemble as much information as could reasonably be expected and afterward attempt to gain from that learning which was picked up from different encounters.

Artificial Intelligence (AI) tries to reenact procedure of realizing which occurs in humans as well as in other living things, in inert machine. Counterfeit consciousness empowers the machines to play out the undertaking with most astounding measure of exactness and additionally exactness given to them without requiring human obstruction.

Machine learning is a sub-field of AI and the primary territory in which machine learning works is to grow new calculations and in addition comprehend and assess calculations which empower the machine to learn. Nowadays in ventures, machine learning is a standout amongst the most mainstream territory of intrigue/work.

Machine learning tries to bring different fields like cerebrum modeling, human brain research and insights together to construct a shrewd framework. Neural systems which are roused from working of cerebrum are utilized generally in machine taking in for gaining from information.

Machine learning utilizes investigation of information and machine learning calculations too utilize examination aptitudes, in this way insight has a critical impact in machine learning. At the point when a PC is to settle or manage a specific errand is then that undertaking is known as assignment space or some of the time likewise alluded to as information base. Data that is created by or gotten from the assignment constitutes its information base.

To speak to learning base we utilize numerical, discrete esteem, social literals and Boolean or at times their mix is additionally utilized. Info yield sets are utilized to speak to learning base, here info given to errand is information and results which we get from that assignment are yield.

Information base's information can be utilized to characterize yield for given information. Information base is insufficient to know the interior working of an errand however, it is sufficient for arranging an offered contribution to some yield.

As when we have a considerable measure of data it is by inconceivable for people to get data from it, machine learning, then again can without much of a stretch do this. With the assistance of something beyond information a computational model is made which can speak to that errand with huge exactness without knowing interior working of that errand.

A calculation can utilize computational model to foresee yield for some unabsorbed contribution for that specific assignment. The computational model can be of any sort it can be basically a few principles, a recipe, or some numerical operations which when connected to input give a yield.

Each machine learning calculation utilizes diverse method to make computational display from information base however objective of each machine learning calculation is to construe information from learning base.

To find out about a procedure, machine learning calculations require dataset. Dataset have information with respect to, which yields was given for a specific info Each information has a few characteristics in it, which enlighten us regarding properties of that specific info, an information can have two quality or in some cases, there can be a huge number of such traits. A quality can either be consistent or discrete. Discrete characteristics as the name infers have unmistakable esteems, for example, state of a protest can either be square, rectangle and so forth; then again consistent have numeric esteems, for example, territory of shape. Each dataset has some info and some yield characteristic[18].

Information is fundamentally given to the learning calculation and the objective of learning calculation is to delineate offered contribution to the yield comparing to that specific assignment. It is expected in machine discovering that estimations of information and yield are reliant. Info traits given in dataset are known as elements in machine learning.

The computational model can likewise be thought as about a capacity which basically maps our contribution to a yield. Machine learning has numerous applications nowadays, for instance, we can prepare a computational model from messages with the end goal that calculation model can figure out how to recognize critical and spam email. Once a computational model is prepared, at that point that model can be utilized to keep vital mails in one envelope and spam on other.

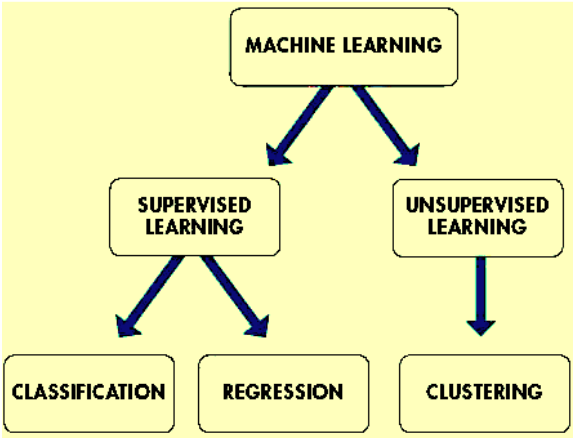

Fig. 2.1 Categorization of Machine Learning

Figure 2.1 demonstrates the order of machine learning approaches. It is isolated into two classes i.e. regulated learning and unsupervised learning. Administered learning is ordered into 1. Classification - foresee discrete esteemed yield 2. Regression - anticipate nonstop esteemed yield. Grouping is unsupervised learning. In managed taking in a computational model is prepared to delineate given in learning base to comparing yield.

In regulated learning, names are given along information, on other hand; in unsupervised learning we are not given marks alongside input. A case of regulated learning is to order whether an email is vital or not or for anticipating what number of runs a cricketer will make in his next match and cases on unsupervised learning incorporates grouping of archives.

Order calculations are those which are utilized to arrange in which class guaranteed input has a place with. Here classes are discrete in nature and some arrangement of standards or a model can be made to order an offered contribution to a specific class.

Information given to a characterization calculation can be either discrete or it can be persistent, however there are a few calculations which just take contribution to type of discrete properties yet yield of such calculations are dependably discrete.

Regression calculations make forecasts in light of some numerical operations or some condition makes a model which gives a nonstop incentive as yield while taking some information. Here information can either be ceaseless or discrete.

Table 2.1 Demonstration of previously designed regression and classification models and these models are accessible in R.

S.No.	Model Type	Model	Package	Method	Tuning Parameter
1	Classification	C5.0	C50	C5.0	Trials, winnow, model
2	Classification	GAMens	GAMens	GAMens	Iter, fusion, rsm_size
3	Classification	ada	Ada	ada	Iter, maxdepth, nu
4	Classification	J48	RWeka	J48	C
5	Classification	JRip	RWeka	JRip	NumOpt
6	Classification	Decision Tree	Rpart	rpart	Minsplit, maxdepth, minbucket
7	Classification	Random Forest	randomForest	rf	Sampling, mtry
8	Classification	SVM	e1071	svm	Nu, epilson
9	Classification	LM	Glm	lm	-
10	Regression	M5	RWeka	M5	Prooned, smoothed, rules
11	Regression	lm	Stats	lm	-
12	Regression	leapSeq	Leaps	leapSeq	Nvmax
13	Regression	lars	Lars	lars	Fraction
14	Regression	icr	Caret	icr	n.comp
15	Regression	foba	Foba	foba	Lambda, k
16	Regression	enet	Elasticnet	enet	Lambda, fraction
17	Regression	cubist	Cubist	cubist	Neighbors, committees
18	Dual use	bstTree	Bst	bstTree	Nu, maxdepth, mstop
19	Dual use	glm	Stats	glm	-
20	Dual use	knn	Caret	knn	K
21	Dual use	cforest	Party	cforest	Mtry
22	Dual use	avNNet	Caret	avNNet	Size, decay, bag
23	Dual use	gbm	Gbm	gbm	Shrinkage, trees, depth
24	Dual use	ctree	Party	ctree	mincriterion
25	Dual use	gcvEarth	Earth	gcvEarth	degree

Chapter 3

Problem Formulation

There are numerous sorts of microbes without which we couldn't live. They are significant to the nearness of life on earth. Survival of plants and hence the animals is made possible due to the presence of bacterial organisms. They fill a twofold need. In the main occasion they go about as scavengers expelling hurtful waste from the earth.

Also, they return it to the dirt as plant sustenance. Dead matter and squanders of living beings are deteriorated by the exercises of the saprophytic microscopic organisms. Bacteria are a necessity for agriculture, industries and medicine.

Scientists have classified them on the basis of their shape. These basic shapes are:

1. Rod
2. Spherical
3. Comma
4. Spiral
5. Corkscrew

The right distinguishing proof of microorganisms is of essential significance to microbial systematists and in addition to researchers required in numerous different zones of connected research and industry (e.g., clinical microbiology, horticulture and food development). Not all the time we can classify bacteria by examining their shape, size and other properties as they are needed to be seen under microscopic instruments as there might be unavailability of such instruments at certain times. Moreover it is difficult to classify them when a specific type of bacteria is mixed in a solution.

So, we tried to use machine approach to predict the class of bacteria. By using machine learning approach, we can save our time and it might also be cheap. So, methods need to be developed to categorize bacteria. Clustering of micro-organisms in groups with certain similarities is a crucial process. Proper organization of the microbial diversity is must.

3.1 Research Gaps

Following are the gaps that are recognized amid writing literature review:

1. Machine learning approach hasn't yet been used to find out the class of each bacteria.

2. Availability of the dataset that can be utilized for building up of base model is not there.

3. Development of the technique that can predict the class of bacteria with lesser amount of attributes is in need.

4. There is a need to study the spectrophotometric techniques in order to classify bacterial species.

5. Classification of bacteria is mainly based on their shape, size and some other chemical properties.

3.2 Research Objectives

Following are the research objectives:

1. Collection of data from the simulation and then use it for training the model.

2. Creation and analysis of different machine learning algorithms on acquired dataset and recognize best model which can classify bacteria most precisely.

3. To examine whether our preferred model is strong or not.

Chapter 4

Dataset and its features

Nowadays so as to comprehend complex biochemical procedures, modeling and simulation are being vastly used. Bacterial classification can possibly expand our comprehension of biochemical processes and metabolism.

4.1 Features and Dataset

Following features were noted down on checking the anti-bacterial activity of the silver nano-particles of tulsi on reacting them with the bacterial samples that we used i.e. *E.coli, Lactobacillus* and *Bacillus subtilis*:

1. Amount of Silver nano-particles of tulsi
2. pH 5
3. pH 7
4. pH 9
5. pH 11
6. pH 12
7. Absorbance

Solutions with varied concentrations of nano-particles of tulsi were made with the different bacterial samples. Concentration plays an important part in the reaction process as it will let us know how much amount is optimum for the anti-bacterial activity. Ten different concentrations of the solution were made.
pH value tells about the acidic, basic and neutral property of the given solution.
pH value ranging between 0 and 7 shows the acidic nature.
pH value ranging between 7 and 14 shows the basic nature.
pH at 7 is considered to be neutral. Solutions were tested at five different values of pH.
Absorbance too plays a vital role. It is the measure of substance's capacity to absorb light of a particular wavelength.

Some part of dataset is given in the following table.

Table 4.1 Sample dataset

Bacterial class	**Amount of silver NP**	**Abs. of pH 5**	**Abs. of pH 7**	**Abs. of pH 9**	**Abs. of pH 11**	**Abs. of pH 12**
		Wavelength = 600 nm				
Lactobacillus sporogenes	0.025	1.9434	1.8659	1.6198	1.7938	1.7856
Lactobacillus sporogenes	0.75	1.1769	0.9274	0.6328	0.5637	0.172
E.coli	0.025	1.8434	1.6659	1.4198	1.3938	1.4856
E.coli	0.75	1.0769	0.4274	0.4328	0.1637	0.172
Lactobacillus casei	0.025	1.4784	1.4225	1.3024	1.257	1.0028
Lactobacillus casei	0.75	0.5689	0.2822	0.0806	0.055	0.0149
Bacillus subtilis	0.025	1.3784	1.2225	1.024	0.5257	1.0051
Bacillus subtilis	0.75	0.6689	0.2822	0.1806	0.015	0.02849

4.2 Feature importance

Feature importance was carried out with the help of random forest model and the following values were obtained.

This method showed which feature is the most important one in determining the classification of bacterial species.

Table 4.2 Feature importance

Features	***Lactobacillus sporogenes***	***E.coli***	***Lactobacillus casei***	***Bacillus subtilis***	**Mean Decrease accuracy**	**Mean Decrease Gini**
Amount of silver NP	0.33	3.45	0.37	2.40	2.33	2.02
pH 5	4.74	5.15	-3.31	8.25	8.64	2.80
pH 7	0.51	-0.32	-3.30	7.68	5.07	2.37
pH 9	-2.05	0.26	-1.40	1.86	-0.71	1.67
pH 11	0.28	0.24	-0.91	7.11	4.05	2.21
pH 12	0.42	1.95	0.31	1.31	1.26	1.91

The above table shows that pH 5 is the most important feature due to the reason that it has the highest Mean Decrease Accuracy and Mean Decrease Gini values.

Chapter 5

Methodology and Models

5.1 Approach

In fig 5.1, methodology has been shown. We take all the required data i.e. absorbance readings corresponding to each pH and concentration value and arrange them in an ordered manner in an excel sheet and then simulate them in R to classify the bacterial species.

A separate dataset was made in which the data of each species was combined and then made ready for R simulation.

After the cleansing of data, we trained our models with 70 % data and the remaining was used as test data to verify our results.

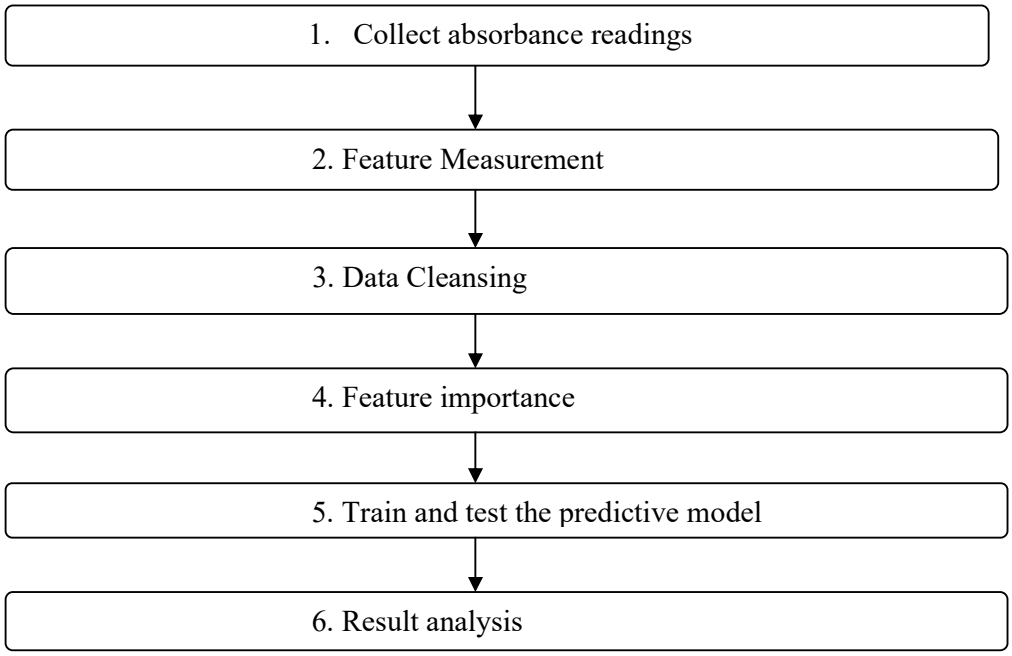

Fig. 5.1 Methodology used

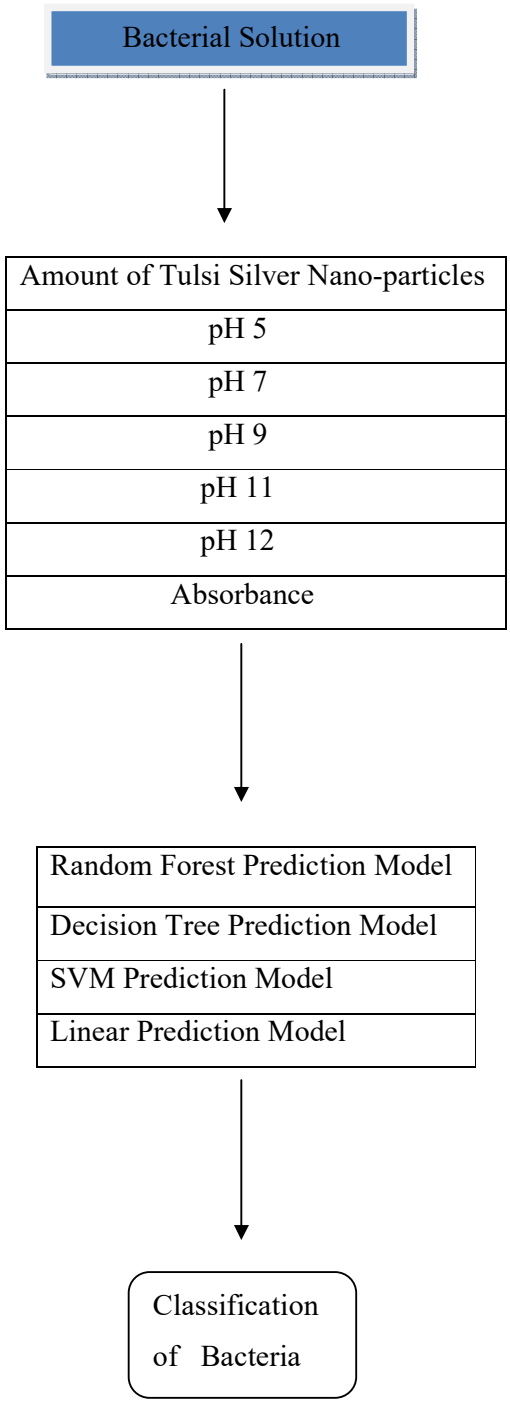

Fig. 5.2 Method of Prediction

Figure 5.2 describes the prediction method.

5.2 Model Evaluation

We can measure the performance of prediction by number of ways. A few strategies perform better in some application, while others perform better in some other application. Therefore, contingent upon the application, we pick technique to assess our prediction. A concise exchange on some of strategies which are utilized to assess execution of prediction is given in following subsections.

Our work uses specificity, sensitivity, Positive Predictive Value (PPV), Area under Curve (AUC), Prevalence, Negative Predictive Value (NPV), Detection Prevalence and Detection Rate in building the performance score.

We have used four machine learning models for the classification of bacterial classes and those models are Decision tree, Random forest, linear model and SVM.

It is a kind of table which describes the performance of a model used for classification on a set of test data whose true values are already known.

Table 5.2 describes the confusion matrix. All the performance parameters i.e. specificity, sensitivity, Positive Predictive Value (PPV), Area under Curve (AUC), Prevalence, Negative Predictive Value (NPV), Detection Prevalence and Detection Rate can be explained using the confusion matrix.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table 5.2 Representation of Confusion Matrix

5.2.1 Sensitivity

It is also known as true positive rate.

It measures the proportion of positives which are identified correctly.

$$\boldsymbol{Sensitivity} = \frac{TP}{TP + FN}$$

5.2.2 Specificity

It is also known as true negative rate.

It measures the proportion of negatives which are identified correctly.

$$\boldsymbol{Specificity} = \frac{TN}{TN + FP}$$

5.2.3 Positive Predictive Value (PPV)

It is the proportion of predicted positives that are actual positives.

It shows the probability that predicted positive is true positive.

$$\boldsymbol{PPV} = \frac{TP}{TP + FP}$$

5.2.4 Negative Predictive Value (NPV)

It is the proportion of predicted negatives that are actual negatives.

It shows the probability that predicted negative is true negative.

$$\boldsymbol{NPV} = \frac{TN}{TN + FN}$$

5.2.5 Detection rate

It is the ratio of the total number of actual cases to the total number of predicted cases.

$$\boldsymbol{Detection\ rate} = \frac{TP}{TN + FN + TP + FP}$$

5.2.6 Prevalence and Detection prevalence

$$\boldsymbol{Prevalence} = \frac{TP + FN}{TN + FN + TP + FP}$$

$$\boldsymbol{Detection\ prevalence} = \frac{TP + FP}{TN + FN + TP + FP}$$

5.2.7 Area under Curve (AUC)

It gives the analysis of the models that we used for the prediction and shows which one is the best.

$$\boldsymbol{AUC} = \frac{Sensitivity + Specificity}{2}$$

5.2.8 k fold cross validation

k fold cross validation is used to verify how robust our model is. In this system, dataset is divided into k parallel estimated tests and out of these k tests, k-1 tests are utilized for preparing and remaining 1 test is utilized for testing our outcome and this procedure is rerun k times (the folds) with the end goal that every last k test is utilized for testing once.
The k comes about because of the folds at that point can be arrived at the midpoint of (or generally joined) to deliver a solitary estimation. The upper hand of this strategy over rerun irregular sub-examining is that all perceptions are utilized for both preparing and approval, and every perception is utilized for approval precisely once[19].

5.3 Models

5.3.1 Decision tree

Tree based learning calculations are thought to be one of the best and for the most utilized regulated learning strategies. Tree based strategies engage predictive models with high precision, strength and simplicity of elucidation. Dissimilar to direct models, they delineate straight connections greatly. They are versatile at taking care of any sort of issue within reach (regression or classification).
It is a sort of regulated learning calculation (having a pre-characterized target variable) that is generally utilized as a part of classification issues. Continuous and categorical input and output variables are being worked by this model. Sample or population is being divided into two or more homogeneous sub-populations that are based on the most important differentiator or splitter in input variables[20].
Output of this method can be easily understood. Even the people having non-analytical background can use it. No requirement of statistical knowledge for interpreting and reading it. Its graphical portrayal is extremely natural and users without much of a stretch can relate their speculation. For identifying variables that are most significant, this method is one of the fastest and also it shows the relationship between different variables. It is also useful in the stage of data exploration. Also, very less cleansing of data is required.

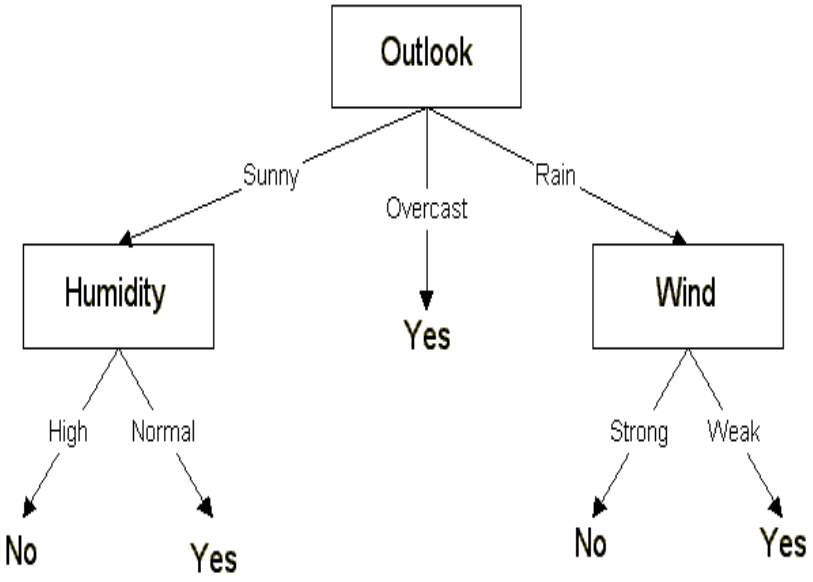

Fig. 5.3.1 Representation of decision tree model

5.3.2 Random forest

Random Forest is a standout amongst the most famous machine learning calculations which was made by Breiman. It is only a gathering of numerous straightforward decision trees and every one of these trees can anticipate the result for any input. These trees can anticipate to which class a specific input has a place, if our issue is of classification, and if issue is of regression, these trees can anticipate a continuous number.

In the classification event each tree in random forest votes in favor of a specific class and the class which has greater votes is given as result for that specific input, then again, in regression yield of each tree is found to be the middle value to get the output for that specific input.

Random Forest can be viewed as group of numerous basic decision trees. Ensembling of numerous decision trees in Random Forest has demonstrated surprising change in execution of model.

Random Forest is likewise ready to beat the issue of over-fitting, which is one of most concerning issue in single decision tree. Amid model training, every decision tree in the model is prepared on arbitrary subset of components of training data.

Another process of ensembling , bagging, chooses random sub-tests of the training data and trains the model on them yet random forest is unique in relation to stowing as here we are picking random examples of training data as well as we are picking random specimen of the features too[21].

Random forest with help of different decision tree are a great deal and are more summed up when contrasted with single decision tree as there is less possibility of over-fitting. Random Forest can likewise be utilized to rank elements. Feature determination utilizing random forest was given in the original paper of random forest.

Fig. 5.3.2 Representation of random forest model

5.3.3 SVM

Right now SVM (Support Vector Machines) is a hot topic in machine learning. SVM shows a capable method for regression, general (nonlinear) classification and outlier detection with a natural model representation. It was basically developed for the classification of binary elements.

In flexible modeling, it is a popular technique and provides an easy to use interface. SVM might be extremely sensible to the best possible selection of parameters, so scope of parameter combinations are needed to be checked, in any event on a sensible subset of your information. Due to better performance and less parameters, C-classification was mainly used for carrying out classification processes[22].

In order to get good results, grid search over all parameters is used. Large datasets may increase the training times. Scaling of data is needed to be done. Class weighting can also be done using this model. In the event that one wishes to weight the classes in a distinguishing way, weights might be determined in a vector with named parts.

Cross-classification can also be done using this model. To evaluate the nature of the training result, we can do a k-fold cross-arrangement on the training data by setting the parameter cross to k, the SVM model will then have some extra values, contingent upon whether regression or classification is carried out.

Fundamentally, svm can just tackle binary classification issues. To take into consideration multi-class classification, this model utilizes the one-against-one method by fitting all binary sub-classifiers and finding the right class by a voting process.

5.3.4 Linear Model

Linear model uses a single autonomous variable to foresee the result of a dependent variable. By knowing this, the most essential type of regression, various complex modeling procedures can be learnt. In order to model binary data, this model is mostly used.

Using glm function, generalized linear models are fitted in R.

Poisson and Logistic regression both belong to the generalized linear model family.

These models are expansions of customary regression models that enable the mean to rely upon the explanatory variables through a link function, and the reaction variable to be any individual from an arrangement of circulations called the exponential family[23].

We can utilize the glm() function to work with GLMs in R. It's utilization is like that of the lm() function which we beforehand utilized for multiple linear regression. The principle distinction is that we have to incorporate an extra argument family to depict the mistake and link function to be utilized as a part of the model.

Chapter 6

Results and Discussion

In this section, values of each parameter that we used for classification process is given with the four models that we used in R.

Table 6.1 Value of parameters using Decision Tree Model

Parameters	**Decision Tree values**
Sensitivity	0.5000
Specificity	0.9091
Positive Predictive Value	0.8333
Negative Predictive Value	0.6667
Prevalence	0.4762
Detection Rate	0.2381
Detection Prevalence	0.2857
Area under Curve	0.7045

Table 6.2 Value of parameters using Random Forest Model

Parameters	**Random Forest values**
Sensitivity	1.00000
Specificity	0.85000
Positive Predictive Value	0.25000
Negative Predictive Value	1.00000
Prevalence	0.04762
Detection Rate	0.04762
Detection Prevalence	0.19048
Area under Curve	0.92500

Table 6.3 Value of parameters using SVM Model

Parameters	SVM values
Sensitivity	1.00000
Specificity	0.75000
Positive Predictive Value	0.16667
Negative Predictive Value	1.00000
Prevalence	0.04762
Detection Rate	0.04762
Detection Prevalence	0.28571
Area under Curve	0.87500

Table 6.4 Value of parameters using Linear Model

Parameters	Linear model values
Sensitivity	0.5714
Specificity	0.8571
Positive Predictive Value	0.6667
Negative Predictive Value	0.8000
Prevalence	0.3333
Detection Rate	0.1905
Detection Prevalence	0.2857
Area under Curve	0.7143

Cross validation of the parameters was done for the Random Forest model and the graphs were plotted.

The graphs were plotted between the parameter and the number of times that model was run.

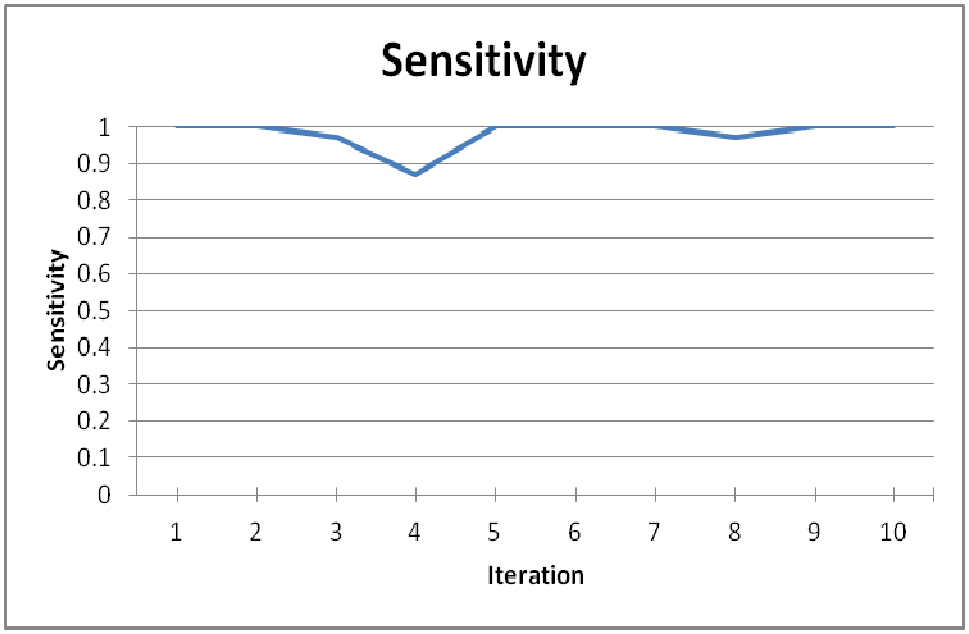

Fig. 6.1 Cross validation of sensitivity parameter using Random Forest model

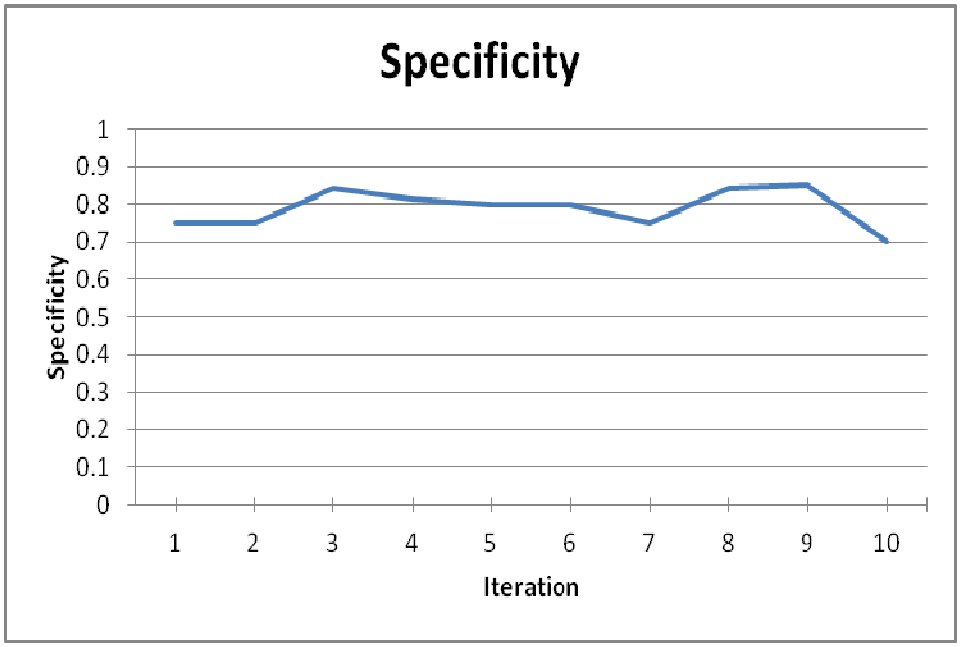

Fig. 6.2 Cross validation of specificity parameter using Random Forest model

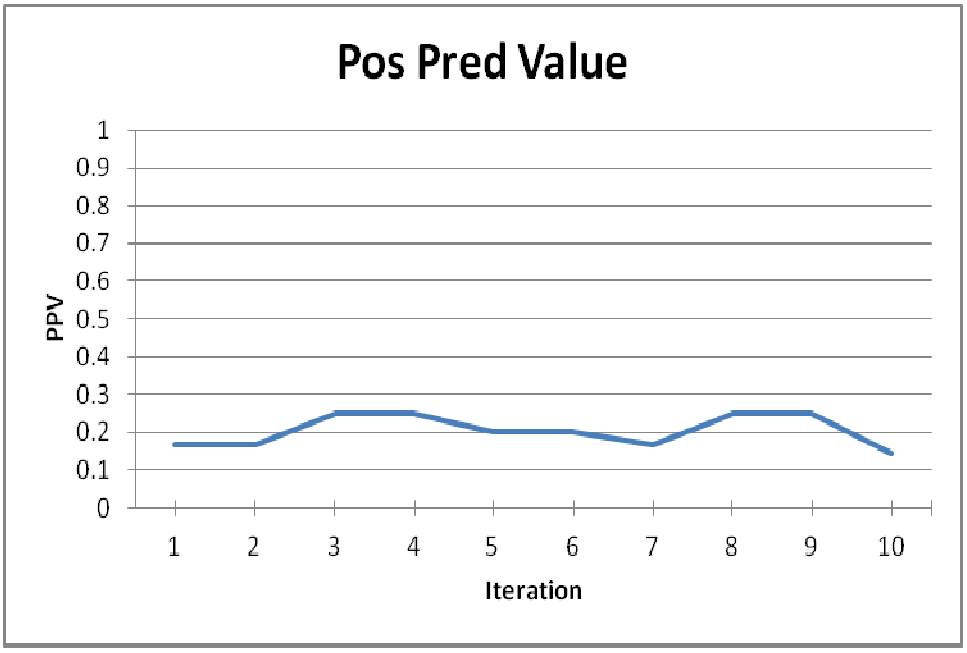

Fig. 6.3 Cross validation of Positive Predictive Value parameter using Random Forest model

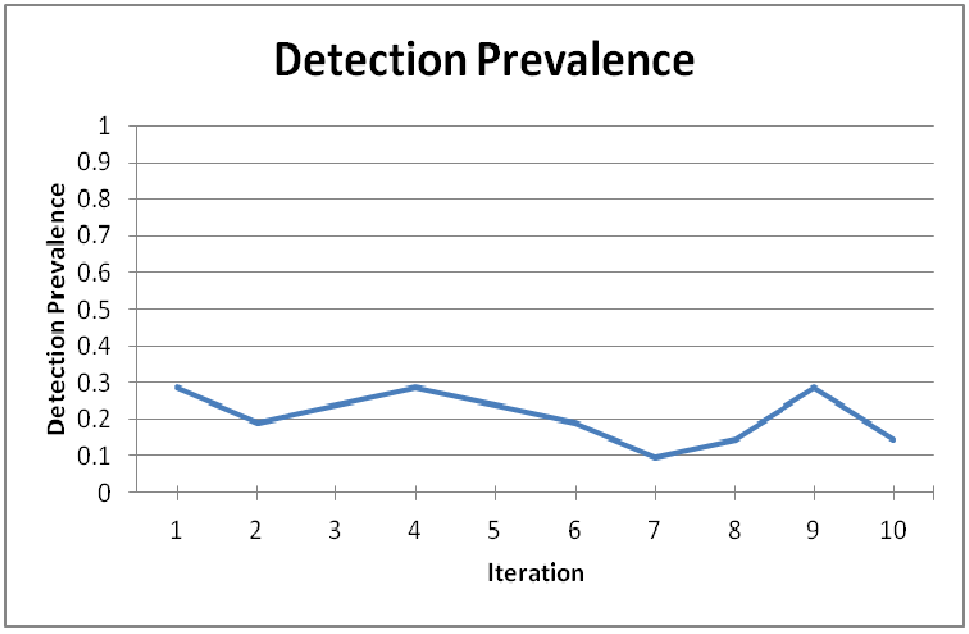

Fig. 6.4 Cross validation of Detection Prevalence parameter using Random Forest model

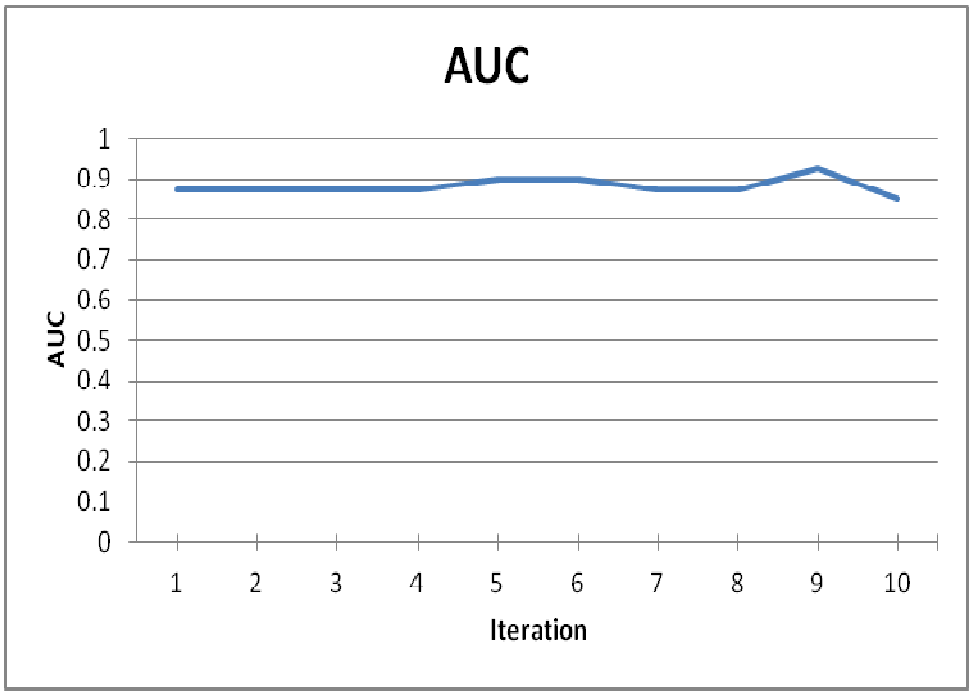

Fig. 6.5 Cross validation of Area under Curve parameter using Random Forest model

The above results show that Random Forest model is the best model that can be used for the prediction of the bacterial classes as this model has the highest value of AUC i.e. 0.92500.

AUC accounts for the best accuracy so after Random Forest model, SVM model (AUC = 0.87500) is the best suited for the classification process followed by linear model (AUC = 0.7143) and Decision Tree model (AUC = 0.7045).

Moreover, the cross validation also shows that Random Forest model is the most favorable model to carry out the classification as the model was run ten times for different parameters and showed the best consistency comparing to other models.

Chapter 7

Conclusion and Future scope

7.1 Conclusion

We developed a machine learning approach for the prediction of different classes of bacteria. Depicting the bacterial classes from different solutions can be time consuming as well as costly. So, our proposed method can be utilized as it can solve the above problems.

This thesis is restricted to only the four types of bacteria on which prediction is done but research can be extended to several other classes of bacteria as well which will be very helpful in future. Number of bacterial samples can be taken and put in different types of solutions and then a machine learning model can be built.

It is not easy all the time to predict what kind of bacteria is present in the given solution as number of chemical processes need to be repeated for that purpose and it will take time. So, in order to save the wastage of chemicals and time, machine learning methods can be built and in no time we will get our results.

7.2 Thesis Contributions

1. For the prediction of bacterial classes, a machine learning technique is proposed.

2. We used R for the simulation process of the features for the generation of dataset.

3. In order to check the robustness of the models used, k cross validation was utilized.

4. We did an investigation of models on acquired dataset and attempted to discover which one is the better for the prediction of bacterial classes.

7.3 Future Scope

1. Four machine learning models are used for the prediction of bacterial classes in our thesis; there are several other machine learning methods that are needed to be explored and can be used for fast and accurate predictions.

2. More and more bacterial species under different environments such as different pH values, temperature, medicinal concentrations and other solutions can be classified using machine learning models.

References

[1] E. Rybicki, "The classification of organisms at the edge of life or problems with virus systematics," *South African Journal of Science,* vol. 86, pp. 182-186, 1990.

[2] K. Dose, A. Bieger-Dose, R. Dillmann, M. Gill, O. Kerz, A. Klein*, et al.*, "ERA-experiment "space biochemistry"," *Advances in Space Research,* vol. 16, pp. 119-129, 1995.

[3] J. K. Fredrickson, J. M. Zachara, D. L. Balkwill, D. Kennedy, W. L. Shu-mei, H. M. Kostandarithes*, et al.*, "Geomicrobiology of high-level nuclear waste-contaminated vadose sediments at the Hanford Site, Washington State," *Applied and environmental microbiology,* vol. 70, pp. 4230-4241, 2004.

[4] M. S. Rappé and S. J. Giovannoni, "The uncultured microbial majority," *Annual Reviews in Microbiology,* vol. 57, pp. 369-394, 2003.

[5] R. N. Glud, F. Wenzhöfer, M. Middelboe, K. Oguri, R. Turnewitsch, D. E. Canfield*, et al.*, "High rates of microbial carbon turnover in sediments in the deepest oceanic trench on Earth," *Nature Geoscience,* vol. 6, pp. 284-288, 2013.

[6] C. L. Sears, "A dynamic partnership: celebrating our gut flora," *Anaerobe,* vol. 11, pp. 247-251, 2005.

[7] T. Ishige, K. Honda, and S. Shimizu, "Whole organism biocatalysis," *Current opinion in chemical biology,* vol. 9, pp. 174-180, 2005.

[8] O. Tenaillon, D. Skurnik, B. Picard, and E. Denamur, "The population genetics of commensal Escherichia coli," *Nature reviews. Microbiology,* vol. 8, p. 207, 2010.

[9] J. B. Russell and G. N. Jarvis, "Practical mechanisms for interrupting the oral-fecal lifecycle of Escherichia coli," *Journal of Molecular Microbiology and Biotechnology,* vol. 3, pp. 265-272, 2001.

[10] E. N. Nazar, "Effect of Some Physical Factors on Natural Biosynthesis by Streptomyces spp," *British Journal of Applied Science & Technology,* vol. 4, p. 2762, 2014.

[11] A. R. Varlan, W. Sansen, A. Van Loey, and M. Hendrickx, "Covalent enzyme immobilization on paramagnetic polyacrolein beads," *Biosensors and Bioelectronics,* vol. 11, pp. 443-448, 1996.

[12] X. Chen and H. Schluesener, "Nanosilver: a nanoproduct in medical application," *Toxicology letters,* vol. 176, pp. 1-12, 2008.

[13] W. S. Hayes and M. Borodovsky, "How to interpret an anonymous bacterial genome: machine learning approach to gene identification," *Genome research,* vol. 8, pp. 1154-1171, 1998.

[14] B. Slabbinck, B. De Baets, P. Dawyndt, and P. De Vos, "Towards large-scale FAME-based bacterial species identification using machine learning techniques," *Systematic and applied microbiology,* vol. 32, pp. 163-176, 2009.

[15] R. M. Jarvis and R. Goodacre, "Characterisation and identification of bacteria using SERS," *Chemical Society Reviews,* vol. 37, pp. 931-936, 2008.

[16] K. De Bruyne, B. Slabbinck, W. Waegeman, P. Vauterin, B. De Baets, and P. Vandamme, "Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning," *Systematic and applied microbiology,* vol. 34, pp. 20-29, 2011.

[17] M. Trincavelli, S. Coradeschi, A. Loutfi, B. Soderquist, and P. Thunberg, "Direct identification of bacteria in blood culture samples using an electronic nose," *IEEE Transactions on Biomedical Engineering,* vol. 57, pp. 2884-2890, 2010.
[18] P. Pantola, A. Bala, and P. S. Rana, "Consensus based ensemble model for spam detection," in *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, 2015, pp. 1724-1727.
[19] P. S. Rana, H. Sharma, M. Bhattacharya, and A. Shukla, "Quality assessment of modeled protein structure using physicochemical properties," *Journal of bioinformatics and computational biology,* vol. 13, p. 1550005, 2015.
[20] J. R. Quinlan, "Induction of decision trees," *Machine learning,* vol. 1, pp. 81-106, 1986.
[21] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news,* vol. 2, pp. 18-22, 2002.
[22] S. S. Keerthi and E. G. Gilbert, "Convergence of a generalized SMO algorithm for SVM classifier design," *Machine Learning,* vol. 46, pp. 351-360, 2002.
[23] J. M. J. M. Chambers, "Computational methods for data analysis," 1977.

Prediction M.Sc Biochemistry - Chirag Sharma

ORIGINALITY REPORT

%8	%6	%4	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	en.wikipedia.org Internet Source	%1
2	genome.cshlp.org Internet Source	%1
3	Sinikallio, Sanna, Timo Aalto, Olavi Airaksinen, Arto Herno, Heikki Kröger, Sakari Savolainen, Veli Turunen, and Heimo Viinamäki. "Depression and associated factors in patients with lumbar spinal stenosis", Disability and Rehabilitation, 2006. Publication	<%1
4	dspace.thapar.edu:8080 Internet Source	<%1
5	Overby, Casey L.. "Predictive Medicine", Encyclopedia of Systems Biology, 2013. Publication	<%1
6	ira.lib.polyu.edu.hk Internet Source	<%1
7	www.readbag.com	

17/7/2017

Chirag