Comprehensive Molecular Diagnostics and Advanced Gene Expression Analysis

Prof. Arindam Ghosh

Dr. B.C. Roy Multi-Speciality Medical Research Centre

Indian Institute of Technology Kharagpur

Lecture 59 : Artificial Intelligence and Machine learning in Genomics

. Hello everyone. Welcome back to lecture series on Comprehensive Molecular Diagnostics and Advanced Gene Expression Analysis. We are in the penultimate lecture. So, it is lecture 59 and we will be discussing the role of artificial intelligence and machine learning that is AI and ML in genetics right. So, it is a definitely futuristic topic the module deals with futuristic trends.

So, we will be discussing the these concepts, we will be giving an overview regarding big data and genomical data right. We will be introducing the concept of machine learning, we will be analyzing the various subtypes how it is useful in genomics, we will be giving a brief overview of R programming, what are its features, what is the exact schematic to develop a machine learning model in any health research informatics study. We will be over viewing artificial intelligence as well as neural networking as well as the applications of AI and neural networks in genetics. So, lot to learn it might be new to if you do not have a computer science background, but you know the domain of informatics and bioinformatics is actually interfacing the biology as well as computer science into a coherent platform integrated platform.

So, it is very interesting for those who want to develop or want to take up these projects and so let us learn. First the concept of big data you know big data lot of data simple. So, how lot large data how large it is not only large, but it is also very complex. Now this data lots of data very complex data when do we call it a big data it is so complex that it cannot be traditionally analyzed by conventional tools. For example, most of us uses Microsoft excel or any statistical program for example, SPSS to analyze data, but it is so much that is very near impossible to manually input all of them and to analyze.

So, traditional in tools for data processing cannot be used for these type of data right. Now these type of data massive in volume can be multiple types can be structured properly labeled or semi structured or totally unstructured unlabeled data right depending on the sources. So, what are the common sources of big data very important social media, twitter, youtube lot of data is being generated every day multiple sensors right.

For example, you do not know if you are wearing a smartwatch right now the simple smartwatch sensor generating so many data based on your position sense what notification is coming so on and so forth right. And most important a business transaction definitely lot of trading everything is happening all over the world so many data are being generated every day and our point of interest scientific research lot and lot of data are being generated every day every hour globally right.

So, if you just look at the I mean features or characteristic of big data generally they are denoted by three V's. So, what are the three V's from the definition itself or from the idea itself it is must be very clear to you by now that it is big in volume the first V is volume big data a lot of data total right voluminous data velocity it not only large amount of data it is being generated in a very rapid pace it is being generated so much speed that normal traditional tools cannot keep up when analyzing all of that. And lastly variety there is no single type of data multiple type of data structured semi structured unstructured so many things and here are multiple example for example, structured data databases right there can be many such databases which needs to be analyzed at a time semi structured data for example, the data that we input into those databases for example, XML files it is a file format system of computer science and unstructured data for example, text document images videos specially in health care delivery system. So, a lot of things so when so many of these things are being generated at a very rapid pace to deal such a volume of data we call it as big data. Now regarding genetics that is our area of interest this is just a pictorial illustration to remind you what all techniques we have learnt in this entire lecture series and even one single of them if it is being run daily at a clinical diagnostic lab it will be generating lots and lots of data all right there can be anything will be.

So, there is proteomics, genomics, transcriptomics, metabolomics, pharmacogenomics. So, everything ultimately needs to be integrated at one front which is our topic of next lecture right, but here we are getting into that topic how to analyze and integrate them. So, considering we are going into one key study just one example of it not key study for example, this cancer genome atlas program again done by national institute of health NIH America right. So, they have done a study or they are actually the study is actually going on. So, multiple data types across 11000 patients across 33 different tumor types right.

So, at this point till this point there are multiple attributes for example, more than 2000 plus categorical attributes of data and it has generated more than 2.5 petabytes of data all right. So, if you do not have any most of you might be having the idea about the unit of data right. So, if you do not let us give you an idea first look at this do not look at the picture over here right. So, 1 GB you already know gigabyte right back 10 years back our ram used to be teen megabyte right 128 MB ram 256 MB ram 512 MB megabyte right.

So, that MB has been conveniently replaced by gigabyte right now even our phones have so many gigabytes and we are casually encroaching the domain of terabytes. So, terabyte is 1000 gigabyte right and 1000 terabyte makes 1 petabyte and very soon it will be again exhausted. So, 101024 petabyte makes 1 exabyte and 1000 exabyte makes 1 zettabyte right. So, just for those who are interested in unit of data here is the unit for you. Now over I mean in traditional English literature we use the term astronomical astronomical proportion right astronomical data you have astronomical volume of work right astronomical responsibility why do we use the term astronomical the word astronomical means huge it signifies huge because astronomical data is accused right you see the here is a comparative chart between the data that is generated by astronomy some social media and genomics right just to put things in perspective the speed of acquisition in astronomy is very fast we have most high tech equipment from all the space program for example, NASA for America right.

So, 25 zettabytes per year of data is I mean taken up right you know already data acquisition method in genomics right anyway. So, one storage 1 exabyte per year right how they are analyzed multiple process of analysis is done is I mean shown over here right and depending on the distribution we already know for example, dedicated lines are there how that is circulator of all across the globe. So, 600 terabyte per second line very high speed line only reserved for astronomical purpose, but I mean compared to that twitter 0.5 to 15 billion tweets per year, but the amount of data that is stored is generally 1 to 17 petabyte per year very less because it is text files only might be images. If you just compare it to youtube its huge 1 to 2 petabyte 1 to 2 exabyte per year because there are video files right 4 k videos 8 k videos now very huge right.

We are not looking to distribution we just looking at the volume of work and lastly if you compare genomics right the acquisition rate is 1 zettabases every year right and the storage is 2 to 40 eta bytes right per year. So, just to put things in perspective you can see it actually puts astronomical data into same even youtube data is much more than astronomical data right. So, very soon you might expect English literature to use the term genomics data in I mean replacing astronomical data it might. So, happen and it might happen very soon when everyone realizes how huge it is right. So, this is just a comparative graph of the same study that was been done and the units are also being shown on multiple stage of I mean how the volume of data increased manifold and what is the storage capacity over the years right.

How genomics has gone from miniscule data when it was under domain of human genome project early human genome project this is has increased exponentially right. So, genomical data is a big data there is no denying that what are the sources of big data multi dimensional data set analysis from various tissues databases from multiple cell analysis functional annotation one the once the whole pathway analysis done multiple functional annotation is a branch of bio informatics where multiple functions are annotated to various pathways that are being analyzed again that generates a huge amount of data right. So, the thing is we need all this data to the final goal is this right to improve disease prevention diagnosis prognosis and treatment efficacy is the blanket goal for which we are actually learning this course with the improve health right. So, we have lots of data we have complex problems with this data we want to make and what do we do with this data we need to make we need to be in such a position that we can use this data to make a better future. How can we make better future we can analyze those data and make data driven decision and we can predict what might happen in future right.

So, you guess the answer whenever there is complex problem and there is a big data that is available to read to analyze the answer to all of that in today's data is machine learning. So, what is machine learning? So, that is how we get into machine learning right. Now, machine learning is a basically it is a data analysis what can I say method right that automates right analytical model building right. So, we are I mean reading this data or analyzing this data we are predicting, but it is done on it is own at some point of time right. So, machine is learning ok.

So, what does it do it helps us to make data driven prediction not only that with such huge amount of data it can decipher patterns. So, this appears similar to this right. So, identification of patterns among those huge astronomical sea of data again very useful property of machine learning without explicit human intervention right active human intervention is not needed in an ML model once it is properly developed right. And therefore, definitely it is helping in analysis of complex data and big data just to put things into perspective a very lucid explanation can be. So, in traditional programming what do we do we have got the components data we have got data we have got the program and we have got the computer that is a calculator if you can just visualize calculate as a computer.

So, what do it in traditional programming we put those data we mention the program what do we want to use and the computer gives you the answer. So, we are inputting 2 plus 2 we are inputting plus and the computer gives 5 right. Machine learning is the approach opposite to that we have got all the programs we have got data we have got the program and we have got the output right. So, we tell the machine that there are this data and I have got the final result 5 how did we reach it. So, now, the machine thinks and answers you oh you need to input the plus program right.

So, you see there is a intelligence involved this is a very early very very primitive illustration of machine learning, but I hope you can appreciate the process. Now, if we got to understand machine learning in detail a bit detail not full detail is not possible in this course in this one lecture. We need to know about supervised and unsupervised learning in general then we get into machine learning. So, supervised learning a teacher tells you a kid what to learn and unsupervised learning adult learning we have got multiple reference text book we learn there is no limit we learn and then we think and we can categorize the data based on our understanding right. Machine learning also works in a similar way.

So, there are supervised learning there is are unsupervised learning models and programs right and under supervised learning I am going to explain all of them in brief. Classification and regression models typically fall under supervised learning and in case of unsupervised learning where the data is not labeled right. Supervised learning means we give the machine labeled data and the machine does with those data and in unsupervised learning we let the machine learn and sum we input unlabeled data and the machine done thus clustering and association. So, what are these terminologies right? So, to make it easier for you see again the brief classification how we get there here we used labeled data in supervised learning very important and in unsupervised learning we used unlabeled data very very very important multiple choice question related to ML. Unsupervised learning unlabeled data right data which is not labeled and supervised learning labeled data.

So, if we have got labeled data the machine knows already we tell the data this data belongs to this subtype or this data is associated with this attribute there is some label then there are multiple applications and algorithms by which machine learning can analyze those data. Number one is classification so, what does classification do in ML model? Assigning test data into specific categories based on features such as distinguished spam I am giving the general example right not healthcare related example. So, spam from legitimate emails right. So, example of them are linear classified decision tree these are all designs or these are all specific models that uses classic that fall under classification category of machine learning. So, you got the point a very important feature often the your Gmail right if you are using Gmail will often learn or even what calls are supposed to go I mean what emails are supposed to go in your spam folder even for those who are using mobile handsets you know in India there is an application that is true caller right and it can actually filter out it can show you what are the spam calls in red color what are the important calls in blue color right.

So, you got the idea now regression is another method another approach of dealing labeled data all right where that is to understand the relationship between the dependent and independent variable. So, we give label data and the machine finds out relationship all right. So, independent dependent variables you already you should be knowing that. So, for example, to predict numerical variable like forecasting sales revenue methods such as using polynomial or linear regression mathematics definitely is a lot of mathematics involved in machine learning, but we will get there we will get how to get into genomics by just understanding the simple facts about basic facts about ML machine learning. So, clustering in clustering what happens now we fall into this unsupervised learning unlabeled data.

Now the we machine actually groups unlabeled data based on similarities or differences because the machine finds out now the we are. So, mind it unsupervised learning is when the machine has gained some level of intelligence right there are stages how we do it for example, algorithm like k means clustering right. Now this type of algorithm finds association, but how clustering is different from association right clustering is similar to classification when the machine groups or categorize or classifies unlabeled data right. And just like regression where we are giving labeled data and machine was finding relationship association is the terminology that is used when a machine associates relationships between various variables of a data set by acquiring unlabeled data. So, it is one of the highest form of machine learning model that can be achieved right.

One example is recommendation right I will tell you what do you mean by customers who bought this item also bought. So, once we look into the real world application of machine learning we can understand I mean all of these are used sequentially to develop final programs and if we look into it everything around us is actually we are immersed in AI and ML. For example, unified payment interface UPI I mean most of you might be using those who are watching for online payment right. So, fraud detection very important in UPI ML is integrated. Again when you are have you seen magical fact that you are suppose searching any product in Google right the moment you log into social media the same type of ads will be featured across all social media you search a flower bouquet immediately you will get ads for flower bouquet where to buy or you search a mobile immediately Facebook if you open Facebook or Instagram or Twitter ads will be featured right again machine learning predictive machine learning and this is done by association at the level of customers who bought this also all item also bought.

So, you might find all of these things in Amazon and that might that he the machine is learning what is your likes and dislikes very important you are giving any food online food order you might also like this menu and you feel tempted to order that right. Again show you have watched one movie the platform social media platform will recommend to another movie and will invariably end up liking because the machine knows from your habits what you like. Game AI you simple very it is very easy to relate right AI is most related to games you are shooting one individual the individual is computer he or she simply dodges like it was shown in matrix movie right self driven car even the face unlock its machine learning right. So, these are the real world application of machine learning and finally, we need to know what are the application of machine learning in our area of interest and why why do we need machine how we can utilize all of these in our area of interest that is genomics. Now, for those who are seeking the answer what type of machine learning classification is suitable for healthcare delivery system the answer is semi clustered right or semi supervised.

For example, we have to first train the model with known data all right 100 data points on multiple known cases control data we need to populate the machine. So, that it learns and then when we give the machine some unknown data unknown images right it will automatically tell you how it is done. This is the basic model how any healthcare development health related development model is developed using ML we will be getting into it, but first let us look what are the main important area why we need ML or AI in genomics. Number one wherever there are sources of big data we need ML to answer we need ML as an answer to solve that problem and the most important is next generation sequencing. I told you in bio bio informatics is a very important part of next generation sequencing DNA sequence RNA sequence analysis the process is actually very easy the data analysis is the difficult part that takes a lot of time because we are using multiple models right.

A huge amount of data genomical amount of data is generated every year as you saw and very important. So, unraveling multiple genetic variation which is various very crucial for genomic research as well as personnel and precision medicine you already know that term from pharmacogenomics lecture. Identifying multiple variants pathological variants from all the sea of data definitely machine learning has got very important role in predicting and understanding clinical diagnosis right. Not only that we have got data from multiple fields genomic data, proteomic data, transcriptomic data when we need to integrate them proper interfacing there also machine learning plays a very important crucial role right. Predictive disease model nowadays multiple disease models are being populated multiple research projects are going on for example, your simple apple watch right or any android smart watch or any device healthcare device or sensor based device you are wearing they have been already populated with multiple data right from your body a human physiological system.

So, when the machine now collects some data from you it can easily predict what disease process you might have or what is suitable for you. So, it predicts and gives you some suggestions. So, early intervention very important. So, predictive model for disease risk as well as early intervention ML has got a huge role. Drug discovery and development again potential drug targets to tackle drug resistance multiple new drugs are being developed one.

So, the main answer to tackle resistance is pharmacogenomics and precision medicine right and again the answer to all of them lies in machine learning models. And needless to say population genetics evolutionary studies constantly the thing is evolving and with lots and lots of data in order to analyze them there is no other way than to involve or get into machine learning. So, that we can find any loss of function any missing link any genetic adaptation etcetera can now be easily predicted by machine learning compared to how it was done 20 years back. So, how to get into machine learning right? Now we know about machine learning the answer to that is computer programming right. First we need to start with computer programming traditional programming and then we need to develop this machine learning model with the help of programming.

So, where to start from the pole of multiple data scientists across all fields specially health research and healthcare delivery system and genetic and data analysis the number one choice is R programming language right. Now python comes close second it is actually a great choice for those who are trying to get into programming because python is easier to learn the annotations are easier to learn the syntax is much easier right. However, eventually in order to develop machine learning model related to disease and genetics R is the answer and understanding python helps us to easily understand R programming right. Now here are few facts about R program why R is important number one it is open source it is very free originally it was developed as an S platform S which was actually paid at bell lab right. Now they have release the source code and now you can download R from Google R studio right and R and you will be able to download it all right.

It is a free programming language and most important it has got lovely statistical programs all statistical analysis statisticians always favor R because there are so many good biological graphic data interfaces that can be represented with R for example, ggplot there is a module known as ggplot 2 there are module known as bio conductor all of them are very user friendly you just input the big data from any public repository and they will give you lovely looking colorful data images right just to start with and then you need to go into develop the further program when you learn more of the computer programming language. So, it is a functional program that is primarily written in C and Fortran languages right if you have an idea about computer science you can appreciate very important cross platform compatible. So, you across all platform you are in windows you are working Macintosh system R programming can be done in all even in android you can do R program there are a multiple apps by which you can explore R studio in your handheld device. There are multiple data packages alright multiple developers have already developed the program by which you can simply get going I mean even without knowing much you can get go. There are references in the reference section there are websites where you can download the R studio you can download the repository for getting into machine learning you can just try test those data and fill automate and it will show you how the things are getting just for your understanding right and everything is there for free right.

Now, of course, I told you it is extensively used by statistician and academicians and its popularity substantially increasing in recent years because the programming language is not only limited in the domain of computer science right. It is now superseded the boundaries of computer science it is extending the bioinformatics the health research cancer research genomic research everything needs machine learning and that whole statistical analysis and ML is standing on R platform. Definitely there are drawbacks because there is a steep learning curve it is not easy to learn right, but recent times with specially graphical user interface it is a bit better compared to how it was still the graphical user interface is limited. So, GUI stands for graphical user interface R studio is a software which you can download in your desktop which has got all the R commands and it is it has a tutorial sort of thing where you can easily learn right and you can start from there. And moreover you need powerful computers again to some memory allocation etcetera are an issue there are less documentation compared to other established platforms like Java and Python right.

However, the much I mean the thing that we already have is enough to get you started in developing any ML model related to any genomic software or any genomical data in your laboratory. So, how do you develop the phases of data development or model development in genomics is actually very easy not only genomics any ML model related to health science research. So, basically first we need to collect the data right. So, demography medical history genetic information whether this data belongs to the cases or whether this data belongs to the control. So, basically first data collection then clustering the data or categorizing that a labeling the data based on the features all right.

And then we train we input those data in the machine right in the model right in the program where it learns. So, over time with a lot of trials this is called data training right. So, the first step is actually data collection is also data mining right when we gather the data. So, what happens after the machine is properly trained ok, then what we do then it is time to test the efficacy. So, then we also give some unknown data right some unlabeled data and let the machine predict fine.

So, once we see all of them. So, once we the it is tested everything is being tested. So, with training set and test set everything is being tested then the finally, the model is built and it is ready to be validated. What is validation? Separate data sets right not training data set or not the test data set for which we already know the values right. So, real data from patients in a control scenario are collected and the machine is asked to predict them mind this is very different from test data. Test data is when we are developing the model, validation is already model is already developed and we are testing the model with known patient data right.

So, to test the robustness and reliability. So, once it is validated using real world patient then the program can be disseminated into clinical practice for disease prediction and management. Mind it very important thing you can see the goal is not to make a perfect guess. So, there are many who will argue that then what is the need of clinician or I mean there is an error in your model. So, how can you how can you think of it to replace the issue is we need to understand the things has to go hand in hand. Nobody can replace anyone the clinician cannot be replaced by machine and the machine learning expertise will always augment the clinician in order to take a good decision to take an informed decision to make good guesses to make good prediction so that it is the future is always better right.

So, it may not be perfect at this time and the machine might need more training right the feedback is also important. So, whatever we might think today the machine may have some domain which is already unexplored. So, as more and more RNDs developed we can train the model further with another I mean additional rare diseases data right and then it will be good. Lastly we will discuss in brief what is neural networking and artificial intelligence. So, neural networking right is a computer program the structure of which is based on brain, neuron, synapses.

So, it consists of multiple nodes right. Now each neuron or each node receives data it processes and then it delivers it to the next layer right. Now there is an activation function by which when the data this node specifically derives I mean acquires the data it delivers it to the next phase for further processing right and this goes in multiple layers. So, with multiple training and repeated data sets finally, we get the out there is a schematic representation of how a neural network actually looks as a multiple interconnections are there. So, there are sensors which acquire the data and it goes to multiple layers of machine learning and data processing and then we finally, get the result right. And you must have heard this neural networking where the machine is actively learning and adapting right.

So, they excel at multiple task for example, speech recognition natural language processing you have heard of chat GPT right the AI program where you can actually tell chat GPT to write a poem or write a text like some poet and it will actually mimic you can ask chat GPT to make a sound like that rapper and it will mimic how you write an answer script the chat GPT can easily learn right. Not only that now neural networking since it mimics the ability to learn it has got immense use in any genomics data analysis and healthcare analysis. So, whatever ML we are expecting the output from ML it is actually done via neural networking and neural networking can do it better for example, multiple encoding of multiple genomic data right. So, whenever there are multiple data the representation huge source multiple category multiple population extracting features from all of them is much better done using a ML model or using a machine or using an

AI that uses neural networking right. For example, extracting meaningful features from genetic sequences or genomic data right.

So, it is done by neural networking disease prediction therefore, disease prediction so variant analysis population genetics all the thing that you already discussed that is the use of AI. Now we know that it is use of ML. So, both artificial intelligence and machine learning via this neural networking are acting in tandem to make our healthcare delivery system better right. And so, these are one of the few examples where we do have a lot of genomic data. Now we already know in next generation sequencing analysis the sequences which are lost right loss of function the program can actually predict the missing areas again how it is done? It is you done by first of all feature extraction by knowing what can be there and done done by predictive modeling right.

So, rare variant everything everything you can think of the answer is in AI and ML. So, to summarize so, genomic generates big data right we already know and that data the final goal is to improve disease prevention diagnosis prognosis and treatment efficacy right. Machine learning is subfield of computer science, but now that is used and that is actually was generally used in analysis of the big data. However, and nowadays as you know machine learning has been incorporated into healthcare system to deliver predictive modeling to discover patterns to discover for helping drug discovery pharmacogenomics what not right with specially with the help of neural network. And the best program to start to develop any ML model is R programming right.

And R programming basic foundation along with knowledge about various genomic model of machine learning right has improved and is improving health informatics as well as bioinformatic research. So, I thank you for your patient hearing and these are my references and I will see you soon in the next class.