

PREDICTING THE VIRULENCE OF BACTERIAL PATHOGENS

**KANMANIVISHWAA
PARAMASIVAM
(720001564)**



Abstract


Infectious diseases that are on the rise or expanding into new regions are known as emerging diseases. Infections with viruses like Ebola and Zika are instances of emerging illnesses. Undiscovered infectious organisms may cause previously unknown illnesses. Emerging pathogens are becoming recognised as a serious microbiologic public health threat, despite the fact that medical communities have had to deal with emerging and reemerging infectious diseases since at least the 1950s. This has only occurred during the last several years. Emergence of more virulent bacterial strains and opportunistic infections, especially affecting populations of immunocompromised individuals, and development of new diagnostic tools like improvements in culture methods, development of molecular techniques, and implementation of mass spectrometry in microbiology are all major concerns. There are many emerging and reemerging bacterial infectious diseases, and we've identified 26 that need special attention. Most of them were spread by animals or water. This apparent evolutionary tradeoff between virulence and transmissibility is still the subject of theoretical debate, but it has been challenging to assess objectively. Another potential predictor of virulence is the mode of transmission. Studies comparing similar illnesses have shown that those transmitted by vectors or those that linger longer in the environment tend to be more lethal to humans. To that end, we hypothesised that modes of transmission other than direct contact, such as vectorborne transmission and routes including environmental components, would be associated with higher virulence. Our initial step was to identify what it is about each bacterial infection that makes them dangerous. Then, we used predictive machine learning models to see whether ecological features of viruses, such their transmissibility and tissue tropism, might be used as risk factors for virulence in humans. To be more precise, we examined the idea that pathogens would pose a greater threat if they were difficult to spread from person to person (a quality known as transmissibility), very severe (as measured by a severity scale), and able to infect a broad variety of tissues. Our findings suggest that there is no universal relationship between the kind of bacterial infection and the degree of sickness in a given patient. The risk variables of tissue tropism and, to a lesser degree, human-to-human transmissibility and transmission route, are significantly more accurate indicators of the severity of a disease than the taxonomy alone. In all classification schemes, bacteria that produced systemic infections, had neural or renal tropism, were transmitted through routes, or had a limited potential to transmit between persons were more reliably predicted to not cause a major disease. This means that a bacterial pathogen with systemic infection potential, brain or renal tropism, pathogen-to-pathogen transmission, or limited inter-host transmission, Our research shows that milder kinds of bacterial infections are far more common than their more severe relatives. There is a very high prevalence of severe pathogens among less severe bacterial pathogens, especially when compared to the prevalence of both the most and least severe pathogens. As compared to the least severe pathogens, there is a much higher proportion of severe pathogens. Nonetheless, the percentage of infections that are really fatal is quite low. However, when we're talking about less dire circumstances, that number drops dramatically.

Introduction

Emerging infectious diseases are illnesses that have either recently appeared in a population or have always existed but are experiencing a rapid increase in either their incidence or geographic range[1]. The epidemic has grown into a major social, economic, and political issue on a worldwide scale. Several notable pandemics, such as the 1918 H1N1 influenza pandemic and the HIV pandemic, claimed the lives of millions of people during the previous century. The Zika virus belongs to the Flaviviridae family of emerging arboviruses. The disease, which is spread by the Aedes mosquito, was discovered for the first time in Uganda in 1947. The concept of virulence dates to the early 19th century, when it was discovered that infectious diseases were caused by specific microorganisms. Since then, the concept of virulence has expanded to include a variety of factors that contribute to the severity of disease caused by infectious agents. One of the earliest examples of virulence research is the work of Louis Pasteur, who discovered in the 19th century that weakened or attenuated forms of microorganisms could be used as vaccines to protect against infectious diseases [2]. This discovery paved the way for the development of modern vaccines, which are used to prevent a variety of infectious diseases. The discovery of virulence factors, which are molecules or structures produced by microorganisms that contribute to their ability to cause disease, was another significant advancement in the study of virulence. For example, in 1890, Emil von Behring and Shibasaburo Kitasato discovered the diphtheria toxin, which helped to explain the pathogenesis of diphtheria and led to the development of antitoxin therapies. Researchers can now study the genetic basis of virulence in infectious agents thanks to recent advances in molecular biology and genomics[3]. There is a possibility that newly developing diseases are caused by infectious agents that have never been seen or studied before. In spite of the fact that medical communities have been dealing with emerging and reemerging infectious disorders since the 1950s, emerging pathogens are now regarded as a severe microbiologic public health hazard. This is the case despite the fact that medical communities have been dealing with emerging and reemerging infectious disorders since the 1950s.

As a result, new virulence factors have been discovered, as well as new therapies and vaccines. It was later detected in humans in several locations of Sub-Saharan Africa, and it was finally discovered in south-east Asia by the middle of the twentieth century[4]. It began in the twenty-first century on Pacific islands and spread to South America by 2014. It has travelled an alarming distance north since then, arriving in Mexico in November of 2015. There have been a variety of zoonotic infections in the twenty-first century, including Ebola, Hendra, and Nipah, and there is presently COVID-19[5]. According to a survey of the relevant literature, there are 1415 unique species of pathogenic microbes capable of causing illness in humans. There are 217 different viruses and prions, 538 different bacteria and rickettsia, 307 different fungi, 66 different protozoa, and 287 different helminths[6]. 868 of these illnesses, or 61%, are zoonotic, which means they may be spread between people and animals, and 175 of these pathogens are linked to "emerging" conditions.

Major factors contributing to the emergence of these infections include the emergence of more virulent bacterial strains and opportunistic infections, especially affecting immunocompromised populations; the development of new diagnostic tools, such as improvements in culture methods, molecular technique development, and the implementation of mass spectrometry in microbiology[7]; and an increase in human exposure to bacterial pathogens as a result of sociodemographic and environmental changes. New diagnostic approaches, like as advances in culture techniques, are another contributor. We found 26 main bacterial infectious illnesses that are on the rise or making a return in the world today. Most of these illnesses are either aquatic in origin or are spread by animals.




The degree to which a microbe is capable of causing disease, also known as its pathogenicity, is referred to as its virulence. According to a review paper that was published in the journal *Virulence* in the year 2010, the term "virulence" may be defined as "the degree of pathogenicity or the capacity of a microorganism to cause disease in a host[8]." It is a measurement of the severity of the disease that is produced by an infectious agent as well as the power of the infectious agent to cause harm to the organism that is being infected. Microorganisms that have a low virulence are more likely to cause moderate symptoms or none at all, whereas those with a high virulence can cause serious sickness or even death. The capacity of a microbe to penetrate host cells, create poisons or other virulence factors, elude the immune system of the host, and efficiently proliferate within the host are all aspects that contribute to the pathogen's virulence. The development of efficient therapies and vaccines, as well as the prevention and control of infectious illnesses, all depend on having a solid understanding of the pathogenicity of infectious agents. The concept of virulence, which plays an important role in infectious disease, has been the subject of a significant amount of research in the field of microbiology. The capacity to penetrate host cells, create poisons or other virulence factors, avoid the host immune system, and proliferate efficiently within the host all contribute to a microbe's virulence. One of the major contributors to the pathogenicity of the bacteria *Staphylococcus aureus* is its capacity to create virulence factors including toxins and enzymes [9]. For the purpose of creating efficient therapies and vaccines, as well as for the prevention and control of infectious illnesses, it is essential to have a firm grasp on the virulence of infectious agents. New virulence factors have been identified, and different strategies for combating infectious illnesses have been developed thanks to this line of inquiry.

Bacterial virulence ranges from non-harmful commensals to extremely dangerous pathogens. And they are so dangerous because they are able to produce poisons, adhere to and invade host tissues, and avoid being destroyed by the body's immune system. *Vibrio cholerae*, the microorganism responsible for the disease cholera, is a very dangerous example of a bacterium[10]. Cholera toxin, which is produced by this bacteria, causes watery stools and, if untreated, can lead to severe dehydration and death. And since it can stick to and colonise the small intestine, *Vibrio cholerae* is able to avoid the host immune response. *Streptococcus pneumoniae* is another instance of a very dangerous bacterium, as it may lead to several different illnesses such as pneumonia, meningitis, and septic shock. Pneumolysin, one of several virulence factors produced by this bacteria, destroys host cells and sets off an inflammatory reaction[11]. Effective therapies and vaccinations require knowledge of what makes germs so dangerous. Several virulence factors and pathogenesis pathways have been discovered, providing therapeutic and preventative targets for research.

The great majority of bacterial strains are completely harmless, and some may even be beneficial to human health. Natural bacteria in the human gut, for example, help digestion and make essential vitamins. It's thought that just about 1% of bacterial species can actually make people sick. *Salmonella*, pneumonia, and meningitis are just few of the many illnesses that may be caused by bacteria. *Mycobacterium tuberculosis* causes TB, the worst bacterial illness affecting humans. More than 1,700,000 people die every year from this illness, making it the leading infectious killer worldwide. Around 6% of cases are resistant to all or practically all antibiotics, and 13% are resistant to the majority of medicines.

In order to keep up with the exponential growth in protein sequences revealed since the dawn of the postgenomic era, it is essential to create computer techniques for discovering virulence factors based only on their sequence data. Notwithstanding the difficulty in providing an unbiased explanation of the likely evolutionary tradeoff between virulence and transmissibility, this is a fundamental problem that is the subject of current theoretical discussion[12]. The mode of transmission of an infection is



another potential indicator of its pathogenicity. Previous research has compared the mortality rates of various diseases and found that those spread by vectors and those with a greater ability for environmental persistence tend to have higher mortality rates overall. This led us to hypothesise that vectorborne transmission, as well as pathways that incorporated environmental factors, would be associated with higher virulence than direct transmission based on contact. It was postulated that there would be an association between phylogenetic closeness/clustering and host pathogenicity. Hence, we hypothesised that infection in nonhuman primate hosts, which is only achievable due to a narrow host range, may serve as a signal of pathogenicity. Our research led us to conclude that a widespread tissue tropism may be an indicator of greater pathogenicity.

One of our goals was to define the characteristics shared by all harmful microorganisms. Finally, we used predictive machine learning models to determine whether ecological features of viruses, such as their transmissibility and tissue tropism, may be used as risk factors for human virulence. To be more specific, we tested the idea that bacteria would be more dangerous if they were more difficult to spread from one person to another (a quality known as transmissibility), had a higher severity score, and could infect a wide variety of organs and tissues.

Trade-off Hypothesis

In evolutionary biology, there is the idea of a "tradeoff hypothesis," which states that organisms must choose between investing in reproductive qualities and investing in other traits, such survival or immunological function. Several researchers in evolutionary biology have found that the tradeoff theory may provide light on how features evolved and how animals prioritise their needs. Bacterial virulence evolution may be explained by the tradeoff hypothesis, which proposes that pathogenic bacteria must choose between host-based and host-independent survival strategies. According to this theory, highly pathogenic bacteria may have a harder time adapting to their natural habitat or may be more vulnerable to other conditions that restrict their development and survival. *Pseudomonas aeruginosa*, a harmful bacterium, provides evidence for the tradeoff theory in regards to bacterial virulence. These bacteria may infect a broad range of hosts due to its abundance of virulence factors. Yet, in order to better survive in the environment, certain strains of *P. aeruginosa* have evolved to be less virulent [13]. Pathogens and the host species they infect have been coevolving for millions of years, mostly as a consequence of the two sets of organisms learning from each other and adapting to one another. The virulence of diseases, which measures how much damage a bacteria can do to its host, is an example of how this coevolutionary process works. The mortality rate is often used as a measure in this context. The "virulence-transmission trade-off idea" was initially presented over 30 years ago and has since been a central tenet of research into host-parasite co-evolution. It is hypothesised that the parasite would incur an ever-increasing cost of virulence as it reproduces. This is due to the fact that it requires the host's resources in order to function. Replication costs will eventually reduce the rate of transmission since a higher replication rate inside a host is linked to a higher host mortality rate. Conflicting results from testing the hypothesis's predictions have raised questions about its use.

Research into the prediction of the effect and spread of infectious illnesses has been profoundly influenced by the assumption that there is an evolutionary trade-off between the severity of a virus and the ease with which it may be passed by Acevedo conducted a meta-analysis of the most

significant underlying relationships and drew attention to the surprisingly small number of empirical research that support this crucially essential hypothesis[14]. The 'trade-off theory' has been proposed as an alternative to the widely held belief that parasites should always develop towards avirulence (the 'avirulence hypothesis') for over twenty years[15].

When two tasks conflict with one another and neither can be completed at the same time, trade-offs arise. When we say "trade-offs," we mean that in order to improve one aspect, another must be diminished. Several factors necessitate the existence of trade-offs. Porter is the middle member in a trio. To begin, there may be incompatibilities between various elements of the product. Hence, the product is wildly popular with one demographic of consumers but fails to resonate with another. Second, you should always think about the potential good and bad results of your actions. The "virulence-transmission trade-off idea" suggests that an intermediate degree of virulence may maximise pathogenicity by swapping one level of virulence for another. Although an increase in pathogen virulence will lead to a faster replication rate, it will have the reverse effect on the amount of time it takes for the infection to spread from one host to another. As a result, the virus will exhibit optimal virulence, meaning that its spread will be facilitated to the fullest extent possible. As a direct consequence of the current situation, this will occur. This is how virulence should appear at its most potent for the spread of an infectious disease. The virulence trade-off hypothesis relies on the existence of antagonistic interactions between strains with different levels of host-damaging potential.


According to the trade-off concept, parasite virulence is an inevitable byproduct of parasite transmission. However, this view has been under increasing scrutiny since the 1990s. The fitness of these strains may be characterised by the basic reproduction number R_0 , which was defined by Anderson and May (1982) as a function of the pathogen's transmission rate (β), the host's population size (N), the host's natural death rate (μ), the host's pathogen-induced mortality rate (i.e., virulence, α), and the recovery rate v such that,

$$R_0 = \frac{\beta N}{\mu + \alpha + v}.$$

Assuming that each parameter develops independently while maintaining constant rates of recovery (v) and natural mortality (μ), R_0 would grow with rising transmission rate (β) or with lowering disease-induced mortality rate (α)[16]. If we follow this line of reasoning, we should anticipate that R_0 would be maximised while virulence will be reduced, or that will be less than 0. The final goal of this study is to establish tradeoff hypotheses between significant parameters in this dataset, such as transmissibility, route, tissue, and severity, as well as the severity scale, which measures the virulence of bacterial pathogens in the supplied data.

Analysis of existing projects (Literature Review)

There is a vast microbial community living in and on every human being, and although most of these bacteria are harmless, there are a few that are absolutely necessary for survival. Human gut bacteria




may be broken down into three groups depending on their preferred way of existence. Allergenic pathogens, opportunistic pathogens, and commensal non-pathogenic microbes all fall within this category. The human body is home to a vast microbial ecosystem consisting of billions of commensal bacteria, many of which are crucial to our survival and the maintenance of our health. There have been several suggested methods in recent years for classifying bacterial genomes as either pathogenic or non-pathogenic to humans. These models, which look at key aspects of categorization, may help us forecast the pathogenicity of new bacterial species and improve our knowledge of the whole pathogenic lifecycle.

We are studying and analysing a broad variety of literature and research from other domains to aid us in our inquiry, prediction, and evaluation of the pathogenicity of bacterial pathogens.

- An early and influential article on the topic, "The Evolutionary Ecology of Pathogen Virulence," was written by Paul Ewald and published in the journal *Trends in Microbiology* in 1991. Using the "trade-off hypothesis," which Ewald coined in this research, the author provides a theoretical framework for analysing how pathogen virulence has changed through time[17]. The trade-off theory proposes that the virulence (how much damage a disease does to its host) and the infectiousness (how easily it spreads) of a pathogen are opposing factors (the ability to spread from host to host). This theory proposes that a pathogen's optimum virulence is set by the trade-off between the costs and advantages of having a highly virulent strain and a less virulent one. In addition to population density and pathogen variety as environmental determinants, Ewald also considered the impact of host immunity on the development of pathogen virulence. The article has had significant impact on evolutionary medicine and our knowledge of the mechanisms behind the increase in disease virulence through time.
- CLAYTON E. CRESSLER makes an attempt to do this by distilling the vast theoretical literature on the topic into a collection of just a few credible predictions in an effort to accomplish this goal[18]. Following this, we will provide a comprehensive analysis of the empirical research that has been conducted before and is now accessible to assess these predictions. Because the presence of many infections might have an effect on the maturation of virulence, it is essential to take this into consideration. concentrating on the several ways in which a trade-off-theory prediction can need certain adjustments to be made owing to the existence of multiple illnesses. When it comes to the theoretical and empirical literatures, bridging the gap between the two is not an easy undertaking, as previous scholars have pointed out. This is because it is notoriously difficult to measure theoretical characteristics like death rate and transmission rate, which are recognised as being essential to models of virulence evolution. This is one of the reasons why this is the case. In addition, the vast majority of experiments are performed at the individual level, but the majority of theoretical considerations are centred on the population level. Because of these challenges, it is necessary to use system-specific proxies for transmission and virulence in what is known as a "test" of theoretical predictions. They believe that it is the responsibility of theorists to improve the links between mathematical models and empirical evidence, particularly given the difficulty of empirically quantifying the parameters identified by current theory as being critical in driving virulence development. They believe that this responsibility lies with theorists. This link may be established through the creation of models of individual systems or through the refinement of overarching theory to better capture widely used empirical metrics of virulence, such as

host morbidity or physiological state. Both of these approaches have the potential to establish this link.

- Michael S. Gilmore and colleagues wrote a review paper titled "The Interplay Between Antibiotic Resistance and Virulence in Bacteria" for the 2014 issue of Cold Spring Harbor Perspectives in Medicine[19]. Antibiotic resistance and pathogenicity in bacteria are discussed in depth in this article. Antibiotic resistance mechanisms and their potential effects on bacterial pathogenicity are introduced to kick off this review. The authors go on to detail the ways in which antibiotic resistance can affect virulence factors like adhesins, toxins, and secretion systems produced by bacteria. Possible outcomes of antibiotic use on bacterial virulence are discussed, such as the selection of more virulent strains and an increased risk of infection. The difficulties in creating novel antibiotics that are effective against virulence and resistance mechanisms are also discussed. Overall, "The Interplay Between Antibiotic Resistance and Virulence in Bacteria" gives valuable insights into the complicated link between antibiotic resistance and virulence in bacterial pathogens and highlights the necessity for a holistic strategy to addressing bacterial disease
- Medical Clinics of North America published a review by Thomas J. Marrie titled "Virulence and Pathogenesis of Bacterial Infections" in 1995. Marrie summarises the involvement of virulence factors and the host immune response in the pathogenesis of human illness by bacterial infections[20]. Marrie elaborates on the role of toxins, adhesins, and capsules, among other bacterially generated virulence factors, in the development of bacterial disease. Furthermore discussed is how knowledge of host-pathogen interactions might affect treatment outcomes. Emergence of antibiotic-resistant strains and the need for novel treatment tactics are only two of the difficulties in treating bacterial infections that are discussed in the article. The variables that contribute to the virulence of bacterial pathogens and the pathogenesis of bacterial illnesses are discussed at length in Marrie's review.
- In 2015, the journal Microbiology and Molecular Biology Reviews published a thorough review paper on the "Molecular Mechanisms of Bacterial Virulence" by José R. Penadés [21]. Bacterial virulence is discussed in detail, with the authors dissecting the function of virulence factors, quorum sensing, and bacterial secretion systems. This study focuses on the molecular pathways that contribute to the pathogenicity of several bacterial infections, both Gram-positive and Gram-negative. The authors also stress the significance of learning about the genetic and regulatory processes that govern bacterial virulence and the ways in which the surrounding environment may affect the expression of virulence genes. This study also sheds light on how bacterial pathogenicity has changed over time, from the processes of horizontal gene transfer to the development of antibiotic resistance. The authors also explore ways in which this information may be used to advance the creation of novel treatments for treating bacterial infections. When taken as a whole, "Molecular Mechanisms of Bacterial Virulence" is an excellent reference for scientists and medical professionals studying bacterial pathogenesis and trying to create new treatments.
- In this section, we will be referring to the study that Helen.C. Leggett carried out in 2017 in order to identify the link between the various virulence characteristics. This article is included in a themed edition of the journal titled "Opening the black box: reexamining the ecology and evolution of parasite transmission," and the title of that issue is also the name of the issue[22]. When everything is taken into consideration, the results provide insight on how the mechanical components of a parasite's life cycle, especially its mode of transmission, may



have an impact on the virulence of the parasite. In light of the fact that comparative analyses can confirm some predicted relationships with virulence while calling into question the effect of other variables, it is clear that this method has a great potential for elucidating the evolution of virulence across species and in a variety of ecological, evolutionary, and epidemiological contexts. This is because comparative analyses can confirm some predicted relationships with virulence while calling into question the effect of other variables.

- According to S. Alizon and A. Hurford, who assert that this is at the centre of the current dispute in the field, the trade-off hypothesis and its fundamental expansions are required to evaluate the qualitative effects of virulence control measures[15]. These two researchers claim that this is one of the most contentious issues in the field. As a means of providing context, they provide a concise summary of the development of virulence research and the path that led to the trade-off hypothesis. They overcame this obstacle by analysing the developments that had taken place over the course of the previous 10 years and argued that, in light of these developments, the trade-off paradigm ought not to be abandoned. Alternately, they suggested that these novel viewpoints should be included into the already accepted theory in order to pave the way for productive study in the years to come. The lack of evidence supporting the trade-off hypothesis is more of a call to enhance the theory and research rather than a call to reject the hypothesis altogether. The identification of trade-off curves will continue to be challenging. In most cases, transmission and virulence cancel one other out, although recovery may also play a vital role. The trade-off hypothesis provides a framework for comparing qualitative or theoretical results, which assists in the process of finding solutions to new problems. This is a theoretical study that we are adopting, and we are using it in our research, in order to compare and associate all of the many components with one another.
- In their study on the evolution of emerging pathogens, Camille Bonneaud and Ben Longdon relied on evolutionary theory to estimate how new infectious illnesses would emerge during the course of their investigation. In this publication, the authors offered a concise assessment of our current knowledge of the growth of pathogens in new host species, shedding light on the potential evolutionary repercussions that may be caused by human control tactics during a pandemic[4]. They discussed what virulence is and how it functions, defined virulence in the context of a disease outbreak, and brought attention to the tension that exists between exploitation and mobility. Increasing the transmission barriers via behaviour in order to decrease contamination, such as staying away from sick people and changing how you prepare food. They also showed how easy it is for a disease to adapt to a new host by using a mouse model. Because it can provide a more in-depth knowledge of the behaviour and reactions of infectious diseases, evolutionary biology provides a powerful strategy for addressing issues relating to the health of humans, animals, plants, and even ecosystems. This is because evolutionary biology is able to provide a more in-depth understanding of the behaviour and reactions of infectious diseases. This is due to the fact that evolutionary biology is capable of providing a more in-depth understanding of the behaviours and responses of infectious illnesses. We are able to comprehend and learn how to control and halt the development of bacterial disease pathogens via the host as well as how to enhance the immune level of the host with the assistance of this body of literature. In addition, we are able to learn how to do both of these things.

- Rupanjali Chaudhuri and Srinivasan Ramachandran conducted research on the possibility of predicting virulence variables with the use of bioinformatics techniques. Even though there have been advances in anti-infection technologies, infectious diseases continue to be an issue for humanity[23]. It is absolutely necessary for a pathogen to produce virulence factors in order for the pathogen to be able to cause sickness in the host. These qualities make it easier for the virus to establish a foothold within the host and, in certain instances, to avoid being detected by the host's defence mechanisms. Both of these abilities are facilitated by the infection's capacity to replicate inside the host. These factors allow for colonisation of the host niche, which eventually results in harm to the host tissue. After that, a concise overview of these substances and an explanation of how the use of bioinformatics led to the prediction of their existence is provided. They were able to successfully estimate the virulent factors by utilising the C code, and by using this study, we are able to understand how to analyse and anticipate the severity of the virulent factor by using python.
- The article "Tissue tropism and transmission ecology anticipate virulence of human RNA viruses," which was written by Brierley, is a companion piece to our own study, the purpose of which is to forecast the infectious capacity of RNA viruses. His study is an important addition to the expanding corpus of work that is seeking to analyse the origin of infectious diseases and make educated predictions based on that understanding[1]. In this paper, we provide a novel contribution, to the best of our knowledge, by concentrating on ecological predictors of the virulence of human RNA viruses. These predictors may be included in holistic frameworks alongside other models, such as those that predict the dynamics of emergence. The highlighted random forests provide valuable insight into the evolutionary determinants of virulence in emerging diseases, which is one reason why they are an effective prediction model. We strongly suggest that, in the future, prognostic research and preparedness programmes relevant to emerging illnesses include consideration of the probability of pathogenicity in humans. This is the first research of its kind, and it compares the virulence of every single species of RNA virus that has ever been discovered in humans. We find that the severity of illness varies dramatically from one viral family to the next viral family, and that risk factors of tissue tropism and, to a lesser degree, transmission channel and quantity of human-to-human transmissibility transcend beyond taxonomy to predict severe disease. These are the fundamental reasons why we have decided to continue in the same method as this work, and the majority of the virulence factors are similar to the data that has been given.

OVERVIEW OF THE MODELLING PROCESS

The modelling method for predicting the virulence of bacterial pathogens requires numerous phases, including the collecting of data, the selection of features, the building of the model, and its evaluation. Here is an overview of the modeling process with reference:

- I. **Data collection:** A huge variety of bacterial strains, both pathogenic and non-pathogenic, are used to gather genomic and other forms of data. These data may contain whole genome sequences, transcriptomic or proteomic data, clinical or epidemiological data, or other pertinent information.
- II. **Feature selection:** From the acquired data, relevant characteristics that are connected with virulence are selected. This might entail a variety of approaches, such as computer programmes that do machine learning, statistical studies, or bioinformatics tools.

- III. **Model Building:** Based on the Feature Selections, a Predictive Model is Constructed. This might utilise a number of different machine learning or statistical approaches, such as logistic regression, support vector machines, or random forests. The model is trained using a subset of the data and validated using a different subset of the data.
- IV. **Model Evaluation:** The performance of the model is assessed using a number of different measures, such as sensitivity and specificity, or the area under the receiver operating characteristic curve (AUC-ROC). It is also possible to validate the model by applying it to an other dataset in order to investigate its generalizability.

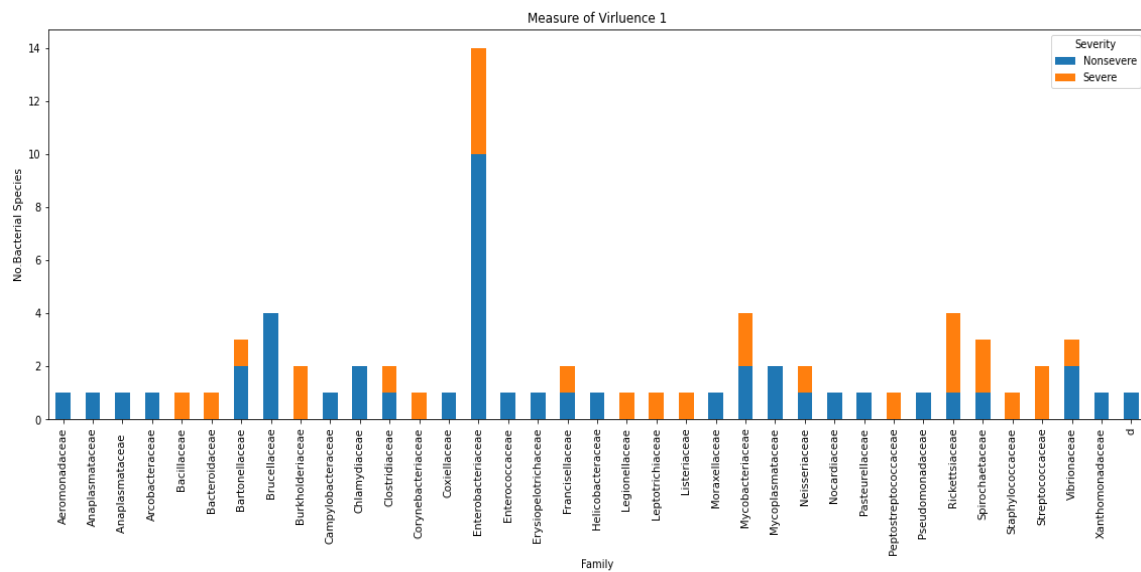


In a research that was conducted by Chen et al., one example of a modelling technique that was used to predict the pathogenicity of bacterial pathogens was presented in 2018. Using genetic and clinical data, the authors of this work constructed a machine learning model in order to predict the potential virulence of *Streptococcus pneumoniae*. In order to construct and assess the prediction model, the authors used a number of distinct strategies for feature selection and machine learning methods.

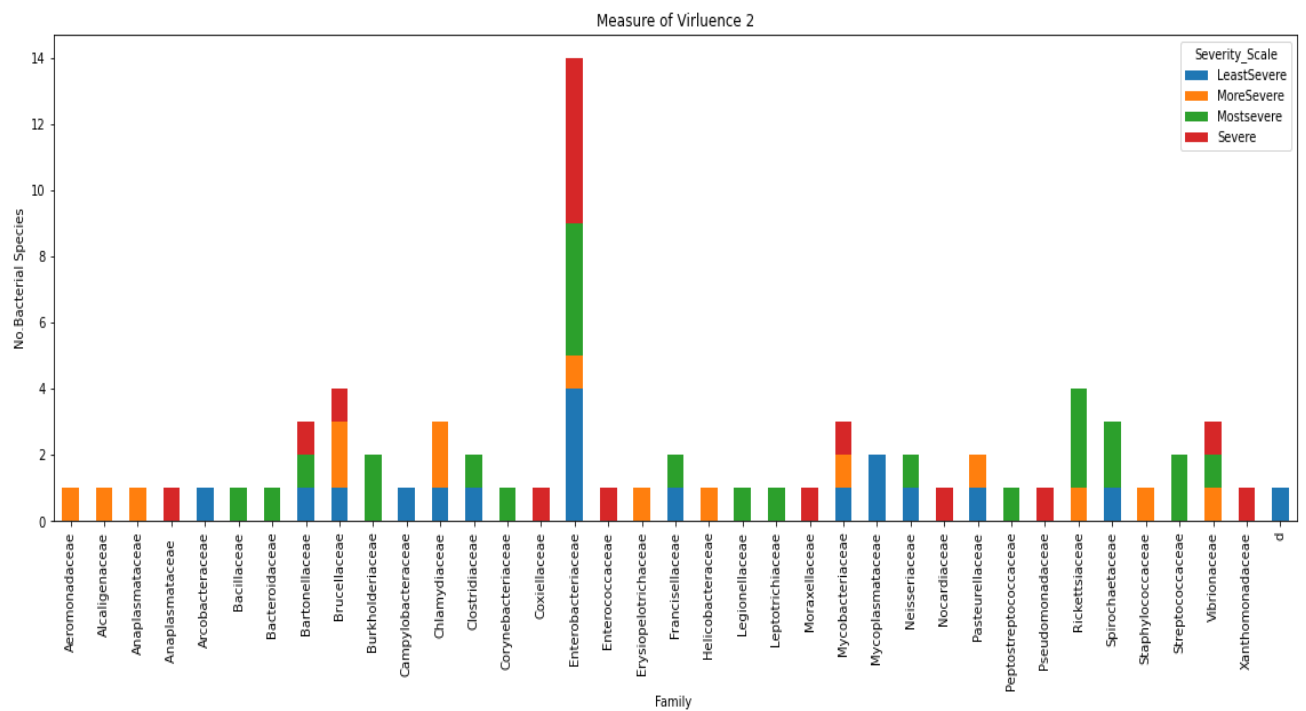
DATA COLLECTION

Virulence of Bacterial Pathogens

According to the information that was presented, there were a total of 36 families and 76 distinct types of bacterial pathogens that were capable of infecting humans. Bacterial infections might be caused by viruses. We use two distinct severity measures, one of which is obtained from the data, and one of which is produced directly from the data, to identify and quantify the pathogenicity of bacterial pathogens. One aspect of virulence is the relationship between the severity of an infecting family and the number of additional bacterial species it may spread. The second aspect of virulence is the correlation between the severity scale of the family and the number of bacterial species that may be transferred. Results from the Measure of Virulence 1 study, which employed a technique with two separate classifications, determined that 28 of them were capable of producing "severe" clinical illness and that 44 were classified as "nonsevere" following a thorough evaluation of the relevant previous research.



The second aspect of the virulence assessment is the connection between the Severity Scale for Families and the Number of Bacterial Species Transmission. It was determined that eighteen of them caused the "least severe" clinical sickness, sixteen caused "severe" illness, fifteen caused "more severe" illness, and twenty-four caused the "most severe" disease. Throughout the course of this research, two different strategies were used.



Model Explanation(used in analysis)

The virulence of bacterial infections is predicted using four different models (Supply vector machine, Random Forest, Linear Regression, and Gradient Boost Classifier) and their relative strengths and weaknesses are compared.

SVM for Supply Vector Machine

Supervised learning systems employ labelled input and output data for classification purposes in the fields of AI and machine learning. After the data has been tagged, these systems may use the tags to decide how to organise the information. Often abbreviated to "SVM," the Support Vector Machine is a well-known supervised method for classifying data that falls under the umbrella of machine learning. The 1980s saw the birth of this technique. The answers obtained by this method are generally reliable when the data may be partitioned in either a linear or non-linear fashion. In cases when the information cannot be partitioned into two distinct types, this technique is still useful. With data that can be partitioned along a straight line, SVMs can create a separation hyperplane that maximises the gap between categories. A separating hyperplane is used to do this. This margin is measured along a line that is perpendicular to the hyperplane.

Random Forest


Random forest is a kind of supervised machine learning that has found widespread use in the fields of classification and regression analysis. Samples with the most votes are used for classification, while for regression, the samples are averaged. Both of these approaches will be broken down into its component parts below. Random Forest is based on the premise that the combined efficacy of several distinct models (or "trees") will outperform that of any single model used in the decision-making process. The use of trees in the decision-making process is symbolic of this. The single most essential piece of advice that can be offered to academic researchers is to consider random forest as the principal approach for carrying out work linked to prediction rather than the traditional regression and individual decision tree analytics tools.

Linear Regression

The purpose of linear regression analysis is to predict one variable from the known or predicted values of other variables. Using a mathematical model that specifies the connection between the variables in question, it is feasible to forecast the future values of two or more variables. Linear regression is often employed in research that aims to make predictions. Predictions are made using linear relationships between the dependent variable (the goal) and one or more independent variables (predictors). Analysis using this method seeks to make the most precise prediction possible about the value of the dependent variable by estimating the coefficients of the linear equation using one or more independent variables. In linear regression, one attempts to find a line or surface that minimises the gap between the anticipated and observed values.

Gradient Boosting Classifier

Machine learning's gradient boosting classifiers are a collection of techniques for combining many weak learning models into a single strong prediction model. It is predicated on the idea that using the best possible model for the future in conjunction with models from the past will lead to the lowest possible prediction error. That's the "lowest overall error in prediction" theory, to put it simply. Rather than providing a single, definitive model for making predictions, it provides an ensemble of models, the vast majority of which are decision trees. This kind of model has been widely criticised for its lack



of realism. To conduct our analysis of the experimental stroke data, we combine two approaches. Among these methods are SMOTE (Minority Oversampling) and the gradient boosting approach (best for asymmetrical data).

Procedure for data analysis(Data Collection and Data Building)

In order to examine risk factors that are predictive of virulence, we first divided the 76 bacterial pathogens into a single training set and a test set according to taxonomic factors and severity. The training set consisted of all of the pathogens with the most severe symptoms. The test set consisted of all of the pathogens with the least severe symptoms. This was done in order to reduce the likelihood of any biases occurring as a result of differences in personality traits across groups. On the other hand, not a lot of information on bacterial pathogens is given in this instance. It is not feasible to do analysis or prediction using lime and shap in Python unless the given data set has at least 200 rows (AUC score and roc curve). As a result, we are using the method of upsampling in this circumstance in order to raise the total number of data entries for the purpose of the process of analysis and prediction. After that, we distinguish the feature columns as a distinct variable and the severity and severity scale as a separate variable for both measurements of virulence. And then the training set and the test set are split, with the test size set to 0.33 and the random state set to 21, both of which are acceptable for the analytic technique that is being used. We construct many different classification machine learning models in order to predict the virulence of bacterial infections. These models all need to be compared to one another in order to discover which one is the most effective.

After the upsampled and encoded data set has been split, the next step, classification using the models described above, will take place. During the process of classification, we are going to assess the function of the classifier and then it fits with the partitioned training dataset. After that, we calculate the prediction, as well as the likelihood of the prediction, by using the dataset that was used for the split test. Next, we do analysis by using the prediction and the likelihood of the prediction in order to get the AUC score. After that, we provide a classification report that consists of recall, precision, f1 score, support, and accuracy, and we conclude with a confusion matrix. We are going to compare the AUC score, the interpretation of the confusion matrix, and the values of the classification components in order to figure out which model is the most effective in predicting the virulence of bacterial pathogens.

RESULTS:

Output and Model Comparison(Data Evaluation)

Classification Report:

Precision, recall, f1 score, support, accuracy, micro average, and weighted average are the 7 components that comprise the Classification report that is derived from the model prediction probability. By considering these 7 parameters, we may choose the most promising of these four models. In the meanwhile, we only learn about those seven elements. The model has a perfect precision of 1.0 since it never gives a false positive identification. If a model has a recall of 1.0, then it accurately classifies 100% of true positives without making any oversights. The optimal model would have an F1 score of 1.0, which is a combination of recall and accuracy. Support is used to determine how many samples are included in each metric. Accuracy of the model expressed as a decimal;

complete accuracy for a model is 1.0. The last two parameters, micro average and weighted average, are derived from the first five considerations and hence are rather unimportant for the comparison.

Classification Report	Precision	Recall	F1 Score	Support
	<i>Severe</i>			
CR1	0.90	0.92	0.91	51
CR2	0.92	0.96	0.94	51
CR3	0.88	0.90	0.89	51
CR4	0.91	0.96	0.93	51
	<i>Non Severe</i>			
CR1	0.87	0.84	0.86	32
CR2	0.93	0.88	0.90	32
CR3	0.84	0.81	0.83	32
CR4	0.93	0.84	0.89	32

We will begin with a discussion of the classification report for all models of virulence measurement for bacterial pathogens based on Severity. When comparing these four classification reports of the four respective models, as we explained earlier, the model that has a precision, recall, f1 score, support, and accuracy value that is close to 1.0 is regarded as a better model than the other models. In this part, we compared the classification of the four models with severity, and the results showed that each model is an improved model. The Random Forest classifier model is superior than other models in terms of its precision, recall, f1 score, support, and accuracy values, all of which are closer to 1 than with the other models. Even Gradient boost Classifier and SVM are both excellent, and the values of their parameters are relatively near to that of random forest. In particular, gradient boost classifier is really close to random forest, but it is not superior to random forest. Therefore, according to the findings of the classification report, the best model is the random forest.

Classification Report	Precision	Recall	F1 Score	Support
	<i>Macro avg</i>			
CR1	0.89	0.88	0.88	83
CR2	0.93	0.92	0.92	83
CR3	0.86	0.86	0.86	83
CR4	0.92	0.90	0.91	83

	<i>Weighted avg</i>			
CR1	0.89	0.88	0.88	83
CR2	0.93	0.93	0.93	83
CR3	0.87	0.87	0.87	83
CR4	0.92	0.92	0.91	83

In this case, we are analyzing the classification report of the measure of virulence of bacterial pathogens based on the Severity scale, which consists of four severity level categories, including "Least severe," "Severe," "More severe," and "Most severe." In this particular instance, while comparing, It is not the same as what was described earlier; in this case, the recall for the most severe condition in SVM is very close to 1.

Classification Report	Precision	Recall	F1 Score	Support
	<i>More Severe</i>			
CR5	1.00	0.17	0.29	12
CR6	1.00	0.75	0.86	12
CR7	0.86	0.50	0.63	12
CR8	1.00	0.75	0.86	12
	<i>Most Severe</i>			
CR5	0.72	0.91	0.81	32
CR6	0.88	0.88	0.88	32
CR7	0.79	0.84	0.82	32
CR8	0.88	0.88	0.88	32

Classification Report	Precision	Recall	F1 Score	Support
	<i>More Severe</i>			
CR5	1.00	0.17	0.29	12
CR6	1.00	0.75	0.86	12
CR7	0.86	0.50	0.63	12
CR8	1.00	0.75	0.86	12
	<i>Most Severe</i>			
CR5	0.72	0.91	0.81	32
CR6	0.88	0.88	0.88	32
CR7	0.79	0.84	0.82	32
CR8	0.88	0.88	0.88	32

Because of this, we are going to examine the accuracy, micro average, and weighted average of the classification reports produced by the four models. Random forest and the Gradient Boost classifier both have values that are extremely close to 1.0, in contrast to the three parameters that were described before. As a result, there is no way for us to determine in this circumstance which model is superior to the others. Therefore, it is necessary for us to look at the AUC scores among those models.

Classification Report	Precision	Recall	F1 Score	Support
	<i>Macro avg</i>			
CR5	0.86	0.72	0.71	83
CR6	0.91	0.88	0.89	83
CR7	0.83	0.78	0.79	83
CR8	0.91	0.88	0.89	83
	<i>Weighted avg</i>			
CR5	0.82	0.72	0.71	83
CR6	0.91	0.88	0.89	83
CR7	0.83	0.78	0.79	83
CR8	0.91	0.88	0.89	83

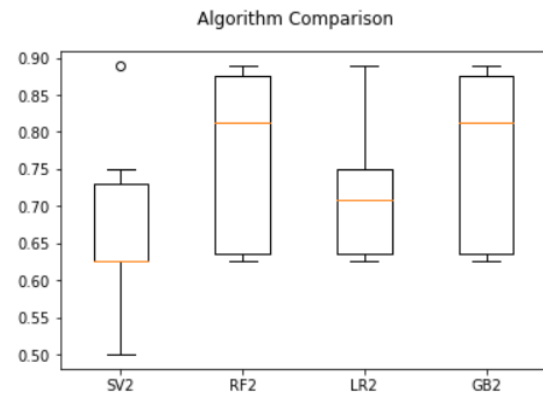
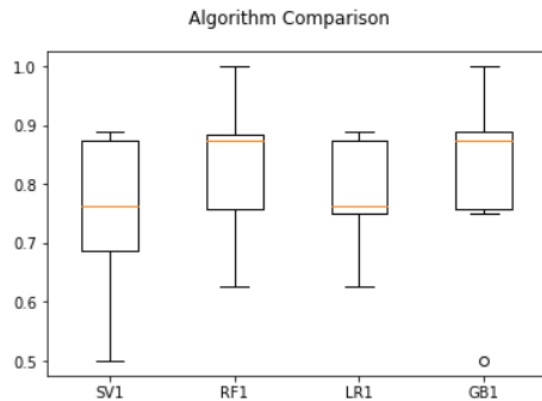
Following the comparison of the classification report, the AUC score of each of the four models will be analysed and compared. In this context, the term "AUC score" refers to the area under the ROC curve (AUC). The results were considered excellent when the AUC value was between 0.9 and 1, good when the AUC value was between 0.8 and 0.9, fair when the AUC value was between 0.7 and 0.8, poor when the AUC value was between 0.6 and 0.7, and failed when the AUC value was between 0.5 and 0.6. In the comparison of the AUC score of the measure of virulence based on Severity, it was found that "the values of AUC of SVM and Linear regression are below 0.95." Because of this, we have decided not to take into consideration both models. When comparing the two models whose AUC scores are both more than 0.95, the Random Forest Classifier emerges as the superior option. This is due to the fact that Random Forest's AUC score value of 0.984 is greater than Gradient boost's AUC score value of 0.971. In this case as well, Random forest proves to be the superior model to others, just as it did in the classification report.

MODEL	AUC (Measure of virulence 1)	AUC (Measure of virulence 2)
SVM	0.882	0.950
Random Forest	0.984	0.983
Linear Regression	0.946	0.907
Gradient Boost	0.971	0.953

Concerning the above scenario, i.e. the Classification report on the the measure of virulence of bacterial pathogens based on severity scale, we did not get any result. In view of the above, it is clear that this AUC score comparison study has the highest concern. Since the AUC score for linear regression is 0.907, which is much lower when compared to the scores of other models and is also lower than 0.95, we will not be looking into this model. The other three models, all of which have an AUC score that is more than 0.95, may be considered superior than the first model. However, in this particular instance, it is not the same as the Classification report since there is a larger gap between Random forest and the other two models. These two models both have an AUC score that is very close to 0.95, but the Random Forest model has an AUC score of 0.983, which is much closer to the value 1.0 than the other two. Now, after much deliberation, we have arrived at the conclusion that the Random Forest model is superior than the others.

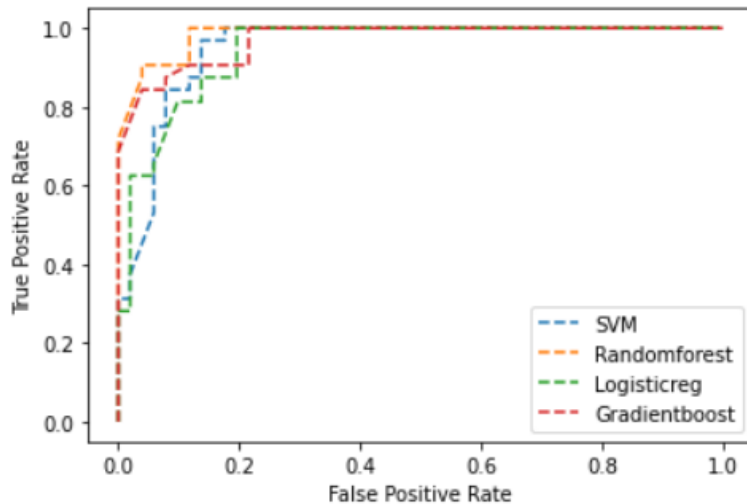
SV1: 0.758333 (0.123009)
 RF1: 0.818056 (0.115612)
 LR1: 0.783333 (0.088541)
 GB1: 0.830556 (0.139000)

SV2: 0.661111 (0.104896)
 RF2: 0.769444 (0.116402)
 LR2: 0.709722 (0.079894)
 GB2: 0.769444 (0.116402)

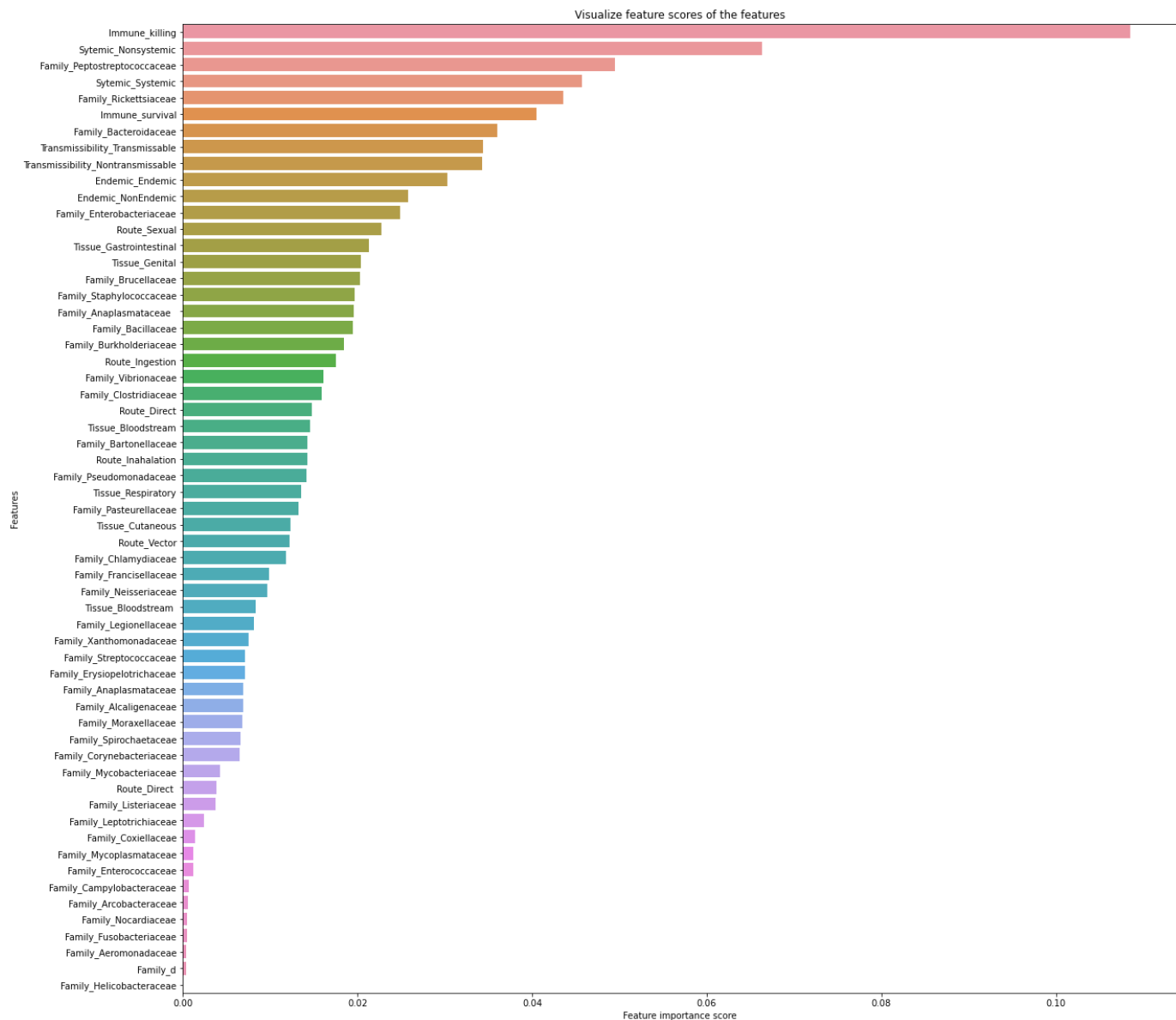


Classification Report	Accuracy	Classification Report	Accuracy
CR1	0.89	CR5	0.80
CR2	0.93	CR6	0.89
CR3	0.87	CR7	0.82
CR4	0.92	CR8	0.89

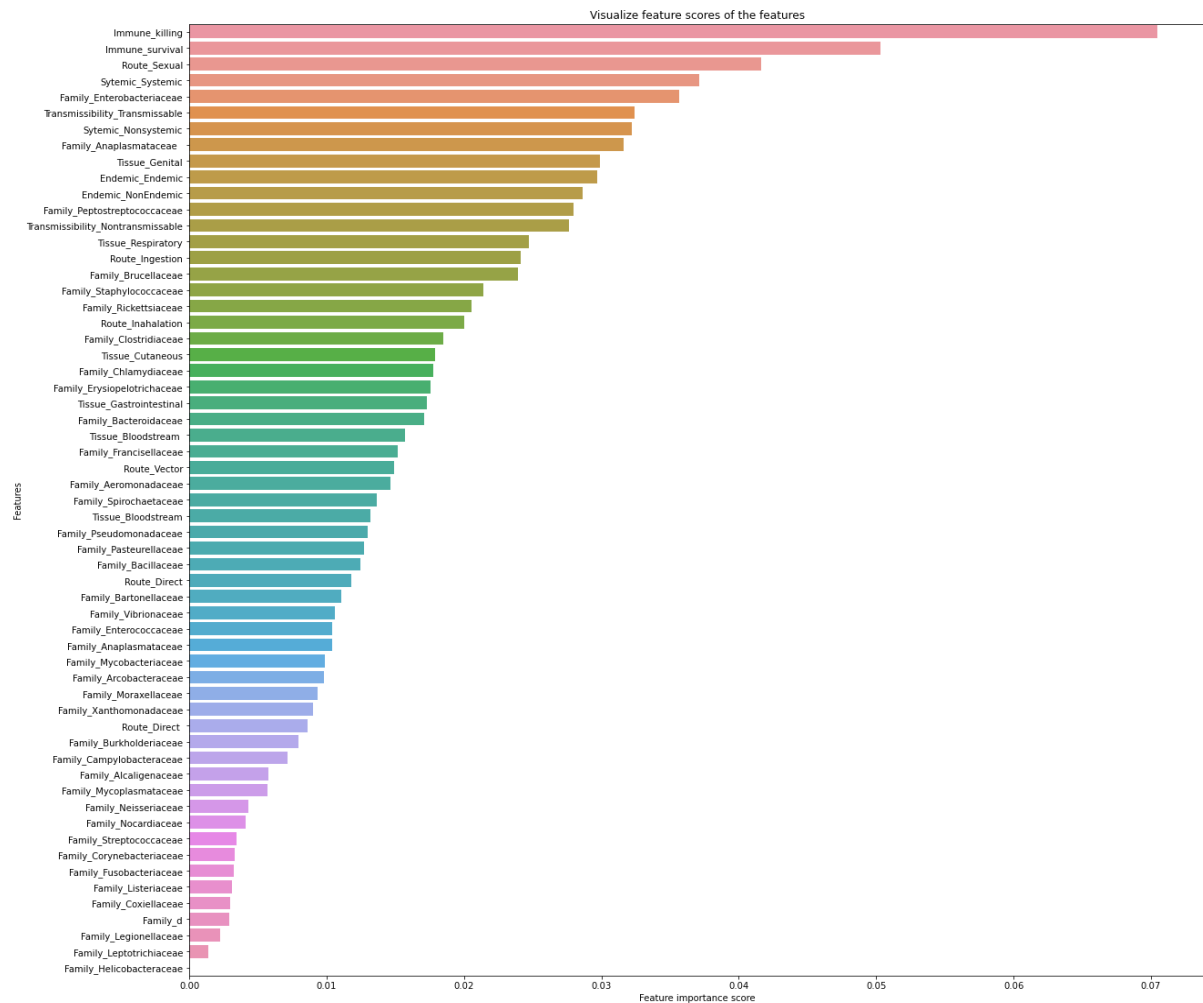
At this point, we have reached the conclusion of our comparison and study of the models, which we are doing only because we want accurate results and predictions. The box plot that is used to compare the algorithms behind the models has a number of parameters. However, in this case, we are simply going to compare the upper fence, the third quartile, and the accuracy number. Following an analysis of the importance of accuracy, the Gradient Boost Classifier has been shown to be the most effective model for determining the virulence of bacterial infections according to their level of severity. However, we performed three separate analyses, and in all of those analyses, random forest came out on top in two of them. In this instance, random forest was also quite close to the gradient boost classifier. In the end, we came to the conclusion that random forest is the optimum model for determining the level of virulence based on severity as well as the roc curve. On the other hand, when it comes to the Severity scale, both Random forest and Gradient boost classifier have the same accuracy score, which is 0.769, and the same box plot values. We have come to the conclusion that Random forest is the best model when compared to the other models after comparing the AUC values obtained from the measurement of virulence based on the severity scale.



Immune_killing	0.108458	Tissue_Cutaneous	0.012302
Sytemic_Nonsystemic	0.066280	Route_Vector	0.012218
Family_Peptostreptococcaceae	0.049520	Family_Chlamydiaceae	0.011860
Sytemic_Systemic	0.045685	Family_Francisellaceae	0.009855
Family_Rickettsiaceae	0.043572	Family_Neisseriaceae	0.009658
Immune_survival	0.040510	Tissue_Bloodstream	0.008313
Family_Bacteroidaceae	0.036057	Family_Legionellaceae	0.008119
Transmissibility_Transmissable	0.034400	Family_Xanthomonadaceae	0.007568
Transmissibility_Nontransmissable	0.034318	Family_Streptococcaceae	0.007168
Endemic_Endemic	0.030348	Family_Erysipelotrichaceae	0.007149
Endemic_NonEndemic	0.025807	Family_Anaplasmataceae	0.006927
Family_Enterobacteriaceae	0.024874	Family_Alcaligenaceae	0.006891
Route_Sexual	0.022787	Family_Moraxellaceae	0.006858
Tissue_Gastrointestinal	0.021328	Family_Spirochaetaceae	0.006622
Tissue_Genital	0.020390	Family_Corynebacteriaceae	0.006513
Family_Brucellaceae	0.020311	Family_Mycobacteriaceae	0.004317
Family_Staphylococcaceae	0.019687	Route_Direct	0.003901
Family_Anaplasmataceae	0.019598	Family_Listeriaceae	0.003728
Family_Bacillaceae	0.019513	Family_Leptotrichiaceae	0.002475
Family_Burkholderiaceae	0.018446	Family_Coxiellaceae	0.001459
Route_Ingestion	0.017520	Family_Mycoplasmataceae	0.001257
Family_Vibrionaceae	0.016156	Family_Enterococcaceae	0.001228
Family_Clostridiaceae	0.015865	Family_Campylobacteraceae	0.000669
Route_Direct	0.014814	Family_Arcobacteraceae	0.000653
Tissue_Bloodstream	0.014622	Family_Nocardiaceae	0.000527
Family_Bartonellaceae	0.014322	Family_Fusobacteriaceae	0.000485
Route_Inhalation	0.014301	Family_Aeromonadaceae	0.000435
Family_Pseudomonadaceae	0.014223	Family_d	0.000364
Tissue_Respiratory	0.013537	Family_Helicobacteraceae	0.000000
Family_Pasteurellaceae	0.013234		

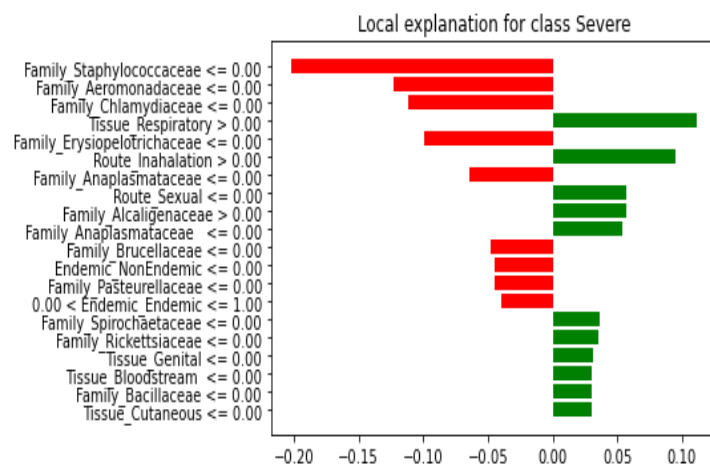
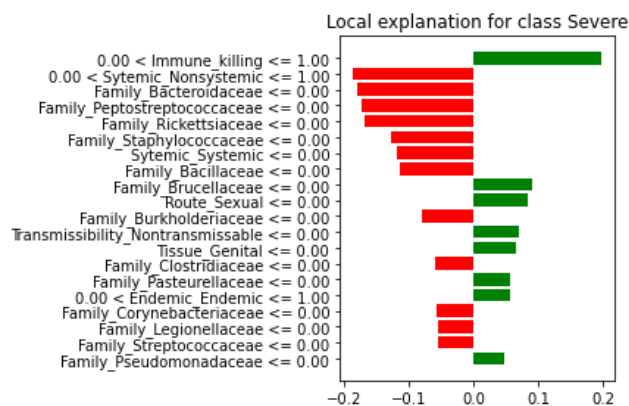


Selecting the right features for these models is an important but frequently overlooked aspect. Wasted data introduces bias, which screws up our machine learning's ultimate outcomes. Methods that assign a value to each feature used as input to a model are collectively referred to as "Feature Importance" methods. These values simply describe the "importance" of each feature. Feature significance is determined by multiplying the magnitude of the reduction in node impurity by the likelihood of reaching that node. By dividing the total number of samples by the number of samples that make it to the node, we can get the node probability. If a characteristic has a greater value, it is more crucial.



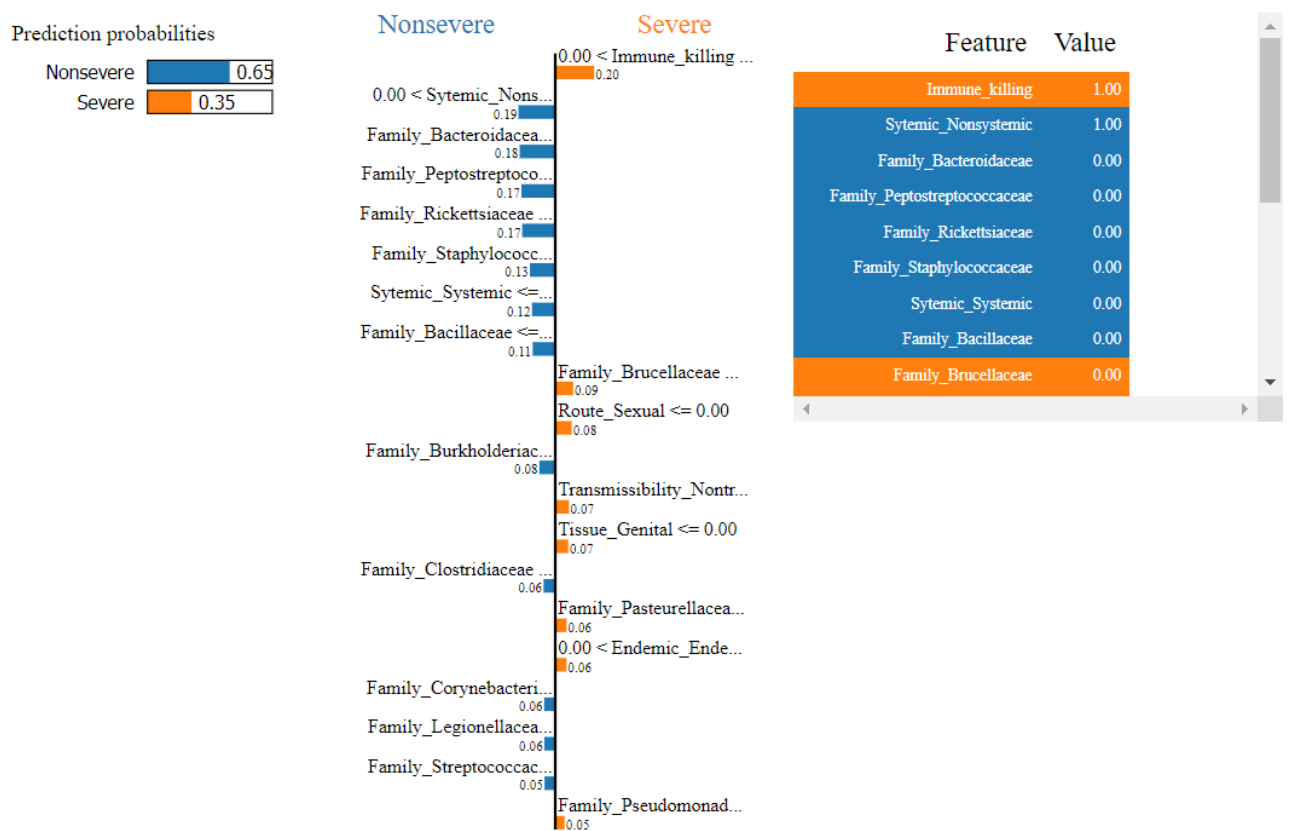
Immune_killing	0.070458	Tissue_Bloodstream	0.013200
Immune_survival	0.050312	Family_Pseudomonadaceae	0.013015
Route_Sexual	0.041622	Family_Pasteurellaceae	0.012752
Sytemic_Systemic	0.037137	Family_Bacillaceae	0.012428
Family_Enterobacteriaceae	0.035627	Route_Direct	0.011787
Transmissibility_Transmissable	0.032418	Family_Bartonellaceae	0.011041
Sytemic_Nonsystemic	0.032241	Family_Vibrionaceae	0.010594
Family_Anaplasmatataceae	0.031592	Family_Enterococcaceae	0.010384
Tissue_Genital	0.029909	Family_Anaplasmatataceae	0.010380
Endemic_Endemic	0.029725	Family_Mycobacteriaceae	0.009877
Endemic_NonEndemic	0.028608	Family_Arcobacteraceae	0.009817
Family_Peptostreptococcaceae	0.027966	Family_Moraxellaceae	0.009312
Transmissibility_Nontransmissable	0.027614	Family_Xanthomonadaceae	0.009030
Tissue_Respiratory	0.024711	Route_Direct	0.008591
Route_Ingestion	0.024112	Family_Burkholderiaceae	0.007978
Family_Brucellaceae	0.023955	Family_Campylobacteraceae	0.007137
Family_Staphylococcaceae	0.021438	Family_Alcaligenaceae	0.005792
Family_Rickettsiaceae	0.020525	Family_Mycoplasmataceae	0.005666
Route_Inhalation	0.020036	Family_Neisseriaceae	0.004331
Family_Clostridiaceae	0.018503	Family_Nocardiaceae	0.004078
Tissue_Cutaneous	0.017923	Family_Streptococcaceae	0.003414
Family_Chlamydiaceae	0.017749	Family_Corynebacteriaceae	0.003301
Family_Erysipelotrichaceae	0.017565	Family_Fusobacteriaceae	0.003220
Tissue_Gastrointestinal	0.017322	Family_Listeriaceae	0.003134
Family_Bacteroidaceae	0.017123	Family_Coxiellaceae	0.002960
Tissue_Bloodstream	0.015716	Family_d	0.002894
Family_Francisellaceae	0.015146	Family_Legionellaceae	0.002248
Route_Vector	0.014902	Family_Leptotrichiaceae	0.001405
Family_Aeromonadaceae	0.014646	Family_Helicobacteraceae	0.000000
Family_Spirochaetaceae	0.013629		

Choosing the appropriate characteristics for these models is an essential step, but one that is usually disregarded. The results of our machine learning are hampered by the introduction of bias, which is caused by discarded data. The collective name for the sets of procedures known as "Feature Importance" methods is "Feature Importance" methods. These procedures assign a value to each feature that is used as an input to a model. These values are only a simple description of each feature's "importance." The relevance of a feature may be calculated by multiplying the degree to which a decrease in node impurity occurs by the probability of arriving at that node. The node probability may be calculated by taking the total number of samples and dividing that number by the number of samples that actually reach the node. If a quality has a higher value, then it must have a more significant impact.



These plots allow us to differentiate between severe and non-severe cases using the Random forest model, which was determined to be the best model based on the study that we carried out. Now that we have established the severity and severity scale, we can go on to the next step, which is predicting the virulence of bacterial pathogens. In this particular lime plot, we are keeping an eye on that explanation for severity. In both lime plots, the traits that are not considered to be severe are represented by the colour red, whereas severe aspects are shown by the colour green. The second plot, however, is based on the severity scale, which consists of four severity levels such as the least severe, the most severe, the most severe, and the least severe. As a result, we need the comprehensive forecast values as well as the plot for the characteristics of the data. Because of this, we decided to do a notebook lime plot for it, after which we had the specific prediction values and a visual representation.

At long last, the lime notebook experiment is providing us with the expected outcomes. The findings that we obtained from this plot are the prediction probabilities, and these probabilities are based on severity in the first plot and severity scale in the second plot. Non-severe cases have a prediction probability of measure of virulence of bacterial pathogen of 0.65, whereas severe cases have a probability of 0.35, depending on the severity.

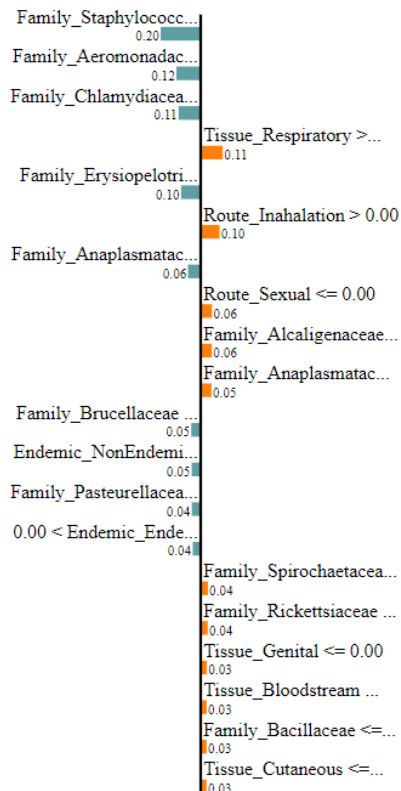


In this second plot, as we were doing the observations, we obtained the findings that we expected, which were four levels of severity according to the severity scale. The data shown in this figure demonstrates that severe and not severe, more severe and not more severe, most severe and not most severe, and least severe and not severe are all present. and 0.72 respectively.

Prediction probabilities

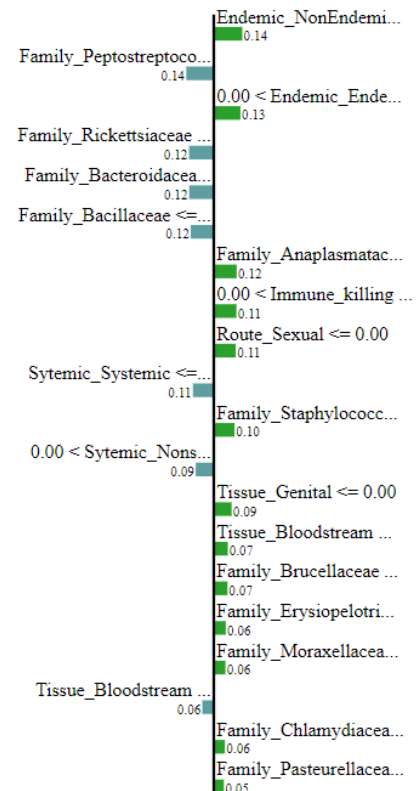
Leastsevere	0.00
Severe	0.72
Moresevere	0.28
Mostsevere	0.00

NOT Severe



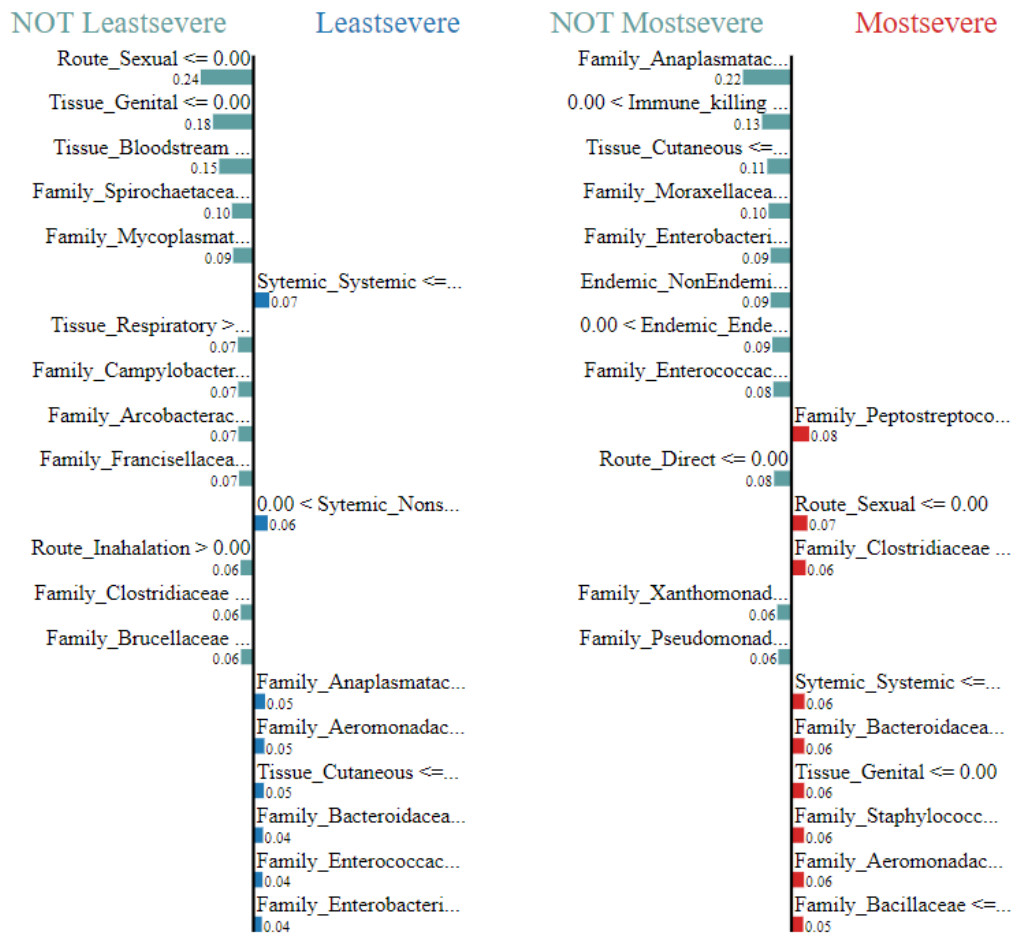
Severe

NOT Moresevere



Moresevere

It is exceedingly difficult to generate a plot that looks like that since it requires several outputs and classifications, which makes it impossible to have four levels in the same area at the same time. This makes simple observation impossible. Consequently, this is one of the potential stories that may be told utilising the lime notepad. Because the prediction probability, which is the result they requested, has extremely low values for both the least severe and the most severe outcomes, the plot displays the prediction probability as 0.00, while the prediction probability for the most severe and severe outcomes is 0.28



Applications

An essential part of fighting infectious illnesses is studying to predict the aggressiveness of bacterial infections. Predicting virulence may be used to inform the design of diagnostic and therapeutic approaches that aim to inhibit the production of virulence factors like toxins and adhesins. Clinicians may improve the success rate and decrease the likelihood of side effects by identifying particularly aggressive strains of bacteria and then developing individualised treatment strategies around those findings. Predicting the virulence of bacterial pathogens may also be used in microbial ecology to provide light on how host-pathogen interactions and virulence have evolved through time. Researchers can better combat the spread of infectious illnesses if they understand what drives the development and spread of particularly dangerous types. These are some examples of applications of bacterial pathogen virulence prediction.

- Forecasting the virulence of bacterial infections may aid in the creation of diagnostic tests for detecting high-risk strains that are more likely to cause severe illness. For instance, a prediction model for *Staphylococcus aureus* virulence was established by Abudahab et al. (2017), which might be utilised to create quick diagnostic tests for detecting high-risk strains.
- Treatment choices may be better informed if the virulence of bacterial infections can be predicted. Specifically, Yang et al. (2018) created a prediction model for *Streptococcus pneumoniae* virulence that may be used to detect high-risk strains and guide treatment choices, such as the selection of medicines.

- Evolution of pathogenicity studies may benefit from predicting the virulence of bacterial infections. McNally et al. (2016), for instance, used a genomic technique to predict the virulence potential of *Escherichia coli* strains and investigate the development of virulence in this species.


Discussion

Bacterial pathogen virulence prediction is a promising field of study with implications for clinical diagnostics, therapeutics, and public health. As we've seen, however, predictive models come with their fair share of caveats and unknowns. Our research demonstrates that the severity of disease is not consistently distributed across families of bacterial pathogens. Data availability and quality is a major hurdle in the modelling process. To build reliable prediction models, scientists need access to a wealth of information about bacterial pathogens, including genetic and phenotypic data, as well as clinical and epidemiological information. Feature selection, or determining which data points are most useful for making a virulence prediction, is another obstacle that must be overcome throughout the modelling process. Furthermore, risk factors of tissue tropism and, to a lesser extent, levels of human-to-human transmissibility and transmission route are better predictors of severe disease than taxonomy alone. Across all classification models, bacterial pathogens were more likely to be predicted to cause not much of a severe disease if they caused systemic infections, had neural or renal tropism, were transmitted via routes, or had a limited ability to transmit between humans. This was the case even if the pathogens had a limited ability to transmit between humans (SVM, random forest, linear regression, and gradient boost). Understanding the underlying biological principles of virulence, as well as using statistical and machine learning approaches, is essential for this complicated and iterative process. Yet, data availability and quality, as well as the selection of assessment measures, might place constraints on model evaluation. In conclusion, predicting the virulence of bacterial pathogens is a difficult and nuanced task that must take into account several constraints and uncertainties. To overcome these obstacles and create accurate and reliable prediction models that may be utilised to enhance clinical diagnosis, treatment, and public health, it is vital to employ a variety of methodologies and technologies. When subjected to a variety of modelling methods, these risk factors remained consistent throughout the testing process.

Limitation

The data that are presented is the primary source of concern in this investigation due to the fact that there are very less of them. The dataset must have at least more than 200 individual data rows. Therefore, doing the analytical component and obtaining the ROC curve might be quite challenging. Python and the R programming language are examples of the most fundamental coding platforms that we have available. Because of the limitations of these platforms, we are unable to do multioutput multiclassification analysis and prediction. According to the findings of our research, the Multioutput in the Multiclassification is a measure of the virulence of bacterial pathogens based on a severity scale with four severity levels. Predicting the virulence of bacterial infections has a number of caveats that must be taken into account before the findings of prediction algorithms can be trusted. Some of the restrictions, along with instances, are as follows:

- Insufficient data: Predictive models cannot provide reliable results without access to huge, varied datasets. On the other hand, certain bacterial diseases or strains may have little



information accessible. For instance, there may be insufficient genetic or clinical data to analyse for certain uncommon or developing infections.

- Complicated host-pathogen interactions: Complex interactions with the host immune system and other host variables determine the pathogenicity of bacterial infections. There is a chance that genetic data and other sorts of information don't adequately capture these relationships. It's possible that genomic data alone won't reveal that some bacterial infections use complex tactics to elude the host immune response.
- Incomplete knowledge of virulence factors: Our knowledge of the variables that contribute to the virulence of bacterial infections is both imperfect and constantly expanding. There is a possibility that predictive models do not take into consideration new or previously unknown virulence variables, which may have an impact on the accuracy of forecasts.
- Changes in evolution: Bacterial pathogens can change quickly and modify their surroundings, which can cause their virulence to change over time. Prediction models may be flawed if they fail to take into consideration these evolutionary shifts. The pathogenicity potential of some bacterial strains, for instance, may be altered if they acquired antibiotic resistance or other virulence characteristics through horizontal gene transfer.
- False positives and false negatives: Predictive models have the potential to provide findings that are false positive or false negative, either of which might result in an incorrect diagnosis or unsuitable therapy. Overreaction or inappropriate treatment might result, for instance, if a model incorrectly predicts that a particular bacterial strain is extremely pathogenic. On the other hand, a pathogen may be treated too slowly or not at all since the model predicts it has a low pathogenicity potential.

Also, writing such complex code on this platform and on our own PCs is not exactly the easiest thing in the world to accomplish. As a result of this investigation and the forecast that was obtained from the analysis, these are some of the major limitations that we encountered.

Conclusion

In conclusion, the study of bacterial pathogen virulence predictive models revealed that these models are effective in detecting possible virulence variables and making virulence predictions. Complex algorithms, data gathering, and feature selection methods are required for these models, along with a high level of computing skill. This study makes a contribution to the continuing efforts that are being made to measure the transmission of emerging infectious diseases and to draw comparisons between other models that already exist. Within the scope of this research, we provide an innovative strategy for ecological The virulence of bacterial pathogens may be anticipated using tropism and transmission ecology, and these models can be linked in more extensive settings with others, such as those that predict origin processes. The highlighted random forests are an excellent prediction model because they give helpful insight into the evolutionary determinants of virulence in developing diseases. According to the findings of our research, less dangerous bacterial infections are found in greater numbers than more dangerous ones. When compared to the very least number of most severe and least severe pathogens, the number of severe pathogens is quite large in the less severe bacterial pathogens. On the other hand, the number of more severe pathogens is relatively low. However, when we are talking about less severe cases, it drops to a very low level. Further work is required to address these limitations and create strong prediction models applicable to a broad range of bacterial pathogens.

Bibliographies

1. L. Brierley, A. B. Pedersen, & M. E. J. Woolhouse, Tissue tropism and transmission ecology predict virulence of human RNA viruses. *PLOS Biology*, **17** (2019) e3000206. <https://doi.org/10.1371/journal.pbio.3000206>.
2. S. Plotkin, History of vaccination. *Proceedings of the National Academy of Sciences*, **111** (2014) 12283–12287. <https://doi.org/10.1073/pnas.1400472111>.
3. Randall K. Holmes, Biology and Molecular Epidemiology of Diphtheria Toxin and the toxin gene. *The Journal of Infectious Diseases*, **181** (2000) S156–S167. <https://doi.org/10.1086/315554>.
4. C. Bonneaud & B. Longdon, Emerging pathogen evolution. *EMBO reports*, **21** (2020). <https://doi.org/10.15252/embr.202051374>.
5. B. Greenwood, The contribution of vaccination to global health: past, present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **369** (2014) 20130433–20130433. <https://doi.org/10.1098/rstb.2013.0433>.
6. L. H. Taylor, S. M. Latham, & M. E. J. Woolhouse, Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **356** (2001) 983–989. <https://doi.org/10.1098/rstb.2001.0888>.
7. M. Vouga & G. Greub, Emerging bacterial pathogens: the past and beyond. *Clinical Microbiology and Infection*, **22** (2016) 12–21. <https://doi.org/10.1016/j.cmi.2015.10.010>.
8. A. Casadevall & L. Pirofski, What Is a Host? Attributes of Individual Susceptibility. *Infection and Immunity*, **86** (2017). <https://doi.org/10.1128/iai.00636-17>.
9. M. Otto, Basis of Virulence in Community-Associated Methicillin-Resistant *Staphylococcus aureus*. *Annual Review of Microbiology*, **64** (2010) 143–162. <https://doi.org/10.1146/annurev.micro.112408.134309>.
10. S. M. Patel, M. A. Rahman, M. Mohasin, M. A. Riyadh, D. T. Leung, M. M. Alam, F. Chowdhury, A. I. Khan, A. A. Weil, A. Aktar, M. Nazim, R. C. LaRocque, E. T. Ryan, S. B. Calderwood, F. Qadri, & J. B. Harris, Memory B Cell Responses to *Vibrio cholerae* O1 Lipopolysaccharide Are Associated with Protection against Infection from Household Contacts of Patients with Cholera in Bangladesh. *Clinical and Vaccine Immunology*, **19** (2012) 842–848. <https://doi.org/10.1128/cvi.00037-12>.
11. A. J. Loughran, C. J. Orihuela, & E. I. Tuomanen, *Streptococcus pneumoniae*: Invasion and Inflammation. *Microbiology Spectrum*, **7** (2019). <https://doi.org/10.1128/microbiolspec.gpp3-0004-2018>.
12. F. Blanquart, M. K. Grabowski, J. Herbeck, F. Nalugoda, D. Serwadda, M. A. Eller, M. L. Robb, R. Gray, G. Kigozi, O. Laeyendecker, K. A. Lythgoe, G. Nakigozi, T. C. Quinn, S. J. Reynolds, M. J. Wawer, & C. Fraser, A transmission-virulence evolutionary trade-off explains attenuation of HIV-1 in Uganda. *eLife*, **5** (2016). <https://doi.org/10.7554/eLife.20492>.
13. D. T. Kenna, S. E. Darch, & M. J. Alston, The role of microbial competition and virulence in the evolution of *Pseudomonas aeruginosa*. (2017).
14. M. A. Acevedo, F. P. Dilleuth, A. J. Flick, M. J. Faldyn, & B. D. Elder, Virulence-driven trade-offs in disease transmission: A meta-analysis*. *Evolution*, **73** (2019) 636–647. <https://doi.org/10.1111/evo.13692>.

15. S. ALIZON, A. HURFORD, N. MIDEO, & M. VAN BAALEN, Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future. *Journal of Evolutionary Biology*, **22** (2008) 245–259. <https://doi.org/10.1111/j.1420-9101.2008.01658.x>.
16. R. M. Anderson & R. M. May, Population biology of infectious diseases: Part I. *Nature*, **280** (1979) 361–367. <https://doi.org/10.1038/280361a0>.
17. P. W. Ewald, The Evolutionary Ecology of Virulence. *The Quarterly Review of Biology*, **69** (1994) 381–384.
18. C. E. CRESSLER, D. V. MCLEOD, C. ROZINS, J. VAN DEN HOOGEN, & T. DAY, The adaptive evolution of virulence: a review of theoretical predictions and empirical tests. *Parasitology*, **143** (2015) 915–930. <https://doi.org/10.1017/s003118201500092x>.
19. D. Van Tyne & M. S. Gilmore, Friend Turned Foe: Evolution of Enterococcal Virulence and Antibiotic Resistance. *Annual Review of Microbiology*, **68** (2014) 337–356. <https://doi.org/10.1146/annurev-micro-091213-113003>.
20. T. J. Marrie, Virulence and pathogenesis of bacterial infections. (1995).
21. R. P. Novick, Pathogenicity Islands and Their Role in Staphylococcal Biology. *Microbiology Spectrum*, **7** (2019). <https://doi.org/10.1128/microbiolspec.gpp3-0062-2019>.
22. H. C. Leggett, C. K. Cornwallis, A. Buckling, & S. A. West, Growth rate, transmission mode and virulence in human pathogens. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **372** (2017) 20160094. <https://doi.org/10.1098/rstb.2016.0094>.
23. R. Chaudhuri & S. Ramachandran, Prediction of Virulence Factors Using Bioinformatics Approaches. *Methods in Molecular Biology*, (2014) 389–400. https://doi.org/10.1007/978-1-4939-1115-8_22.

Appendix code

```
import pandas as pd
import numpy as np
```

```
pd.set_option('display.max_rows', 1000)
pd.set_option('display.max_columns', 1000)
pd.set_option('display.width', 5000)
```

Import Datasets

```
bact=pd.read_excel("C:/Users/KANMANIVISHWAA/Downloads/ML_Dataset_Oct22.xlsx")
bact
```

```
bact = bact.drop(labels=0,axis=0)
bact
```

Check Value Count of Row Family

```
bact['Family'].value_counts()
```

Adding new column using family value count

```
bact1 = pd.DataFrame(bact, columns=["Severity", "Family","No.bacterial pathogen"])
bact1
```

Numerical to Categorical

```
replace_dict= {0:'Nonsevere',1:'Severe'}
print(replace_dict)
```

```
bact1['Severity']=bact['Severity'].map(replace_dict)
bact1
```

```
bact2=pd.crosstab(bact1.Family,bact1.Severity)
bact2
```

```
sum(bact2.Nonsevere)
sum(bact2.Severe)
```

Bar Plot

```
import matplotlib.pyplot as plt
bact2.plot(kind='bar', stacked=True,figsize=(19.5,6.5))
plt.title("Measure of Virulence 1")
plt.xlabel("Family")
plt.ylabel("No.Bacterial Species")

bact3 = pd.DataFrame(bact, columns=["Severity_Scale",
"Family","No.bacterial pathogen"])
bact3
```

Numerical to categorical

```
replace_dict1= {1:'Mostsevere',2:'MoreSevere',3:'Severe',4:'LeastSevere'}
print(replace_dict1)
```

```
bact3['Severity_Scale']=bact['Severity_Scale'].map(replace_dict1)
bact3
```

```
bact4=pd.crosstab(bact3.Family,bact3.Severity_Scale)
```

```
bact4
```

```
sum(bact4.LeastSevere)
```

```
sum(bact4.Severe)
```

```
sum(bact4.MoreSevere)
```

```
sum(bact4.Mostsevere)
```

Box Plot

```
bact4.plot(kind='bar', stacked=True,figsize=(19.5,6.5))
```

```
plt.title("Measure of Virulence 2")
```

```
plt.xlabel("Family ")
```

```
plt.ylabel("No.Bacterial Species")
```

Numerical to Categorical

```
replace_dict2= {0:'NonEndemic',1:'Endemic'}
```

```
print(replace_dict2)
```

```
replace_dict3= {0:'Nonsystemic',1:'Systemic'}
```

```
print(replace_dict3)
```

```
replace_dict4= {0:'Nontransmissable',1:'Transmissable'}
```

```
print(replace_dict4)
```

```
replace_dict5= {0:'Motile',1:'Nonmotile'}
```

```
print(replace_dict5)
```

```
bact['Severity']=bact['Severity'].map(replace_dict)
```

```
bact['Severity_Scale']=bact['Severity_Scale'].map(replace_dict1)
```

```
bact['Endemic']=bact['Endemic'].map(replace_dict2)
```

```
bact['Sytemic']=bact['Sytemic'].map(replace_dict3)
```

```
bact['Transmissibility']=bact['Transmissibility'].map(replace_dict4)
```

```
bact['Motility']=bact['Motility'].map(replace_dict5)
```

```
bact
```

Checking Na value

```
print(bact.isnull().sum())
```

Remove na value using interpolating

```
bact_1 = bact.interpolate()
```

```
bact_new = bact_1.apply(lambda x: x.fillna(x.value_counts().index[0]))
```

```
bact_new
```

Checking Na Value

```
print(bact_new.isnull().sum())
```

Upsampling

```
from sklearn.utils import resample
```

```
bact_new_upsampled= resample(bact_new,
```

```
        replace=True,
        n_samples=250,
        random_state=21)
bact_new_upsampled
```

MEASURE OF VIRULENCE 1

#split dataset in features and target variable

```
feature_cols = ['Family', 'Endemic', 'Sytemic', 'Route','Tissue',
'Transmissibility','Immune_survival','Immune_killing']
X = bact_new_upsampled[feature_cols] # Features
y = bact_new_upsampled.Severity # Target variable
```

Encoding x and y

```
X_encoded = pd.get_dummies(X)
X_encoded
```

```
from sklearn.preprocessing import LabelEncoder
```

```
encoder = LabelEncoder()
```

apply on df

```
y_encoded1 = encoder.fit_transform(y)
y_encoded1
```

```
from sklearn.model_selection import train_test_split
```

Split dataset into training set and test set

```
X_train, X_test, y_train, y_test = train_test_split(X_encoded, y_encoded1,
test_size=0.33,random_state=21)
```

Model Building

Model 1

```
from sklearn import svm
```

```
SV1 = svm.SVC(probability=True)
```

Train Decision Tree Classifier

```
SV1.fit(X_train, y_train)
```

```
pred_y1_sv = SV1.predict(X_test)
```

```
pred_y1 = SV1.predict_proba(X_test)[: ,1]
```

auc scores

```
from sklearn.metrics import
```

```
roc_auc_score,classification_report,confusion_matrix
```

```
auc_score1 = roc_auc_score(y_test, pred_y1_sv)
```


```
auc_score1
```

```
cr1 = classification_report(y_test, pred_y1_sv)
```

```
print('cr1:',cr1)
```

Model 2

```
from sklearn.ensemble import RandomForestClassifier
```



```
RF1 = RandomForestClassifier()
RF1.fit(X_train, y_train)

pred_y2_rf = RF1.predict(X_test)
#Predict the response for test dataset
pred_y2 = RF1.predict_proba(X_test)[: ,1]
# auc scores
from sklearn.metrics import roc_auc_score
auc_score2 = roc_auc_score(y_test, pred_y2)
auc_score2
cr2 = classification_report(y_test, pred_y2_rf)
print('cr2:',cr2)
```

Model 3

```
from sklearn.linear_model import LogisticRegression
LR1= LogisticRegression()
LR1.fit(X_train,y_train)
pred_y3_lr = LR1.predict(X_test)
#Predict the response for test dataset
pred_y3 = LR1.predict_proba(X_test)[: ,1]
# auc scores
from sklearn.metrics import roc_auc_score
auc_score3 = roc_auc_score(y_test, pred_y3)
auc_score3
cr3 = classification_report(y_test, pred_y3_lr)
print('cr3:',cr3)
```

Model4

```
from sklearn.ensemble import GradientBoostingClassifier
GB1 = GradientBoostingClassifier()
GB1.fit(X_train,y_train)
pred_y4_gb = GB1.predict(X_test)
#Predict the response for test dataset
pred_y4 = GB1.predict_proba(X_test)[: ,1]
# auc scores
from sklearn.metrics import roc_auc_score
auc_score4 = roc_auc_score(y_test, pred_y4)
auc_score4

cr4 = classification_report(y_test, pred_y4_gb)
print('cr4:',cr4)
```

Box plot

```
from sklearn import model_selection
# prepare models
models = []
models.append(('SV1', svm.SVC()))
models.append(('RF1', RandomForestClassifier()))
models.append(('LR1', LogisticRegression()))
models.append(('GB1', GradientBoostingClassifier()))

# evaluate each model in turn
results = []
names = []
scoring = 'accuracy'
```

```

for name, model in models:
    kfold = model_selection.KFold(n_splits=10)
    cv_results = model_selection.cross_val_score(model, X_test, y_test,
cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
# boxplot algorithm comparison
fig = plt.figure()
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()

```

AUC Curve

```

from matplotlib import pyplot
from sklearn import metrics
# calculate roc curves
sv_fpr, sv_tpr, _ = metrics.roc_curve(y_test, pred_y1)
rf_fpr, rf_tpr, _ = metrics.roc_curve(y_test, pred_y2)
lr_fpr, lr_tpr, _ = metrics.roc_curve(y_test, pred_y3)
gb_fpr, gb_tpr, _ = metrics.roc_curve(y_test, pred_y4)
# plot the roc curve for the model
pyplot.plot(sv_fpr, sv_tpr, linestyle='--', label='SVM')
pyplot.plot(rf_fpr, rf_tpr, linestyle='--', label='Randomforest')
pyplot.plot(lr_fpr, lr_tpr, linestyle='--', label='Logisticreg')
pyplot.plot(gb_fpr, gb_tpr, linestyle='--', label='Gradientboost')
# axis labels
pyplot.xlabel('False Positive Rate')
pyplot.ylabel('True Positive Rate')
# show the legend
pyplot.legend()
# show the plot
pyplot.show()

```

Feature importance score and its box plot

```

feature_scores = pd.Series(RF1.feature_importances_,
index=X_train.columns).sort_values(ascending=False)

feature_scores

import seaborn as sns
f, ax = plt.subplots(figsize=(20, 20))
ax = sns.barplot(x=feature_scores, y=feature_scores.index, data=bact_new)
ax.set_title("Visualize feature scores of the features")
ax.set_yticklabels(feature_scores.index)
ax.set_xlabel("Feature importance score")
ax.set_ylabel("Features")
fig.tight_layout()
plt.show()

```

MEASURE OF VIRULENCE 2

#split dataset in features and target variable

```

feature_cols2 = [ 'Family','Endemic', 'Sytemic', 'Route','Tissue',
'Transmissibility','Immune_survival','Immune_killing']
X2 = bact_new_upsampled[feature_cols2] # Features
y2 = bact_new_upsampled.Severity_Scale # Target variable

Encoding Y and x
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()
# apply on df
y_encoded2 = encoder.fit_transform(y2)
y_encoded2

X2_encoded = pd.get_dummies(X2)
X2_encoded

from sklearn.model_selection import train_test_split
# Split dataset into training set and test set
X_train2, X_test2, y_train2, y_test2 = train_test_split(X2_encoded, y2,
test_size=0.33, random_state=21)

```

Model Building

Model 5

```

from sklearn import svm
SV2 = svm.SVC(probability=True)

# Train Decision Tree Classifier
SV2.fit(X_train2, y_train2)

pred_y1_sv1 = SV2.predict(X_test2)
#Predict the response for test dataset
y_pred_proba1=SV2.predict_proba(X_test2)

#roc_auc
auc_score_1= roc_auc_score(y_test2, y_pred_proba1,multi_class='ovr')
auc_score_1

cr5 = classification_report(y_test2, pred_y1_sv1)
print('cr5:',cr5)

```

Model 6

```

RF2 = RandomForestClassifier()
RF2.fit(X_train2, y_train2)
pred_y2_rf1 = RF2.predict(X_test)
#Predict the response for test dataset
y_pred_proba2=RF2.predict_proba(X_test2)
#roc_auc
auc_score_2 = roc_auc_score(y_test2, y_pred_proba2,multi_class='ovr')
auc_score_2

cr6 = classification_report(y_test2, pred_y2_rf1)
print('cr6:',cr6)

```


Model 7

```
from sklearn.linear_model import LogisticRegression
LR2= LogisticRegression()
LR2.fit(X_train2,y_train2)
pred_y3_lr1 = LR2.predict(X_test2)
#Predict the response for test dataset
y_pred_proba3=LR2.predict_proba(X_test2)
#roc_auc
auc_score_3 = roc_auc_score(y_test2, y_pred_proba3,multi_class='ovr')
auc_score_3

cr7 = classification_report(y_test2, pred_y3_lr1)
print('cr7:',cr7)
```

Model 8


```
from sklearn.ensemble import GradientBoostingClassifier
GB2 = GradientBoostingClassifier()
GB2.fit(X_train2,y_train2)
pred_y4_gbl = GB2.predict(X_test2)
#Predict the response for test dataset
y_pred_proba4 = GB2.predict_proba(X_test2)
# auc scores
from sklearn.metrics import roc_auc_score
auc_score_4 = roc_auc_score(y_test2, y_pred_proba4,multi_class='ovr')
auc_score_4

cr8 = classification_report(y_test2, pred_y4_gbl)
print('cr8:',cr8)
```

Box Plot

```
from sklearn import model_selection
# prepare models
models = []
models.append(('SV2', svm.SVC()))
models.append(('RF2', RandomForestClassifier()))
models.append(('LR2', LogisticRegression()))
models.append(('GB2', GradientBoostingClassifier()))

# evaluate each model in turn
results = []
names = []
scoring = 'accuracy'
for name, model in models:
    kfold = model_selection.KFold(n_splits=10)
    cv_results = model_selection.cross_val_score(model, X_test2, y_test2,
cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
# boxplot algorithm comparison
fig = plt.figure()
fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
```



```
plt.show()
```

Feature importance score and its box plot

```
feature_scores = pd.Series(RF2.feature_importances_,
index=X_train.columns).sort_values(ascending=False)
feature_scores
```

```
import seaborn as sns
f, ax = plt.subplots(figsize=(20, 20))
ax = sns.barplot(x=feature_scores, y=feature_scores.index, data=bact_new)
ax.set_title("Visualize feature scores of the features")
ax.set_yticklabels(feature_scores.index)
ax.set_xlabel("Feature importance score")
ax.set_ylabel("Features")
plt.show()
```

Lime function for Measureof virulence 1

```
num=1
test_sample=X_test.iloc[num,:]
```

```
import lime
from lime import lime_tabular
explainer1 = lime_tabular.LimeTabularExplainer(
    training_data=np.array(X_train),
    feature_names=X_train.columns.values,
    class_names=['Nonsevere', 'Severe'],
    mode='classification',
    verbose=True,
    random_state=21)
expl1 = explainer1.explain_instance(
    data_row=test_sample,
    predict_fn=RF1.predict_proba,num_features=20)
```

Lime notebook plot

```
plt=expl1.as_pyplot_figure()
plt.tight_layout()
```


Detailed lime plot

```
expl1.show_in_notebook(show_table=True, show_all=False)
```

Lime function for Measureof virulence 2

```
num>1
test_sample_2=X_test.iloc[num,:]

import lime
from lime import lime_tabular
explainer2 = lime_tabular.LimeTabularExplainer(
    training_data=np.array(X_train2),
    feature_names=X_train2.columns.values,
    class_names=['Leastsevere','Severe', 'Moresevere','Mostsevere'],
    mode='classification',
    verbose=True,
```



```
random_state=21)
exp2 = explainer2.explain_instance(
    data_row=test_sample_2,
    predict_fn=RF2.predict_proba,num_features=20,top_labels=4)
```

Notebook lime plot

```
plt=exp2.as_pyplot_figure()
```

Detailed lime plot

```
exp2.show_in_notebook(show_table=True, show_all=False)
```