CHAPTER 3

REVIEW OF LITERATURE

3.1 Importance of disease forecasting

3.1.1 Requisites for an effective forecasting of communicable diseases

World Health Organization’s (WHO) regional committee for the Eastern Mediterranean has discussed various aspects of communicable disease forecasting, including approaches to forecasting, uses and limitations of forecasting, prerequisites for forecasting and steps for effective forecasting in its technical paper on ‘Forecasting in communicable diseases’ (1999). The committee is of the opinion that a comprehensive and feasible strategic plan is needed to address the problem of communicable diseases and that future health scenarios of communicable diseases that can predict likely, probable or even merely possible trends i.e., forecasting can play an important role in developing such a strategic plan. It suggested that forecasting should be regarded as a tool to increase understanding and highlight important processes. The paper discussed several issues that need to be well defined before developing a meaningful forecast in communicable diseases such as, availability of historical data and the accuracy of such data, proper understanding of the causes of changes in patterns in the past and determination of the factors that might effect changes in the pattern of the communicable diseases in future. The committee opines that forecasting plays key role in health services planning and the disease prevention programmes require some knowledge of the future burden of disease in order to guide the health policy process and that decisions are to be taken in advance in order to cope with the expected disease burden in the future. They advocate that forecasting helps in preventing unusual occurrence of communicable diseases or at least in being prepared for it so that it can be controlled efficiently.

While discussing prerequisites for forecasting, the authors convey that the important requisites are the basic factors determining the epidemiology of the disease and the past experience with that disease in a population. Summary of the requisite factors are as follows.

- Availability of historical data about the communicable disease and the accuracy of such data:
Sufficient historical data are required to reveal significant patterns of changes by time, such as long-term trend, seasonal and cyclical fluctuations. There are different sources for collecting data, of which, disease surveillance is an important tool to provide the data from which forecasting can be done. The sources of data are historical records, surveys and longitudinal studies.

- Data integrity and accuracy:
  Another requisite for good forecasting is integrity and accuracy of the data. This includes transparency, consistency and completeness in data reporting. Data reporting is critical in developing good forecasts and enabling appropriate steps to be taken to plan for the future.

- Proper understanding of the causes of changes in the past:
  According to the committee, the next requisite is that the changes reflected in the trend or seasonal or cyclical or irregular fluctuations detected in a communicable disease time series should be explained by identifying the epidemiological factors responsible for such changes and that it will entail identification of the agent-related, host and environmental factors that may be shaping the observed past pattern of the communicable disease time series.

- Factors that might effect change in the patterns of communicable diseases in the future:
  Another requisite expressed by the committee is that proper understanding of the relevance of any of the factors and their possible role in shaping the epidemiology of communicable diseases at the regional and national levels requires continuous monitoring using proper epidemiological approaches and indicators.

The committee further condensed various steps involved in effective forecasting as

- Defining the scope of forecasting
- Identifying the variables of interest
- Checking reliability of data and preparing it for analysis
- Analysing the data
- Selecting one or more appropriate forecasting method(s)
- Applying the selected forecasting method(s) to determine tentative forecasts
- Comparing the results of the forecasting methods
- Determining the final forecast
Finally, the committee concluded that to develop a more comprehensive strategic plan to address the problem of communicable diseases, information is required about the future outlook of priority communicable diseases so that decisions and choices can be taken to cope with such diseases and that forecasting is the process by which such information can be obtained, provided the prerequisites and the steps for effective forecasting are met.

The Committee further inferred that forecasting in communicable diseases necessitates good epidemiological knowledge and practice on one hand and a functional surveillance system on the other. The paper also mentions that statistical and mathematical expertise is needed to develop the models.

The committee recommended that the WHO member states with the support of WHO should continue to improve their disease surveillance systems by implementing the regional strategy for development and strengthening of epidemiological surveillance and introduce the concept of forecasting for priority communicable diseases.

3.1.2 Early Warning System (EWS)

Early Warning system is providing effective response at the right time through the recognized institutions that allows persons exposed to risks to take actions to avoid or reduce risk and prepare for an effective response (Grasso and Singh, 2011).

Kshirsagar et al. (2013) have analysed the importance of Global Early Warning System (GLEWS) and its use in forecasting outbreak of animal diseases in India. According to them, Early Warning and Response (EWS) is based on the concept of dealing with a disease epidemic in its early stages and concluded that EWS will reduce the morbidity and mortality due to a disease and help to take up preventive measures for its control and eradication.

The International organizations and their functionalities are highlighted in the following sections.

- Global early warning and response system (GLEWS) of World Organisation of Animal Health (OIE):
- The GLEWS is the joint OIE-Food and Agriculture Organisation (FAO)-WHO system that assists in prediction, prevention and control of animal disease threats
through sharing of information, epidemiological analysis and joint field missions to assess and control the outbreak, whenever needed.

Aims of GLEWS:

- Better detection of epidemiological episodes at country level
- Better transparency among countries reporting to OIE
- Improving the quality of field animal health information
- Strengthening the system between veterinary & medical laboratories
- Provide fast and efficient aid to the affected countries

- Early warning system developed by Project Directorate on Animal Disease Monitoring and Surveillance (PD_ADMAS):

PD_ADMAS was established in 1987 in India to develop a system for national disease monitoring and surveillance of economically important livestock diseases with a mandate to carry out surveillance and monitoring of major livestock diseases in the country with an aim to design the strategic control measures for their control.

Animal Disease Surveillance is a key for improving disease analysis, early warning and prevention of the spread of diseases. PD_ADMAS is now known as ICAR- National Institute of Veterinary Epidemiology and Disease Informatics (NIVEDI). It has developed National Animal Disease Referral Expert System (NADRES), a web-based relational software for forecasting of major animal diseases two months in advance so that effective measures can be taken up for prevention and/or control of the diseases. It uses weather parameters as risk factors for the forecast of livestock diseases at district level. NIVEDI has established a good network of state animal husbandry departments in India to carry out surveillance of animal diseases and guides important national policy decisions in animal health. Successful implementation of EWS is dependent on efficacy of national disease surveillance program, degree of awareness among field veterinarians, technicians, extension specialists and farmers about clinical and epidemiological features of diseases.

Finally, the authors concluded that early detection and response provide better preparedness for effective control and containment of disease outbreaks.
3.2 Epidemiological models for disease forecasting

Dubé et al. (2007), while discussing the use of epidemiological models for the management of animal diseases felt that modelling is a widely used tool to support the evaluation of various disease management activities and the value of epidemiological models lies in their ability to study ‘what if’ scenarios and provide decision-makers with a prior knowledge of consequence of disease incursions and impact of control strategies. Further, they were of the opinion that the modelling would be most useful when applied to pre-outbreak, particularly in the areas of retrospective analysis of previous outbreaks, contingency planning, resource planning, risk assessment and training.

Huppert et al. (2013) observed that an important role of modelling enterprises is that they can alert the researcher about the deficiencies in current understanding of the epidemiology of various infectious diseases, and raise crucial questions for extensive investigation and additional data that need to be collected.

Siettos et al. (2013), while reviewing ‘Mathematical modelling of infectious disease dynamics’ presented key modelling methods used for the surveillance and forecasting of infectious disease outbreaks. They categorized epidemiological models into three classes viz., statistical, mathematical-mechanistic state space, and machine-learning based models. They suggested that public health organisations throughout the world to use such models to evaluate and develop intervention disease outbreak policies for ever-emerging epidemics. They stated that simulation allows for rapid assessment and decision making. It provided quantification and insight into the spatiotemporal dynamics of spread of the disease. Further they opined that an intensive inter and multidisciplinary research is speeding up the developments in the field, integrating advances from epidemiology, molecular biology, computational engineering, science, applied mathematics and sociology.

Garrett (2012) demonstrated that it is possible to predict future disease burden using Box-Jenkins forecasting techniques and the overall accuracy of the forecast and disproportionate number of forecast values contained within the first forecast interval validate this as a method that may be used to monitor disease trends and potentially facilitate the early identification of an outbreak through a study on Box-Jenkins modelling techniques which can forecast future disease burden and identify disease aberrations. Their results of the study indicated that it is possible to predict future disease burdens within a community based upon the historical
surveillance data within local health jurisdictions of varying size using Box-Jenkins forecasting techniques.

3.3 Modelling on anthrax

Forecasting the outbreak of diseases by modelling is very helpful in saving the livestock population and thereby averting economic loss to the livestock farmer in particular and country in general. The ability for early detection of the outbreaks is important to minimize the morbidity and mortality through timely implementation of disease prevention and control measures.

3.3.1 Ecological niche modelling with the Genetic Algorithm for Rule-set Production

Blackburn (2006) explored the spatial ecology and potential pathways of infection of anthrax in North America by adopting a multi-scale approach to evaluate the components required for the survival of anthrax *bacillus* in the environment, interactions with wildlife, and the potential role that vectors play in anthrax transmission. He used ecological niche modelling with the Genetic Algorithm for Rule-set Production (GARP) to predict the geographic distribution of anthrax in the continental United States of America, using case data from outbreaks between 1957 and 2005. He considered temperature, precipitation, elevation, soil moisture, soil pH and NDVI as risk factors. The results obtained were applied to produce the first quantitative, continental scale predictions of anthrax in Mexico. The study provided the estimates for anthrax distributions in United States of America and Mexico. Performance of the model was evaluated using Area under the Receiver Operating Characteristic (ROC) curve.

Joyner (2010) developed Ecological niche model with the Genetic Algorithm for Rule-set Prediction (GARP) in the model-building process, which can predict the current and future potential distributions of *Bacillus anthracis* in Kazakhstan using annual mean temperature, annual temperature range, annual precipitation, precipitation of wettest month, precipitation of driest month, elevation and normalized difference vegetation index (NDVI) as predictors. He concluded that the output produced by GARP provided a better understanding of how GARP constructs rule-sets and projects presence and absence onto the topography and the output was significant in understanding the geographical and ecological space where *B. anthracis* exist in Kazakhstan. He suggested that other modelling approaches also identify
environmental ranges and should be explored extensively in future studies. Performance of the model was evaluated using Area under the Receiver Operating Characteristic (ROC) curve.

Mullins et al. (2013) modelled the ecological niche of *Bacillus anthracis* in United States of America, Italy and Kazakhstan to understand the geographic distribution of anthrax and its potential association between regional populations and ecology. They developed country-specific ecological-niche models with the Genetic Algorithm for Rule-set Production (GARP) using environmental variables, i.e., elevation, annual temperature range, annual mean temperature, precipitation of driest month, precipitation of wettest month, annual precipitation and normalized difference vegetation index (NDVI) and transferred to other countries to determine if pathogen presence could be accurately predicted on novel landscapes. They found that native models accurately predicted endemic areas within each country, but transferred models failed to predict known occurrences in the outside countries. They suggested that models should be developed regionally, on the native landscape, and with consideration to population genetics. Performance of the model was evaluated using Area under the Receiver Operating Characteristic (ROC) curve.

Steenkamp (2013) developed Species Distribution model for anthrax in the Kruger national park, South Africa. The aim of the study was to identify and map areas within the park that were ecologically suitable for harbouring of *B. anthracis* spores in the soil. This was achieved using maximum entropy as a statistical model, with environmental variables, NDVI, elevation, distance of positive locations from dams, pans, rivers, springs, troughs, bore holes and water holes, soil data and vegetation data as predictors. He recommended that future study should compare the results of the Maxent model with the results of other modelling techniques. Performance of the model was evaluated using Area under the Receiver Operating Characteristic (ROC) curve.

### 3.3.2 Multiple logistic regression models

Gunaseelan et al. (2011) analysed the distribution of the animal anthrax in Tamil Nadu state from 1991 to 2006 using Multiple logistic regression to assess the association between the three effector variables and the occurrence of >25 outbreaks of anthrax. The variables used were population of cattle per district, percent of villages with alkaline soil and presence of a tannery to estimate the odds of experiencing > 25 anthrax outbreaks. They observed that the
only variable with statistical significance was the percent of villages with alkaline soil per
district.

3.3.3 Clinician Detection model

Adamou et al. (2006) developed a probabilistic model which explained how clinicians detect
a disease outbreak due to an outdoor release of anthrax spores, when the clinicians have
access to traditional clinical information only. Through the Clinician Detection model, they
estimated that an outdoor release of anthrax is expected to be detected by clinicians within
nine days of the release. They concluded that estimated clinician detection time can be used
in planning for a worst-case scenario of a bioterrorist induced anthrax release.

3.3.4 Modelling response to anthrax infection

Jamrog et al. (2007), while studying modelling responses to anthrax and smallpox attacks,
examined the impact on fatalities due to a delay in treatment in the event of a large-scale
anthrax attack, like the one that would kill 50,000 people if a public health response was not
mounted. They proposed that reducing the number of potential deaths by 90% requires public
health officials to start distributing antibiotics within two to three days of the attack and
should complete the distribution of antibiotics to the entire population within three days.
Further they expressed that the goal would be impossible to achieve if the first report of the
attack of the disease came from medical diagnosis as it would not occur until some four to six
days after the attack. So, they declared that time is very important and crucial in the event of
an anthrax attack.

3.3.5 A mathematical simulation of the inflammatory response to anthrax infection

Kumar et al. (2008) have suggested through a mathematical simulation model of the
inflammatory response to anthrax infection that an optimal strategy would require both
vaccination and antibiotic administration.

3.3.6 Dose-Response Model

Bartrand et al. (2008) reviewed and analysed Dose-response data from the open literature for
inhalation exposure to Bacillus anthracis to develop Dose-Response models for inhalation of
Bacillus anthracis spores. They identified, via dose-response modelling, that the exponential
model and the Beta-Poisson dose-response model provided the best fit for the data on inhalation exposure to *Bacillus anthracis*. They observed that all dose-response data showed distinct dose-response behavior i.e., low dispersion and apparent low host-to-host variation in susceptibility. They fitted Exponential, Beta Poisson, and Log-Probit dose-response models.

### 3.3.7 Modelling of Inhalational anthrax

Gutting *et al.* (2008), while studying mathematical modelling of inhalational anthrax, observed that in spite of the continuing threat of *Bacillus anthracis* and intensive research on this subject, experts face limited options when assessing the risk of inhalational anthrax. They arrived at a judgement that a single or combination of models might enable experts to predict valid dose-response relationships for inhalational anthrax over pertinent exposure ranges.

### 3.3.8 Modelling for impact of distributing MedKits

Houck *et al.* (2011) developed a model that could estimate the impact of deploying Med Kits in a community and use that model to consider some possible scenarios to get some insight into the effectiveness of the Med Kit strategy in preparing for an anthrax attack.

### 3.3.9 Modelling the host response to inhalation anthrax

Day *et al.* (2011), while modelling the host response to inhalation anthrax, explored the probability of survival of an infected individual. They were of the view that this probability depended on several factors *viz.*, quantity of spores inhaled, rate at which the spores migrate from alveolar spaces to tracheobronchial lymph nodes and germinate, the number of bacteria released from the lysed host cell and the timing of the treatment. Accordingly, their model included the toxin released by the bacteria, the early immune response in the thoracic/mediastinal lymph nodes (TMLN) and the recruitment of neutrophils.

### 3.4 Modelling on other animal diseases

#### 3.4.1 Mathematical Modelling and rank analysis

Singh and Prasad (2008) analysed the economic losses due to diseases and estimated in terms of losses due to mortality, milk yield, body weight and opportunity cost through mathematical modelling of economic losses due to some important diseases in goats in India.
They computed the losses for a 15-year period (1991 - 2005) based on published reports. They ranked seven important diseases with respect to their incidence, deaths and economic losses through rank analysis. They suggested that necessary preventive measures are required to control these diseases of goats.

3.4.2 Combination of statistical models

Ensoy (2013) developed forecast and prediction models applied to the field of demography and infectious animal diseases. Firstly, he observed the influence of animal movements on spread of Bluetongue (BT virus - 8) during the year 2006 in Belgium. Secondly, utilizing the univariate and bivariate prediction models, he developed a model for outbreak detection and prediction and lastly he put forth demographic forecasting of the disease in Belgian population. He established "hot spots" in time and space in terms of movement and the potential for disease spread using weighted negative binomial model. He used a spatio-temporal dynamic model to explore the contribution of different factors to the spreading of the disease. He modelled the probability of a farm being infected using a Binomial model.

3.4.3 Mathematical models for controlling equine influenza

Daly et al. (2013) have brought together key findings from various modelling studies conducted over the past 10 years that are relevant to the disease equine influenza. They elucidated that modelling has particular value in predicting the likely outcome at the population level of implementing different control measures and is increasingly considered an important tool for development of disease prevention and control measures. They were of the opinion that the conclusions drawn from mathematical models must always be critically evaluated as there may be limited data available to test the performance of a model and preconceived ideas may lead to conceptual errors in the construction of a model. They concluded that studies on mathematical modelling do not replace epidemiological studies and animal experiments, but augment them.

3.5 Modelling on human diseases

3.5.1 Multiple Linear Regression and remote sensing

Devi et al. (2003) used Multiple Linear Regression (backward elimination) method to analyse the environmental factors related to malarial incidences. They used remote sensing and
geographic information system (GIS) for monitoring the environmental factors associated with the vector-borne disease, malaria. They conducted an epidemiological and ecological study to determine sustenance of malarial incidences in Salem district of Tamil Nadu, India, considering vegetation cover, rainfall, water body, temperature and humidity as predictors. Their results showed that temperature, water body and the interaction of rainfall and forest cover played an important role in the spread of the disease and concluded that even if the environmental conditions do not favor the spread of the disease, other factors such as sociological and health factors viz., lack of awareness among public and no implementation of precautionary methods play important roles. They suggested that management strategies must be very effectively implemented in these areas.

3.5.2 Multi Step Polynomial Regression

Chatterjee et al. (2009) developed a non linear regression modelling methodology for forecasting incidence of malaria in Chennai, India and predicted future disease incidence using multi-step polynomial regression model. They considered three types of data to develop the regression methodology i.e., a longer time series data of Slide Positivity Rates (SPR) of malaria, a smaller time series data (deaths due to Plasmodium vivax) of one year and spatial data (zonal distribution of Plasmodium vivax deaths) for the city along with the climatic factors, viz., minimum temperature, maximum temperature, minimum humidity, maximum humidity, and rainfall; population and previous incidence of the disease. They performed variable selection by simple correlation study, identification of the initial relationship between variables through non linear curve fitting and used multistep methods for induction of variables in the non linear regression analysis along with applied Gauss-Markov models and ANOVA for testing the prediction, validation and construction of the confidence intervals. Their results showed the applicability of the method for different types of data, the autoregressive nature of forecasting and high prediction power for both SPR and Plasmodium vivax deaths, where one-lag SPR values plays an influential role and proves useful for better prediction. Different climatic factors were identified as playing crucial role on shaping the disease curve. They concluded that with utilizing the readily available excellent models of climatic forecasts in conjunction with their method, one can predict the disease incidence at long forecasting horizons with high degree of efficiency. They were of the opinion that such technique will be a useful early warning system at region or nation level to take up disease prevention and control measures.
In another study, Tuyishimire (2013) identified and mapped factors responsible for causing malaria in Ruhuha sector, Rwanda, through the use of remote sensing, GIS and spatial statistics. He assessed the relationship between malaria prevalence and its causing factors using logistic regression. He considered a number of environmental factors such as altitude, land use and distance to anopheles mosquito breeding sites; demographic factors such as household size, age, gender and distance to household with infected people and economic factors such as house material and animal ownership as malaria causing factors to build the model. He applied Analysis of Variance (ANOVA) and Chi Square tests to assess the possible differences between malaria clusters. The study revealed that infection of malaria increases with closeness to irrigated farmland and to houses with infected people. He also observed that more incidences were noticed with increase in household size and in houses made of mud. The study also suggested that malaria infection is higher in earth floor houses as compared to cement floor houses.

3.5.3 Principal component regression

Nizamuddin et al. (2013) used Advanced Very High Resolution Radiometer (NOAA / AVHRR) environmental satellite data to produce weather seasonal forecasts for predicting malaria epidemics in Tripura state in India, which has one of the highest endemic cases of malaria in the country. They developed an algorithm that uses Vegetation Health Indices (Vegetation Condition Index (VCI) and Temperature Condition Index (TCI)) computed from Advance Very High Resolution Radiometer (AVHRR) from National Oceanic and Atmospheric Administration (NOAA) afternoon polar orbiting satellite. They identified a significant relationship between satellite data and annual malaria incidences at least three months before the major malaria transmission period. The statistical technique, principal component regression (PCR) was used to develop a model to predict malaria as a function of the TCI. Then the results were compared with observed malaria statistics which showed that the error of estimations of malaria is insignificant. They inferred that optical remote sensing is a variable tool to estimate malaria well in advance so that preventive measures can be taken.

3.5.4 Time series analysis

Kumar et al. (2014) used Autoregressive Integrated moving Average (ARIMA) model for forecasting malaria cases using climatic factors in Delhi, India. They fitted the model using
covariates rainfall, relative humidity, mean maximum temperature and taking a lag period of one month for all the variables. ARIMA model was found to be the best fit statistical model for predicting malaria cases in Delhi. They identified that only two climatic factors viz., mean rainfall and relative humidity both lagged at one month were significant predictors of malaria infections in the study area. They suggested that this information could be useful for administrators in effectively implementing preventive and control measures for malaria.

3.5.5 Comparison of time series models with Exponential Smoothing Method and the Box Jenkins Method

Malinga (2015) explored the feasibility to fit appropriate time series models, compare and assess the accuracy of two different methods, i.e., Exponential Smoothing Method and the Box Jenkins Method to forecast malaria in children under 15 years of age from different epidemiological zones in Kenya. He tested the data from three health facilities located in three different regions of Kenya and evaluated the performances of the models through data splitting using error measures such as mean error, mean percentage error and the mean absolute squared error. The statistical forecast accuracy of the models showed that both methods were effective in predicting malaria case admissions in Kenya. He observed that Box Jenkins method was more accurate in long term projections while the exponential smoothing method performed better in short and medium term forecasts and in data which experienced recent abrupt level shifts. Through the study, he showed the limitations of forecasting malaria case admissions using only historical patterns and emphasised the need to develop improved models by incorporating external predictors such as climate-related variables. He concluded that although different methods can be applied to malaria forecasting, they should be tailored to the specific malaria transmission setting to avoid misleading results due to the underlying disease transmission dynamics.

3.5.6 SARIMA model for dengue in Rajasthan

Bhatnagar et al. (2012) developed a prediction model for dengue fever/dengue haemorrhagic fever (DF/DHF) using time series data over the past decade in Rajasthan which can forecast monthly DF/DHF incidence for 2011. They concluded that among all candidate models, SARIMA (0, 0, 1) (0, 1, 1)_{12} was the most suitable predictive model in their study, which showed the highest stationary R-squared and the lowest normalized BIC (Bayesian information criterion) and MAPE (mean absolute percentage error) values. However, they
opined that the predictions might not be credible for forecasting the number of dengue cases in epidemic years, as it could be result of lack of immunity in a population exposed for the first time to a given dengue viral serotype. They advised that further research is required to evaluate the effectiveness of integrating the forecasting model into the existing disease control program in terms of its impact in reducing the disease occurrence.

### 3.5.7 Dengue – Negative Binomial model

Ramachandran *et al.* (2016) conducted retrospective study to investigate whether climatic factors could be used to predict the occurrence of dengue in East Delhi and developed an empirical model for estimating dengue incidence using temperature, rainfall and relative humidity in East Delhi. For this, the retrospective data on the number of monthly dengue cases reported at Guru Teg Bahadur Hospital in East Delhi for a period of 19 years i.e., from January 1997 to December 2015 was obtained. Monthly climatic data on rainfall, temperature and humidity was collected from the Delhi Weather Station, New Delhi. One-way analysis of variance (ANOVA) was performed to determine whether each of the climate variables differed significantly between seasons. As the outcome variable was in counts and was over-dispersed, a negative binomial model with a log-link function in the generalized linear model was used to model for estimating dengue cases based on the independent variables of rainfall, temperature and humidity. Since a significant correlation was present in the variability of dengue cases across seasons, they included season as a covariate in the model. They concluded that temperature, rainfall and relative humidity significantly affected dengue occurrence in East Delhi and that the weather-based dengue Negative binomial model can forecast potential outbreaks two months in advance, providing an early warning system for intensifying dengue control measures.

### 3.5.8 Boosted regression tree (BRT) model for Dengue fever

Liu (2014) investigated the validity of different selection strategies when using *presence only* and *pseudo absence* data while modelling the spatial distribution of dengue fever. He used Boosted regression tree (BRT) model to investigate climatic conditions and human population as possible predictors of dengue fever transmission. For this, he considered disease *presence only* (PO) and disease *pseudo absence* (PA) data. The places where dengue fever had been reported globally were referred to as *presence only* (PO) data and the places where dengue had not been reported were referred to as *pseudo absence* (PA) data. He used
two strategies to randomly select pseudo absence (PA) data. One strategy used a selection based on the geographical distance to presence only (PO) and the other strategy used selection of data based on the report of evidence based consensus regions of dengue absence. He concluded that different pseudo absence selection methods affect the distribution of dengue. He further claimed that the risk maps indicated the risk areas of dengue are more under selection according to evidence-based consensus compared to selection at random.

3.5.9 Seasonal autoregressive model for dengue in Mexico

Johansson et al. (2016) assessed and compared dengue forecasts obtained from different types of models and evaluated the performance of seasonal autoregressive models with and without climate variables for forecasting dengue incidence in Mexico. They observed that climate data did not improve the predictive power of seasonal autoregressive models significantly. They concluded that although seasonal autoregressive models captured a substantial amount of dengue variability, better models are needed to improve dengue forecasting.

3.5.10 Artificial neural network

Penna (2004) tried to evaluate recurrent neural networks as a predictive technique for time-series in the health field through a study on use of an artificial neural network for detecting excess deaths due to cholera in Ceará, Brazil. She carried out the study during a cholera epidemic which took place in 1993 and 1994 in the state of Ceará, Northeastern Brazil and was based on excess deaths having ‘poorly defined intestinal infections’ as the cause (ICD-9). She obtained the monthly number of deaths due to this cause between 1979 and 1995 in the state of Ceará from the Ministry of Health’s Mortality Information System (SIM). She concluded that the artificial neural network showed good predictive ability, especially in the initial period, and was able to detect alterations concomitant and subsequent to the cholera epidemic.

3.5.11 SARIMA for Hand, Foot and Mouth disease

Song et al. (2015) performed Time Series analysis of Hand, Foot and Mouth disease (HFMD) integrating weather variables as part of their research. For this, samples from 1556 inpatients and 11004 outpatients diagnosed with HFMD were collected from January 2009 to October
2013 in Guangzhou city, China. They applied Seasonal Autoregressive Integrated Moving Average (SARIMA) model to establish high predictive model for inpatients and outpatients as well as three viral serotypes (EV71, Pan-EV and CA16). They included meteorological variables obtained from the National Meteorological Information Center, such as Average temperature (°C), Maximum temperature (°C), Minimum temperature (°C), Humidity (%), Visibility (Km), Mean wind speed (Km/h), Maximum sustained wind speed (Km/h) and Precipitation amount (mm) in the analysis. The climate variables which contributed significantly were identified by correlation analysis and were executed to improve time series modelling as external repressors. They proposed that climate patterns and HFMD incidences have been shown to be strongly correlated and that SARIMA model developed for this study can be a helpful tool in developing an early warning system for HFMD.

3.5.12 Decision tree algorithm and Naive bayes classifier: A comparison in the case of Swine flu.

Shinde et al. (2015) compared Decision tree algorithm and Naive bayes classifier for Swine flu prediction. They used clustering algorithm K mean to make a cluster of Swine flu suspects in an area. They applied Decision tree algorithm and Naive Bayes classifier on the same inputs to find out the actual number of suspects and predict the possible surveillance of a Swine flu in a nearby area from suspected area. They compared the performances of these techniques based on accuracy. They observed that Naive Bayes classifier performed better than decision tree algorithm while finding the accurate count of suspects.

3.5.13 SARIMA model for incidence of tuberculosis (TB) in Iran

Moosazadeh et al. (2014) estimated the incidence of tuberculosis (TB) in Iran using Box-Jenkins Models. To forecast TB in Iran, they considered monthly data of tuberculosis cases recorded in the surveillance system of Iran tuberculosis control program from 2005 till 2011. They predicted the overall TB incidence in 2014 to be 1.38 times the year 2005. They observed that the TB peak in Iran is in spring and May. After discussing the work done on forecasting methods by many researchers, they came to the conclusion that SARIMA \((0, 1, 1)(0,1,1)_12\) gave the best fit for predicting TB prevalence. They concluded that the prediction results showed an increasing trend of TB cases in Iran.
3.5.14 SARIMA model and a generalized regression neural network model for tuberculosis

Cao et al. (2013) aimed to develop an appropriate model for predicting tuberculosis (TB) epidemics in China. To develop the model, data on monthly TB incidence cases from January 2005 to December 2011 were obtained from the Ministry of Health, China. They used Seasonal Autoregressive Integrated Moving Average (SARIMA) model and a hybrid model which combined the SARIMA model and a generalized regression neural network model to fit the data. They compared the goodness of fit of these two models using simulation performance parameters of mean square error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). They concluded that the hybrid model showed better TB incidence forecasting than SARIMA model.

3.5.15 Autoregressive integrated moving average model (ARIMA) - Back propagation neural network (BPNN)

Ren et al. (2013) developed a mathematical model to forecast the incidence of hepatitis E in Shanghai, China. For developing the model, they considered morbidity data of hepatitis E in Shanghai from 2000 to 2012, obtained from the China Information System for Disease Control and Prevention. The autoregressive integrated moving average model (ARIMA) - Back propagation neural network (BPNN) combined model was developed to fit the linear and nonlinear patterns of time series data. Residual analysis, Root Mean Square Error (RMSE), normalized Bayesian Information Criterion (BIC), and stationary R square methods were considered to compare the goodness-of-fit among ARIMA models. Goodness-of-fit and parameter estimates were used to determine the best-fitting model and it was found that ARIMA (0, 1, 1)(0 ,1, 1)_{12} gave the best fit to the incidence of hepatitis E infection. They concluded that Time series analysis suggested a seasonal pattern of hepatitis E morbidity in Shanghai, China.

3.5.16 Comparison of negative binomial models with seasonal model

Soyiri et al. (2013) developed two related negative binomial models and compared them with a naive seasonal model to forecast asthma-related hospital admissions in London. For this, data on hospital admissions of asthma patients were obtained from the nationally recorded Hospital Episode Statistics maintained by the National Health Service, England. The data
consisted of an unidentified record of all asthma-related, emergency hospital admissions within London from 1 January 2005 to 31 December 2006. They obtained weather related data such as ambient air temperature, vapour pressure (HPa) and humidity (%) from the UK Meteorological Office database and were based on averaged daily results from the weather monitoring sites across London. Air quality data were based on 24h averages and was obtained from air quality monitoring sites across London. The data was analysed by taking the daily count of asthma admissions as the dependent variable and the averaged 24h daily weather measures (including temperature and humidity) as well as the averaged 24h daily air quality measures of ozone, nitrogen dioxide and nitrogen oxide for London as potential predictors. The data were analysed using negative binomial models to iteratively model the effects of these exposures. They observed that seasonality was the main predictor of asthma daily admissions with little influence of additional meteorological data.

3.5.17 Zero inflated model

Ahmad et al. (2015) examined the factors that are associated directly or indirectly in pneumonia patients among the children between the age group 0 – 12 years through a study on applications of zero inflated models for health sciences data. They considered a five year study period, from 2005 to 2009, which consisted of 1252 patients from a hospital in Malaysia. They considered diagnosis of Pneumonia as the response variable and diabetes, influenza, septicemia, diarrhoea, asthma, anemia, tuberculosis, health services, acute tonsilitis, streptoccocus pneumonia, and heart disease as independent variables to model Pneumonia. They fitted Poisson, Negative Binomial (NB), Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) regression models to pneumonia data. Based on evaluation statistics, they concluded that Zero-Inflated Negative Binomial model gave the best fit to the data.

3.5.18 Multivariate logistic regression for modelling diabetes

Niyikora (2015) used multiple logistic regression to model diabetes. For this, he considered a three year period data i.e., from 2011 to 2013 on diabetes from Gitwe Hospital as dependent variable. Age, gender, occupation status, smoking, alcohol consumption, cholesterol level, hypertension and family history of diabetes were considered as risk factors of diabetes. He performed multiple logistic regression on all the predictors and a test of significance on coefficients. The Wald test revealed that age, occupation status, alcohol consumption,
cholesterol level and Hypertension as statistically significant and gender, smoking and family history of diabetes as not statistically significant predictors. He concluded that the risk factors *viz.*, older age, smoking, alcohol consumption, high cholesterol level and hypertension increase the risk of having diabetes, whereas being employed decreases the risk of having diabetes.

### 3.5.19 Modelling Diabetes Disease Diagnosis Using Multivariate Adaptive Regression Splines

Senthilkumar *et al.* (2013) introduced the significance of Multivariate Adaptive Regression Splines (MARS) in the disease diagnosis through the collected data for diabetes to develop an intelligent decision support system. They observed that MARS model obtained better accuracy with minimum number of predictors and outperformed by handling nonlinearities, missing data and interactions among predictors compared to other methods. They concluded that the proposed model is easily understandable, provides a better and faster model for diagnosing diabetes patients.

### 3.5.20 Artificial neural network (ANN) for thyroid disease

Gharehchopogh *et al.* (2013) used artificial neural network to diagnose thyroid disease. They considered a Multi-layer Perceptron (MLP) artificial neural network (ANN) with back propagation learning algorithm to classify Thyroid disease. Their simulation results indicated that the optimization in Multi-layer Perceptron artificial neural networks can reach the classification accuracy for Thyroid disease up to 98.6%.

### 3.5.21 Support Vector Machine (SVM) and Artificial Neural Network (ANN): A comparison in the case of heart disease.

Vadicherla *et al.* (2013) analysed the application of Support Vector Machine (SVM) and Artificial Neural Network (ANN) for classification and prediction of heart disease. They compared the two methods based on accuracy and training time. They classified the data on heart disease into two classes using Sequential Minimal Optimization (SMO) algorithm in Support Vector machine and Artificial Neural Network (ANN). They compared the performance of both the classification techniques with respect to time needed for classification and accuracy of the system. Based on the analysis, they concluded that
Sequential minimal optimization in Support vector machine is more effective than Resilient back propagation in Artificial neural network.

### 3.5.22 Heart Disease Prediction Model Using Classification Algorithms

Yadav *et al.* (2013) analysed the performance of different classification techniques for the prediction of heart disease from the heart disease data set. The techniques that were compared were function based Logistic, Multilayer perception and Sequential Minimal Optimization algorithms. After analysing the experimental results, they concluded that the logistic classification algorithm technique emerged as the best classifier for heart disease prediction as it is more accurate and has least error rate.

### 3.5.23 Brain Tumor detection and classification using Adaptive boosting

Bargaje *et al.* (2017) proposed a Decision tree with adaptive boosting technique to classify the brain Magnetic Resonance Imaging (MRI) image as normal or abnormal. They claimed that Adaptive boosting technique provided 100% accuracy in classifying the MRI image.

### 3.5.24 Lymph Diseases Prediction Using Random Forest and Particle Swarm Optimization

Almayyan (2016) developed a model to enhance lymphatic diseases diagnosis by the use of random forest technique. He carried the study in two phases. First phase included feature selection and the second phase included classification. In the first phase, he selected a number of discriminative features using Particle Swarm Optimization (PSO). In the second phase, he applied the random forest ensemble classification technique to diagnose lymphatic diseases. He used original and resampled distributions of the dataset to train random forest classifier. He concluded that proposed random forest (RF) + Particle Swarm Optimization (PSO) model improved the accuracy performance and achieved promising results.

### 3.6 Modelling on plant diseases

#### 3.6.1 Models developed by Indian Agricultural Statistics Research Institute (IASRI)

S.C. Mehta *et al.* discussed the contribution of Indian Agricultural Statistics Research Institute (IASRI) in forewarning crop pests and diseases. They reckon that IASRI has played
a significant role in developing methodologies for forewarning of different aspects relating to crop diseases, such as Alternaria Blight, White Rust, Powdery Mildew and Aphid (Mustard), Aphid (potato), American boll worm, Pink boll worm, Spotted boll worm & Whitefly (Cotton), Spodoptera litura, Late leaf blast & Rust (Groundnut), Pyrilla, Early shoot borer & Top borer (Sugarcane), Pod fly, Pod borer, Sterility Mosaic & Phytophthora Blight (Pigeon pea), Fruit fly, Hopper & Powdery Mildew (Mango) and Gall midge (Rice) for selected centers. They discussed various models developed at the institute, mostly weather based, using quantitative as well as qualitative data. Models that were discussed are Regression models, Complex polynomial, Model by deviation method and Artificial Neural Network (ANN) technique. They contend that the performance of the models was found to be good and that the models provided timely forewarning of various aspects of pests and diseases (www.iasri.res.in/sovenior/article_07.pdf).

3.6.2 Support vector machines (SVM)

Kaundal et al. (2006) introduced a new prediction approach based on support vector machines (SVM) for developing weather-based prediction models for plant diseases. They developed a SVM-based web server for rice blast infection prediction. It is first of its kind in the world, which can help the plant science community and farmers in their decision making process. The server is freely available to people who would like to use SVM.

To the best of authors' knowledge, there is no report of using support vector machines in understanding the relationship between disease severity and its associated environmental conditions. So, the present case study was aimed to determine the usefulness of SVM models over the existing artificial neural network (ANN) and conventional multiple regression models to predict rice blast severity based on prevailing weather conditions both within and between the locations/years, and to calculate the overall risk of rice blast infection at these field sites using a set of weights from the trained SVM models. For this, they used 5-year rice blast/weather data collected from five different locations spread over the district Kangra of Himachal Pradesh (India) as part of a National Agricultural Technology Project (NATP) which was implemented at CSK Himachal Pradesh Agricultural University, Palampur, Himachal Pradesh, India. Their main aim in the study was to evaluate the status of currently available approaches for predicting plant diseases in comparison to the latest prediction techniques that are still unexploited like SVM. They observed that SVM performed better than Back propagation neural network (BPNN), generalized regression neural network
(GRNN) and Regression (REG) on the dataset used in this study. However, they do not claim that in general, SVM is better than neural networks or the other methods, but the fact is that their analysis on the present dataset revealed better prediction accuracy with SVM over all the years and locations as compared to the REG, BPNN and GRNN-based validation approaches, which is an important and significant contribution of this study. They contend that the advantages of using SVM over the other methods are that SVMs are robust when one has a sparsely filled high dimensional dataset and that the flexibility allowed by this modelling approach makes it possible to fit nonlinear relationships and complex interactions between variables without requiring complex transformations of variables and trial-and-error searches for interactions as required in neural networks. They opine that for a system like leaf blast of rice for which the associations between the environment and disease development are not completely understood with conventional statistical techniques, SVM serves as an excellent tool for developing prediction models. Finally, they wrap up by suggesting that the present web server could play an important tool in integrated rice blast management system for direct application to plant pathologists, academicians and farmers which will ultimately have direct economic impact in terms of reduction in expenses to develop new fungicides, development of target-site specific and environment-friendly fungicides and increase in savings by reducing fungicide application.

3.7 Species distribution model validation

Allouche et al. (2006) have explained the observed dependence of Kappa statistic on prevalence, and proposed an alternative measure of accuracy, the True Skill Statistic (TSS). They compared the responses of Kappa statistic and True Skill Statistic to prevalence using empirical data. They showed the limitations of Kappa and opined that True Skill Statistic can compensate for the limitations of Kappa, keeping all the advantages of Kappa. They recommend True Skill Statistic as a simple and instinctive measure for assessing the performance accuracy of species distribution models.