GIS-Based Multi-Hazard Susceptibility Assessment using Machine Learning

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

of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Shruti Sachdeva

(Roll No. 801532049)

Under the supervision of:

Ms. Tarunpreet Bhatia

Lecturer

&

Dr. A.K. Verma

Associate Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

June 2017

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, ***"GIS-Based Multi-Hazard Susceptibility Assessment using Machine Learning"***, in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Ms. Tarunpreet Bhatia* and *Dr. A.K. Verma* and refers other researchers' works which are duly listed in the reference section. The matter presented in the thesis has not been submitted for the award of any other degree of this or any other University.

Shruti Sachdeva

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

Ms. Tarunpreet Bhatia
Lecturer

Dr. A. K. Verma
Associate Professor

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my guides **Ms. Tarunpreet Bhatia** and **Dr. A.K. Verma**, Department of Computer Science and Engineering for their continuous support throughout my study and related research, for their patience, motivation, and immense knowledge. Their guidance helped in all the time of research and writing of this thesis. I could not have imagined having better advisors and mentors for my M.E. study.

I am also thankful to **Dr. Maninder Singh**, Head of Department, Computer Science and Engineering Department and **Dr. Ashutosh Mishra**, P.G. Coordinator, for their constant inspiration that kept me motivated for this thesis work. I also want to express my gratitude to **Dr. S.S. Bhatia,** Senior Professor and Dean of Academic Affairs, Thapar University, for making provisions of resources and infrastructure without which the completion of this work would not have been possible.

Last but not the least; I would like to thank my parents for supporting me spiritually throughout the writing of this thesis and my life in general.

Shruti Sachdeva

ABSTRACT

The present century has been the most unfortunate in the context of the casualties and losses caused on account of the disasters induced by natural hazards globally. Words fall short in describing the scale of devastation and havoc brought on Indian soils by floods, landslides, forest fires, avalanches, etc. while the mankind remained a mute spectator in such crucial times. All the while the advancements in science and technology has coined terms such as machine learning and artificial intelligence in its miraculous projectile of evolution. These sciences have empowered us to be better prepared for handling aforementioned adversities caused by natural disasters through indigenous techniques of hazard susceptibility prediction and assessments. These efforts have been enhanced by an incandescent flourishing of technologies like Remote Sensing and Geographic Information Systems. This thesis is an attempt in this direction of predicting a region's susceptibility towards a specific hazard of which the given region has been a victim of. This task involves the use of satellite images to extract characteristic details of the stretch of land under study. The aforementioned specific characteristics are the causative factors which play a role in the unfolding of the specific hazard. The spatial coordinates of the locations which in the past have witnessed the disasters on various scales are mapped with their corresponding values of their causative factors/traits. This problem of mapping is reduced to that of mere pattern detection by identifying the traits and their values that have played a pivotal role in the past episodes of disaster and predicting the susceptibility of any such events in future. This is where the expertise attained in the field of machine learning steps in to rescue. The susceptibility for the region is interpolated for the entire expanse of the region selected, to delineate the regions in terms of their susceptibility on the scale of very highly susceptible to least susceptibility. The accuracy in delineations has been enhanced by virtue of state-of-the-art optimizations and ensembling of a variety of machine learning models. The final outcome is a hazard susceptibility map describing the corresponding region's achieved delineations that could aid in framing a contingency plan, along with, strengthening the efforts of disaster mitigation and management.

Keywords: Remote Sensing, Geographic Information System, Hazard Susceptibility Mapping, Machine Learning, Particle Swarm Optimization, Evolutionary Optimization

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
CART	Classification and Regression Trees
DT	Decision Trees
EO	Evolutionary Optimization
FFSM	Forest Fire Susceptibility Mapping
FFSZ	Forest Fire Susceptibility Zonation
FSM	Flood Susceptibility Mapping
FSZ	Flood Susceptibility Zonation
GBDT	Gradient Boosted Decision Trees
GIS	Geographic Information System
HSM	Hazard Susceptibility Mapping
HSZ	Hazard Susceptibility Zonation
LR	Logistic Regression
LRM	Likelihood Ratio Model
LSM	Landslide Susceptibility Mapping
LSZ	Landslide Susceptibility Zonation
LULC	Land Use Land Cover
NB	Naïve Bayes
NDBR	Nanda Devi Biosphere Reserve
NDVI	Normalized Difference Vegetation Index
PSO	Particle Swarm Optimization
RF	Random Forest
ROC	Receiver Operating Characteristics
RS	Remote Sensing
SPI	Stream Power Index
SVM	Support Vector Machine
TPI	Topographic Position Index
TWI	Topographic Wetness Index
VFI	Voting Feature Intervals
WLC	Weighted Linear Combination
WoE	Weight-of-Evidence

CHAPTER 1
INTRODUCTION

History has been witness to the vagaries of nature time and over. Natural disasters are major inimical events caused by nature resulting in the irreversible loss of human lives and property on wide scales [1]. Due to their enormous scales, they affect lives and properties all around the globe. And as in the majority of cases, the reasons are not typically under our control. The socio-economic status of the households has a direct bearing on the magnitude and the nature of these effects [2]. There is a unanimous agreement that a prior knowledge of a region's level of susceptibility to a particular natural calamity can significantly improve our level of preparedness. This would be highly favorable for drafting out mitigation strategies. Experts from various disciplines and fields namely meteorology, geology, environmental sciences, computer sciences, and various others have time and again put in herculean efforts to predict the time, place and severity of the disasters. Weather forecasting models, data mining models, machine learning algorithms, artificial intelligence are only a few of the mechanisms being employed in accomplishing the above-mentioned tasks worldwide. Finding out the requirements and need of the victims during a natural disaster through a quick and accurate flow of information is another point of research. For the accomplishment of such tasks of research and study, the social media and the internet have been unanimously been agreed upon as the most available sources of information. Apart from these, satellites, wireless and remote sensors, UAVs (unmanned aerial vehicles), national and international meteorological, geographical departments, NGOs, international, government and private bodies also contribute data before, during and after a disaster.

Tasks involved in dealing with disasters are primarily classified into three broad categories of prediction, detection and management [3] as shown in Figure 1.1. Prediction involves prognosticating place, time and scale of the natural disaster that can occur. Apart from this, it also involves prediction of disaster prone area and various other attributes associated with the disaster. Detection involves quickly ascertaining the occurrence of any disaster. Studies so far have indicated that the response of social networking is more prompt in reporting the occurrence of disasters as compared to

conventional means [4]. Management involves identification of communication needs, the requirement of relief materials for the victims and distribution of the same. In a majority of the cases, the categorization is rather not strict or exclusive, but remains highly overlapping [3]. The prediction domain allows a greater chance at survival and recuperation due to an impetus in being prepared, alert and vigilant corresponding to the degree of disaster, as and when it hits.

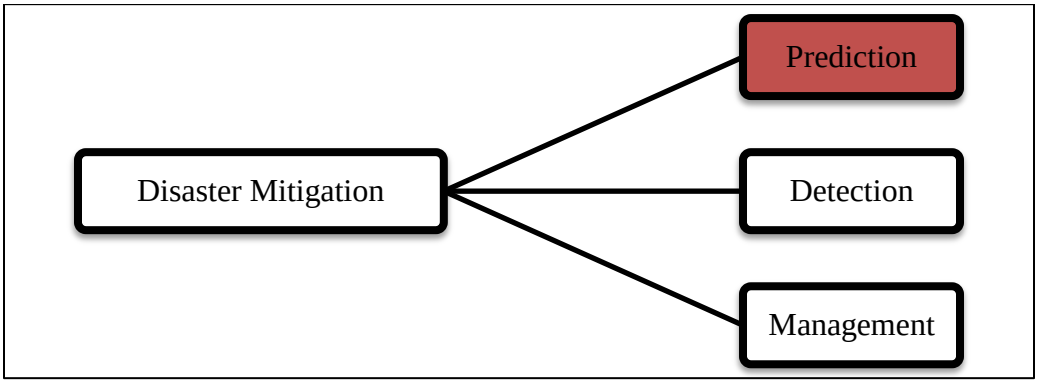

Figure 1.1: Types of Disaster Mitigation Techniques

Machine learning is capable of providing assistance in the domain of spatial prediction of aforementioned potential hazards. These techniques tend to simulate the human intelligence and decision making by analyzing the past data and the factors/circumstances present therein causing those hazards. It then tends to observe the relationship between these factors and the hazards and builds a model accordingly, which is a collection of rules which can then be reiterated over a new data set allowing us to make a new set of hazard predictions. Hence it apes the human thought process at a much faster pace and greater accuracy.

1.1 Motivation for Thesis

Among all the countries in the world, Asian countries are considered to be few of the most frequent witness of a variety of natural hazards as the region has seen the maximum number of disaster-induced casualties and fatalities. The main reason attributed is the dynamic ecosystem in prevalence. The extensive financial overhaul and rural empowerment leading to population remodeling in terms of both their numbers and diversification is a conspicuous manifestation of such activities. Climate change and rapid land-use transformations are other sinister by-products of these exertions. It is not

surprising that these alterations have sparked major flip-flops in the frequency and spatial dissemination of natural hazards. Nowhere is it truer than in the case of South-East Asia, which houses some of the world's most rapidly growing economies.

In India, among the fatal disasters induced by natural hazards, the widest scale of apathy and the havoc actuated has been attributed to Landslides, Floods, Forest Fires, and Avalanches etc. Among these Landslides, Floods and Forest Fires have been given a detailed consideration in this study and a special focus in the case studies on specific regions as per their corresponding proneness to these unfortunate events. Landslides are the most potent catastrophes prevalent, particularly in the hilly regions of India [5]. An approximate 15% of the India's landmass encompassing besides Western Ghats, the Himalayas, the Northeastern hill ranges, the Eastern Ghats, the Nilgiris, and the Vindhyans are vulnerable to major landslides [6]. The North-Eastern part of India has a history of major landslide disasters. The state of Nagaland has experienced major landslides like those of Dimapur (August 2001), Kohima (August 2003), Mokochung (May 2005), Wokha (August 2006), Zunheboto (September 2006), and Kiruphema (October 2007) [7]. In Assam, the city of Guwahati has experienced a number of catastrophic landslides in its hilly areas [8]. In the monsoons of the year 2016, more than 100 people were affected by landslides at Phesama along NH-2, Nagaland. Ten people lost their lives to landslides in Karimganj and Hailakandi districts in Barak Valley in Assam on 18th May 2016. So a region encompassing the states of Assam and Nagaland and surrounding areas has been selected for landslide susceptibility mapping (LSM) in Case Study I.

On the other hand, an estimate between 75 to 820 million hectares of land burns each year globally [9]. In India, the facts are equally staggering. As per the Forest Survey of India, around 50 percent of the country's forests are prone to wildfires [10]. It is reckoned that the proportion of forest areas prone to fires annually ranges from 33% in some states to over 90% in others. Also, 90% of the forest fires are deduced from human factors [11]. As per a study, the 5 year period of 1985-1990 witnessed 17,852 fires affecting 5.7 million hectares, culminating to a yearly average of 1.14 million hectares [12]. As per the Parliamentary Standing Committee on Science and Technology, the year 2016 witnessed a 55% increase in the number of forest fires [13]. Of these, 1,600 were from the state of

Uttarakhand which is the home to the Nanda Devi Peak, the second highest mountain in India. The fires squandered 3,185 hectares of forest area [14]. The year 2017 also witnessed an almost recap of the previous years with instances being reported from Pithoragarh district's Munsiyari, places in Almora, the Pindar valley in Chamoli district and areas in Bageshwar district and the Nanda Devi Biosphere Reserve (NDBR). So, this study emphasized on forest fire susceptibility mapping (FFSM) on the region of NDBR and surrounding areas (districts of Chamoli, Pithoragarh and Bageshwar) as discussed in Case Study II.

All the while, the frequency of floods has increased manifold over the years and so have their intensities. As per the findings of Nations Office for Disaster Risk Reduction (UNISDR) [15], a total of 3,455 flood incidents have been reported worldwide in the period of 1980-2011. Floods also account for affecting the largest number of people worldwide, a total of about 2,437 million in the 20 year period from 1992 to 2012. As per the World Disaster Report [16], floods accounted for 44% of deaths caused by natural hazards which are more than the deaths caused by any other disaster. Almost 100 million people were affected by floods in the year 2013 alone with Asia being the worst affected region of the world as 87% of those affected were Asians. One such unfortunate incident in this region that shook the world was the 2013 floods in Uttarakhand, India. The floods brought major devastation with 4,120 people dead, 1,800 villages affected, 2,500 families homeless, 150 bridges destroyed and 17,000 km^2 roads damaged [17]. The estimate of the losses to bridges and roads went up to 285 million US $, losses associated with dam projects went up to 30 million US $ the state's tourism suffered losses amounting to 195 million US $ [18]. July 2016, brought back the haunting memories of the 2013 floods when similar conditions were almost on the brink of recurrence, causing various landslides and taking a toll of around 30 lives. Taking note of the recurring flooding episodes here, the district Chamoli has been selected for carrying out flood susceptibility mapping (FSM) as specified in the Case Study III.

1.2 Thesis Contribution

This research work uses state-of-the-art technologies of remote sensing (RS) and geographic information systems (GIS) for information extraction from satellite images

pertaining to specific regions in India that are prone to hazards like floods, landslides and forest fires. The information extracted is then processed via unique machine learning techniques for generating Hazard Susceptibility Zonation (HSZ) maps. Overall, a unique amalgamation namely, RS and GIS hand-in-glove with machine learning unravel the mystery behind a region's traits that condition it for the final encounter with an unfortunate disaster. This has been accomplished by an extensive study of the existing hazard susceptibility assessment techniques which was then followed by the design and development of models for LSM, FFSM and FSM. The developed models were verified and validated by comparison of the proposed models with existing machine learning models on the basis of various comparative metrics. The validated models were hence used for respective HSZ map generation. It is aimed that such maps would empower the concerned agencies with information that could aid in disaster mitigation and management efforts.

1.3 Thesis Layout

The thesis has been organized as follows: Chapter 2 gives an insight into the theoretical background associated with the study; Chapter 3 encompasses the related work that has been recognized in the research community while Chapter 4 summarizes the research gaps recognized from Chapter 3 and justifies the selected objectives and the proposed methodology for this research work; Chapter 5 delves into the case studies implementing the proposed strategy. Chapter 6 encloses the final conclusion with further scope for improvisations.

CHAPTER 2

THEORETICAL BACKGROUND

In order to arrive at the problem statement systematically, it is imperative to define the basic terminologies as foundation stones. So the phenomenon of natural hazards, the state-of-the-art machine learning techniques as well as the performance metrics employed for their comparisons and the sophistication associated with the relatively new fields of RS and GIS is introduced in the following sections.

2.1 Natural Hazards

The sustenance of mankind on the earth's countenance is unanimously agreed upon as the byproduct of nature's ecological indulgence. Natural disasters are major inimical events caused by nature resulting in the irreversible loss of human lives and property on humungous scales [1]. The major disasters to have scarred the earth's countenance include floods, earthquakes, droughts, forest fires, tsunamis and landslides. An insight into a few of the major hazards namely landslides, forest fires and floods are provided here.

2.1.1 Landslides

Among hazards, one of the most intriguing and complex natural hazards with relatively higher mortality and morbidity rates is the landslides. The term "landslide" refers to the shifting motion associated with a pile of boulders or debris naturally triggered by factors like earthquakes, cloudbursts, snow melting, etc. leading to the displacement of land mass [19]. They are characterized by a spectrum of surface motions like cliff falls, slope failures, debris topples etc. Other than their natural causes, the triggers catalyzing landslides include human activities of thoughtless excessive mining, excavation, and haphazard infrastructure development. No continent, or ocean has been left untouched by landslides making their impact far reaching and long lasting. Ergo millions and billions have been spent in aid money for recuperating from the effects of such events. Consequently, over the course of time there have been numerous investigations in the field of landslide hazard assessment and in those attempts there has been a constant need to construct maps displaying a region's relative landslide proneness depending upon the

characteristics of the locations that have in the past encountered landslides by many a nation's government and various other national and international research institutes. The probability that the landslide will strike a given area under the presence of a certain set of conditions is known as landslide susceptibility. Although the exact time as to when the landslide will occur cannot be furbished from such information, however, the magnitude/scale/frequency which the landslide might occur is widely accepted as milestone result brought under consideration for land use zoning and environmental planning.

One step towards forming a landslide contingency plan involves LSM which could be defined as an ensemble of processes carried out in succession to achieve what is known as a Landslide Susceptibility Zonation (LSZ) Map [5]. An LSZ map is defined as an HSZ map aimed at predicting those specific locations where the slope failures are most or are highly likely to occur. The modus operandi for LSM involves in-depth study and data collection for the landslide prone area, data pre-processing, building a prediction model on the processed training data set, validating the model built on the testing data set, and ultimately implementing the validated model to predict a new location's susceptibility on the basis of the data. The effectiveness of LSZ map would depend on the method applied and the quality of the data on which the method has been applied. Historical, Geological, topographical and environmental factors must be taken into account while constructing such maps. The historic aspect of the study is accounted for by the inclusion of a landslide inventory that comprises of the coordinates of the locations that have been struck by landslides in the past. While the other factors (geological, topographical and environmental) are collectively termed as landslide conditioning factors.

2.1.1.1 Landslide Inventory

Landslides are conventionally associated with mountainous ranges, river banks and coastlines. Also, landslides tend to repeat themselves at locations where they have struck before. Therefore an area's past landslide occurrence increases the area's susceptibility to a fresh landslide. This historic factor is encompassed by means of a landslide inventory. A landslide inventory indicates the spatial locations which in the past experienced landslides [20]. With past landslide occurrences of a region holding important clues for the region's future landslide susceptibility, thus the landslide inventory is an imperative

part of the landslide studies. Conventionally, these inventories have been built by means of physical field visits ascertaining the exact locations and the damage suffered at these locations. Such field surveys are difficult to undertake due to inaccessibility of the regions and time constraints. However, with the recent technological advancements in the fields of RS and GIS, has made it possible to obtain such results quickly and with equal precision.

2.1.1.2 Landslide Conditioning Factors

The requisite data for LSM allude to the Landslide Conditioning or Landslide Causative factors. These are the parameters that congruently deduce a location's susceptibility to landslides. Over the years, the study in this field has witnessed different factors being adopted with varying degree of relevance derived from a location's inherent attributes. Few of the commonly employed causative factors are as mentioned:

- **Elevation**: It is an important causative factor in landslide assessment which has been included in various landslide susceptibility assessment studies [21, 22]. It refers to the height of the surface terrain above sea level [22]. It has been established that the rocks at higher elevation tend to have more strength in comparison to those at lower altitudes hence justifying their inclusion as a landslide conditioning factor [21].
- **Slope:** It refers to the angle that the landform makes against the horizontal. It has been incorporated in studies like [20, 23]. It has been found to play a crucial role in landslides and henceforth has been in embodied in various studies including the Case Study I.
- **Aspect:** Aspect indicates the direction of the slope [22] and has been referred in works like [20, 22]. Aspect influences major traits of lands like the direction of received rainfall and the amount of sunshine gathered.
- **Curvature:** Curvature has been frequently used in landslide studies attributing to its ability to control the water flow on the surface [24]. It is defined as the inverse of the value of the radius of the curvature of the line [25]. Due to the inverse proportionality tighter curves tend to have higher values of curvature as compared to wider curves.
- **Plan Curvature:** Plan curvature is the value of the curvature associated with a contour formed by the intersection of the landform surface and the horizontal plane, hence representing the bending of the surface in the perpendicular direction [22].

- **Profile Curvature:** Profile Curvature indicates the rate of change of slope or the curvature in the vertical plane [24]. All three factors of curvature, plan and profile curvature interweave, controlling the acceleration, deceleration and viscosity of the flow on the surface [22].
- **Surface Roughness:** The surface roughness is an important aspect of LSM which takes into consideration the satiability aspect of the slope [26].
- **Slope Length (LS):** Slope length is the length of the slope subjected to a sustained overflow considered from the initiating point of the debris flow/rock fall to the point where sedimentation begins or the flow immerses into a sink [22]. A long slope would result in gaining a higher momentum by the flow leading to an increment in its degradation powers. It can be calculated by Equation (2.1) [22].

$$LS = 1.4 \left(\frac{A_S}{22.13}\right)^{0.6} \left(\frac{\sin\beta}{0.0896}\right)^{1.3} \quad (2.1)$$

- **Topographic Wetness Index (TWI):** It interpolates the effect of topography on the slides/flows [25]. It is specified in Equation (2.2) [26].

$$TWI = \ln\left(\frac{A_S}{\tan\beta}\right) \quad (2.2)$$

- **Stream Power Index (SPI):** It indicates the potential erosive power of the overland flow. It is calculated by Equation (2.3) as proposed in [26].

$$SPI = A_S \times \tan\beta \quad (2.3)$$

In Equations (2.1), (2.2), and (2.3) A_S refers to the catchment area and β is the slope in degrees.

- **Normalized Difference Vegetation Index (NDVI):** It is widely accepted that the denser the vegetation lower are the chances of landslides [26]. So, NDVI takes the vegetation density into account for landslide susceptibility. In order to obtain NDVI 4^{th} and 5^{th} band of the LANDSAT 8 OLI images are generally employed, which contain the information regarding Red and Near Infrared Reflection (NIR) respectively. The NDVI is then calculated using Equation (2.4) as proposed by [26].

$$NDVI = \frac{(NIR - RED)}{(NIR + RED)} \quad (2.4)$$

- **Land Use Land Cover (LULC):** It emphasizes the type of activities and arrangements a stretch of land is employed for (e.g. agriculture, pasture, or habitation) thus directly affecting the region's proneness to landslides [22, 26].
- **Distance from Roads, Faults, and Rivers:** Proximity of the previous landslide sites to roadways, fault lines and waterways have revealed the crucial role played by these conditioning factors justifying their inclusion in landslide related research studies. A new factor of Distance from Railways indicating the proximity of locations to railroads has also been taken into account in Case Study I and was found to hold significance.

2.1.2 Forest Fires

Forests are an integral part of this earthly mosaic. Besides supporting the biodiversity, they also are a source of livelihood for millions. Also, they act as harbingers of employment and economic development globally [27]. However, with global climate change, bringing about extremities like rainfall deficit, seasonal irregularities, altering temperature patterns, and the incessant unsustainable human interventions have played havoc with the health of these valuable resources. One such aftereffect of these disruptions is witnessed in the blazing uncontrollable flames of wildfires. Wildfires are complex events diagnosed as fire in an area of combustible vegetation that is attributed to either natural or human factors [28]. Although fires can play an ecologically significant role in biogeochemical cycles and ecosystem functioning, they are considered to be a potential hazard with physical, biological, ecological and environmental consequences. They cause, nature imbalance endangering biodiversity by reducing the faunal and floral wealth. Other than the imminent deforestation and desertification, they cause huge economic losses for people that are dependent on forests as their only means of living, along with the reducing the agriculture and biodiversity to ashes, thus causing huge terrestrial distress. The causes of wildfires are varied like from lightning, volcanic eruptions, sparks from rock falls to those attributed to human activities. The probability of the start and spread of forest fire in a given forest under the presence of a certain set of conditions is known as wildfire susceptibility. A forest fire susceptibility zonation (FFSZ) map defined as an HSZ map aimed at predicting those specific forest locations where the fires are most or are highly likely to occur/initiate/spread. Historical, Geological,

topographical and environmental factors must be taken into account while constructing such maps. The historic aspect of the study is accounted for by the inclusion of a forest fire inventory that comprises of the coordinates of the locations that have witnessed wildfires in the past. While the other factors (geological, topographical and environmental) are collectively termed as Ignition factors.

2.1.2.1 Forest Fire Inventory

The foundation of FFSM is based on the relationship between the spatial coordinates of the historical forest fires and the ignition factors at those locations, hence the role of the forest fire inventory is indispensable [27]. A forest fire inventory indicates the spatial forest locations which in the past experienced wildfires.

2.1.2.2 Ignition Factors

The quality of forest fire prediction depends vastly on the parameters selected for prediction. As the causes of wildfires are varied, the same variety is reflected in the literature for the ignition factors considered for predicting forest fires. Also, the scale of the wildfire does not only depend on its source, but also the factors that would be responsible for its spread and aggravation. A few of the widely considered ignition factors are described below and the formal definitions of the ignition factors of slope, aspect, elevation, plan curvature, TWI, land cover and NDVI can be referred to in Section 2.1.1.2.

- **Slope:** It plays a crucial role in the progression of wildfire as fire tends to spread quickly up the slope as compared to down the slope, coupled with the fact that fire rise on steeper slopes due to the proximity with ground surfaces [29].
- **Aspect:** It is of immense consideration in FFSM for its correlation with an area's share of solar energy encountered [28]. Surfaces with South aspect tend to experience more sunlight, leading to higher temperatures, robust winds, and lower humidity levels.
- **Elevation:** Elevation has a direct impact on temperature, rainfall, moisture, the wind and indirect one on the vegetation and fuel moisture and hence has an effect on wildfire susceptibility and thus is considered an important ignition factor [29].

- **Plan Curvature:** The Curvature represents the morphology of topography and has been considered for the role it plays in the spread of fire [30].
- **TWI:** It represents the impact of hydrological conditions on an area's fire susceptibility [31].
- **Topographic Position Index (TPI):** The TPI reflects the difference in elevation of a focal cell and that of all of its neighboring cells [30] and is used to classify the regions on the basis of its landforms.
- **Climatic Factors:** The climatic factors have an enormous impact on an area's vulnerability to wildfires and hence the factors of Annual Temperature, Annual Rainfall, Wind Speed, Aridity Index, Relative Humidity and Potential Evapotranspiration have been included in Case Study II [27, 30, 31].
- **Soil Texture:** The Soil Texture has an indomitable role to play in the characterization of an instance of wildfire and has an indirect impact on its surrounding environment [30].
- **LULC:** The type of activity a piece of land and its surroundings are employed for, affects the chances for a probable forest fire and hence Land Cover has also been inculcated as an ignition factor [27, 30, 31].
- **NDVI:** NDVI being an indicator of the health status of vegetation is considered as an important ignition factor.
- **Distance to Rivers:** The proximity to rivers plays a direct role in the forest health [30].
- **Anthropogenic Factors:** The anthropogenic factors have long been affirmed to play a big role in the various wildfire incidents world over and therefore the distance from roads and habitations are imperative wildfire ignition factors [11, 27, 28].

2.1.3 Floods

Floods are globally notorious for causing extensive damages by inundating farms, settlements, croplands, villages, etc. and anything that comes in its way. As far as the causes of flooding are concerned, they are varied like flow rate exceeding the channel capacity of a water body, faults in dams triggered by an earthquake or otherwise, rivers overflowing their banks, storm surges in coastal areas associated with tropical cyclones,

tsunamis or high tides. Human casualities, losses to property and business, displaced cattle, submerged farmlands, comprehensive damage to livelihoods of those who survived, coupled with the irreversible mutilation of flora and fauna is all that remains after such an episode. The probability of flooding under the presence of a certain set of conditions is known as flood susceptibility. A flood susceptibility zonation (FSZ) map defined as an HSZ map aimed at predicting those specific locations where the floods are most or are highly likely to occur. Historical, Geological, topographical and environmental factors must be taken into account while constructing such maps. The historic aspect of the study is accounted for by the inclusion of a flood inventory that comprises of the coordinates of the locations that have witnessed flooding incidents in the past. Subsequently, all the other factors (geological, topographical and environmental) are collectively termed as flood conditioning factors.

2.1.3.1 Flood Inventory

The foundation of the FSM is based on the relationship between the spatial coordinates of the historic flood locations and the flood conditioning factors at those locations hence the role of the flood inventory is indispensable. A flood inventory indicates the spatial locations which in the past experienced floods [32].

2.1.3.2 Flood Conditioning Factors

As the causes of floods are varied the same variety is reflected in the literature for the flood conditioning factors considered for predicting floods. A few of the widely considered flood conditioning factors are described below while the formal definitions for slope, aspect, elevation, plan curvature, TWI, land cover and NDVI can be referred to in Section 2.1.1.2.

- **Slope:** The slope plays a crucial role in defining the areas submerged and magnitude of flooding.
- **Aspect:** The Aspect represents the region's geomorphological traits.
- **Elevation:** Elevation has a direct impact on temperature, rainfall, moisture, the wind and indirect one on the vegetation and fuel moisture.
- **Plan Curvature:** The Curvature represents the morphology of topography and has been considered for the role it plays in the magnitude of flooding.

- **TWI:** It has been included to account for the impact of hydrological conditions on an area's flood susceptibility and represents the accumulation flow [32, 33].
- **SPI:** It indicates the potential erosive power of the overland flow [32, 33].
- **Rainfall:** The rainfall is generally one of the most crucial factors due to flooding being rain-induced in the majority of the cases.
- **Soil Texture:** It is included to take geology of the region into account.
- **LULC:** The type of activity a piece of land and its surroundings are employed for, affects the chances of flooding and hence the Land Cover has also been inculcated as a conditioning factor [32].
- **NDVI:** NDVI being an indicator of the health status of vegetation is of utmost importance in influencing if a region is prone to floods.
- **Distance to Rivers:** The close proximity of the flooded areas from rivers can be observed from the many flooding inventories, thus the factor Distance from Rivers is also included as it signifies the impact in the spread and magnitude of floods [32].

2.2 Remote Sensing and Geographic Information Systems

For a very long time, satellites, wireless and remote sensors, UAVs (unmanned aerial vehicles), national and international meteorological and geographical departments, NGOs, various other international, government and private bodies have contributed data before, during and after a disaster [20]. A large part of data gathered from satellites, remote sensors, etc. have come in useful to accomplish tasks such as natural hazard susceptibility prediction [34]. These data collectively come under the domain of GIS. GIS is a computerized technology dealing with coordinate-based/spatial/remote-sensing data which comes in handy at various decision-making junctures [24]. Here the information is available in different layers pertaining to a particular location or stretch of land. The data are available for the attributes associated with coordinates for the region being considered. This data available as parameterized coordinates is exploited in the field of disaster mitigation comprising of disaster prediction, detection and management. Indian Space Research Organization (ISRO)'s Bhuvan, United State Geological Survey (USGS)'s Earth Explorer and Global Visualization (GloVis) are a few of the examples of

online applications that provide an easy access to the satellite data covering various earthly aspects which have been made use of for data gathering.

2.3 Machine Learning and Prediction Techniques

A wide assortment of techniques has been implemented in the past for hazard susceptibility mapping (HSM) [35]. These techniques can be predominantly be classified as Qualitative or Expert-based or Direct techniques and Quantitative or Indirect techniques as shown in Figure 2.1.

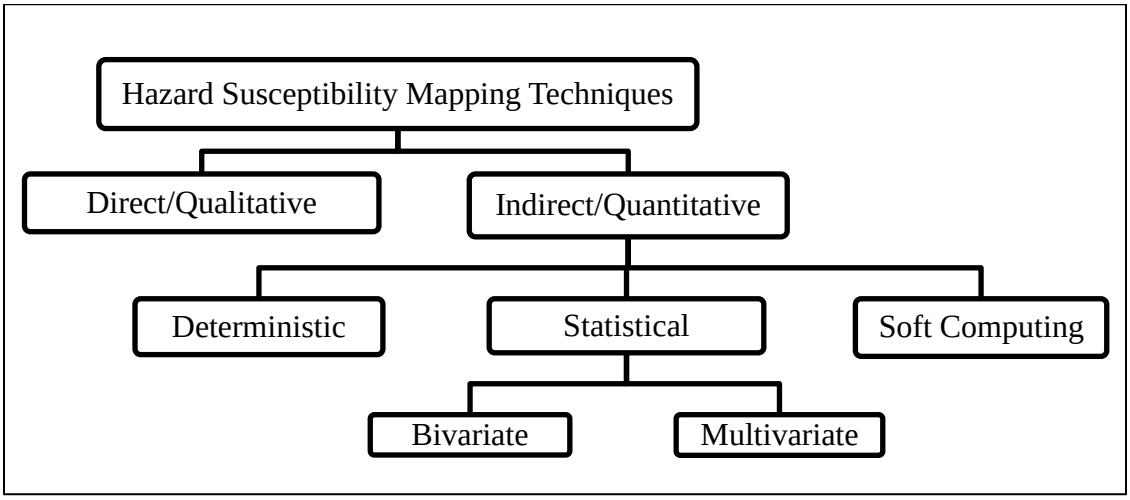

Figure 2.1: Taxonomy of Hazard Prediction Techniques

2.3.1 Direct and Indirect Techniques

Direct Techniques: The direct techniques rely on the knowledge and the expertise of a domain expert to ascertain an area's hazard susceptibility [5]. Due to the dependence of such a technique on a proficient, there is an aperture for the introduction of human errors. Such error could breach the accuracy of the susceptibility mapping.

Indirect Techniques: The Indirect techniques on the other hand base the predictions for a region's susceptibility on the information available for the study area [20]. This involves correlating the causative factors, by means of an abstract binding for the searching pattern in the data extracted. Such techniques are further classified as Deterministic, Statistical and Soft Computing [35].

- **Deterministic Techniques:** The deterministic techniques involve modeling an algebra based on absoluteness without any scope for approximations or randomness.

Such schemes are inspired by physical laws of conservation of energy, mass and momentum and involve expressing the susceptibility as a byproduct of interactions among the conditioning factors [35]. Due to the impenetrable nature of these techniques, such schemes impose a need for comprehensive data of the study area, hence, leaving such techniques unfeasible for large areas although retaining promise for small and homogeneous regions [36].

- **Statistical Techniques:** The statistical techniques employ combinations of variables for the purpose of predictions [37]. These techniques embark upon proposing a hypothesis for the statistical relationships among the variables and then make predictions based on the hypothesis. The hypothetical aspect of these techniques makes them prone to errors. However, there are corrective measures to handle the randomness and the bias introduced. These techniques can further be categorized as bivariate and multivariate [5].
- **Soft Computing Techniques:** Soft computing comprises of a synergistic amalgamation of methodologies that work together to model solutions for those real world problems that are too intricately complex to model and solve algebraically. They tend to provide approximate solutions to problems that do not have any algorithms available to provide exact solutions in polynomial time. Soft computing envelops fields of Machine Learning, Fuzzy Logic, Probabilistic Reasoning and Evolutionary Computation [35].

2.3.2 Bivariate and Multivariate Techniques

Bivariate Techniques: The bivariate statistical techniques in HSM involve overlaying the causal factor's thematic maps with the hazard's location from the inventory which weighs the causal factor's thematic map with the hazard density at the specific location leading to the determination of hazard susceptibilities at those locations. As the name suggests these techniques deal with bivariate data or data involving 2 variables. Hence such techniques revolve around establishing what relation exists between 2 sets of data. These differ from the regular 2 sample data analysis where the sample data are actually not directly related and hence not paired whereas here in the bivariate analysis the data would always be in pairs even if there is independence among them. Regression analysis

and correlation coefficients are the most popular approaches enlisted in these types of techniques. Both tend to analyze the relationship among variables on attributes such as the nature of the relationship, its type and direction, relationship's statistical significance and strength. A parameter often designated for above tasks is the Correlation Coefficient [36] which numerically quantifies the extent to which 2 attributes or variables are associated or related. The Pearson's Correlation Coefficient [36] is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2.5}$$

Here x_i is the i^{th} value for the variable X and similarly, y_i is the i^{th} value for the variable Y . $\overline{x}$ and $\overline{y}$ refer to the respective variable means. n is the number of sample pairs in the analysis. The literature review reveals Weight-of-Evidence (WoE), Frequency Ratio and Information Value as the few of the most frequently enforced strategies in the field of HSM. Following is a brief insight into these approaches.

- **WoE:** It apprises of the prognostic power of an independent variable in influencing the values of the dependent variable [13]. In the case of HSM, the model is expressed in terms of conditioning factors leading to hazard or non-hazard points/events. The modus operandi followed through encompasses splitting data on each factor into parts. This is followed by calculating the count of hazard and non-hazard events in each part created [37]. Employing the above count the percentage of hazard and non-hazard events in each of the parts is determined. The WoE is ultimately furnished by employing the formula as mentioned in Equation (2.6).

$$WoE = \ln\left[\frac{(Percentage\ of\ non-hazard\ events)}{(Percentage\ of\ hazard\ events)}\right] \tag{2.6}$$

  It allows a transition of a continuous independent variable into a set of classes based on the resemblance of dependent variable distribution, i.e. the number of events and non-events.

- **Information Value:** Information Value accentuates the work of WoE model by prioritizing the variables on the basis of their predictive powers in a predictive model [35]. It is calculated as given in Equation (2.7).

$$Information\ Value = \sum(\%\ of\ non-hazard\ points - \%\ of\ hazard\ points) \times WoE \quad (2.7)$$

Multivariate Techniques: These techniques, on the other hand, weigh the conditioning factors on the basis of their relative degree of contribution to the hazard. Logistic Regression (LR), Cluster Analysis and Discriminant Analysis come under the domain of multivariate techniques [35].

- **LR:** The aim of such a technique is to elucidate the functional relationship between the independent and dependent variables. It is used when there is one dependent variable and two or more independent variables. The technique is used to determine the probabilities of the dependent variable and prioritizing the independent variables on the basis of the dominance they have on the dependent variable. The aforementioned task is accomplished by deducing the equation that best predicts a particular value of the dependent variable Y on the basis of particular values of independent variables X. The final outcome of the procedure would be an equation as specified [38]:

$$\ln\left[\frac{Y}{(1-Y)}\right] = a + b_1x_1 + b_2x_2 + b_3x_3 + .. \quad (2.8)$$

  Here in the slope $b_1, b_2, b_3, ..$ and the intercept a is calculated for the best fitting equation using either the least squares method or the maximum likelihood method. Here maximum likelihood method is an extensive computation based scheme which those values of the independent variables which are most likely to give the desired value in the dependent variable.

- **Discriminant Analysis:** Discriminant analysis predicts the group a particular dependent variable would belong to on the basis of the value of the independent variables. Such a prediction is made on the premise of group's prior information availability corresponding to the independent variable's value. Hence it establishes the demarcations among variables which discriminate the dependent variable into the existing classes. Algebraically, it resembles analysis of variance (ANOVA) [39]. For example, in a random sample of heights of 50 males and 50 females it is observed

that the mean of height of females, less than the mean height of males (assumption). So the discriminating independent variable the "height" allow us to differentiate between males and females with a better chance probability. Thus, it enables classification into groups and proclaims group memberships using the averages of variables.

- **Cluster Analysis:** Unlike discriminant analysis which alludes to an apriori grouping of the dependent variable on the basis of the independent variables, clustering has no prior classification information in hand to assist in fresh segregation. It, in fact, has no hint of the classes existing or even of the number of groups that might co-exist [40]. Rather, it attempts at demarcating the groups on dependent variable based on values of independent variables such that intragroup similarity is maximized while the inter-group similarity is minimized. A number of metrics have been established in past to quantify uniformity or the similarity amongst the elements. The most frequently used is the distance metric which in itself generally alludes to Euclidean distance [40]. The techniques of agglomerative and divisive clustering enveloped under the category of Hierarchical Clustering whereas schemes like k- mean clustering come under the category of Non-Hierarchical Clustering.

2.3.3 Soft Computing Techniques

- **Fuzzy Logic:** It is a method of reasoning which does not demarcate outcomes on the hard lines of 0 and 1 or $true$ and $false$ rather is inclusive of various degrees of truth represented with values lying between 0 and 1 [21].

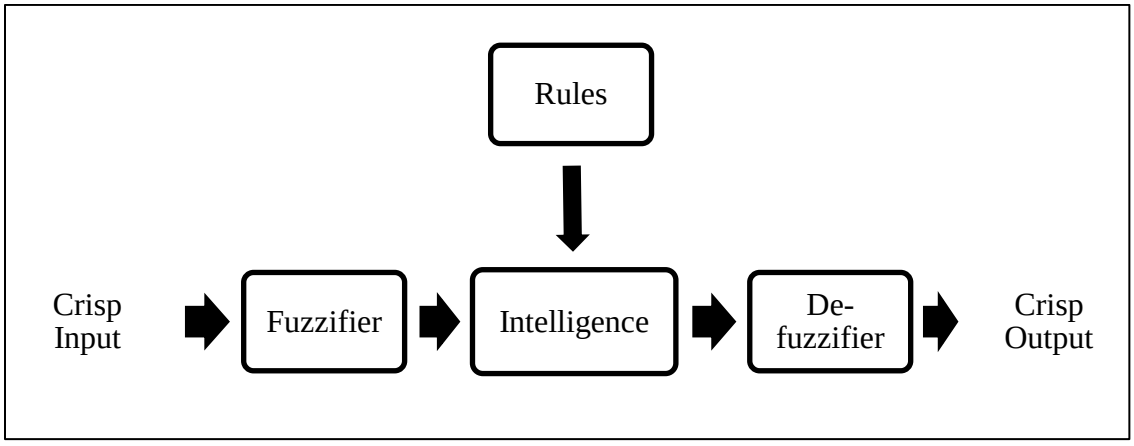

Figure 2.2: Fuzzy Logic System Architecture

So it simulates human thinking and reasoning with decent credibility. As a result, it is able to furnish outcomes in response to inputs whose completeness and accuracy is questionable generating acceptable results from ambiguous or distorted data. Unlike other techniques which have an algebraic essence as their founding base, fuzzy logic is based on "if then else" rule approach. The quantitative factors of relevance here are the error and the rate of change of error which again come with inherent flexibility. These values transition into critical only when a real-time performance is one of the pre-requisites. These involve highly descriptive languages where the input handling mechanism resembles that of a human mind. The basic architecture comprises of the four modules as depicted in Figure 2.2. The fuzzifier is responsible for the transformation of data into fuzzy sets such as large positives, medium positives, small, medium negatives, and large negatives. The intelligence module comprises of all the if-then rules drafted by the experts of the field under consideration. Also, it is this module's responsibility to follow the rules available on the inputs to generate the outputs. The de-fuzzifier is responsible for the transformation of outputs obtained from intelligence module into proper form. Fuzzy Logic for HSM comprises of obtaining fuzzy membership values for all the causative factors that have been normalized to lie between 0 and 1 and then applying various fuzzy operators for obtaining HSZ maps [21].

- **Machine Learning:** Machine learning is the branch of computer science and a principal constituent of Soft Computing which emphasizes on simulating the human thought process for forging out crucial decisions [35]. It delves into the domain of automation for prediction by analysis, self-learning and self-evolution in the context available. In HSM it takes upon the hazard inventory and the corresponding hazard causative factor values for locations in the inventory as inputs and on the basis of these, it determines the susceptibility of regions against the specific hazard [41]. It then focuses on developing mappings from causative thematic maps to hazard inventory and then using those generated mappings to predict a new region's hazard susceptibility. The most commonly used machine learning algorithms for HSM are as follows:

- **Artificial Neural Network (ANN):** These are used to model/simulate the distribution, functions or mappings among variables as modules of a dynamic system associated with a learning rule or a learning algorithm [36].

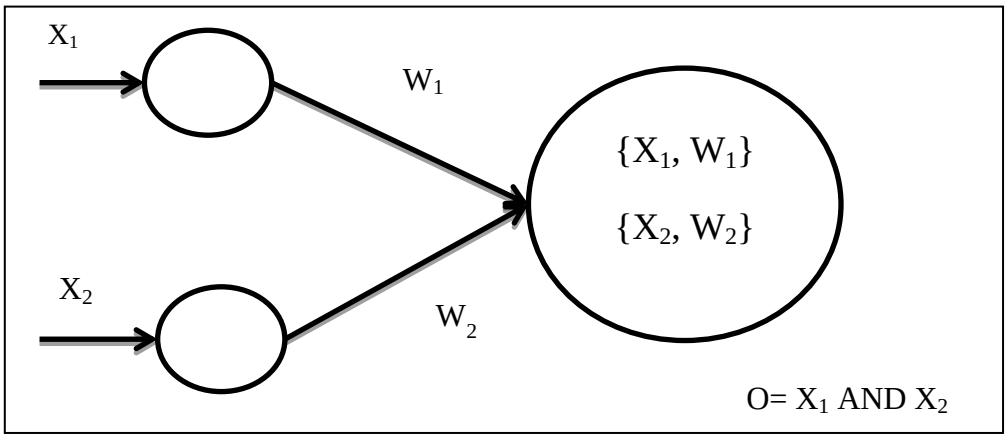

Figure 2.3: AND ANN

The modules here simulate neurons in the nervous system and hence ANN collectively refers to the neuron simulators and their synapsis simulating interconnections between these modules in different layers [36]. The defining aspect of an ANN is the function implemented at each neuron and the learning algorithm for the dynamic weights assigned to the interconnections among neurons. Figure 2.3 depicts a sample ANN simulating the AND logic. In simple terms, it is a computation system which can be represented as an annotated graph whose nodes symbolize modules performing certain computations.

Here the inputs to the network are represented as $X_n$which in the case of HSM would be causative factors. Each of these inputs is then multiplied by connection weights represented as W_n. These products, hence obtained are added and then fed as input into the transfer function. An essential characteristic of this, as represented in the form of a directed graph is that unlike other graphs the weights allocated on the edges do not remain static rather are dynamic giving it the learning capability it is so favorably used for [36]. What makes it stand apart is its ability to simulate human thought process coupled with continuous learning, growth and evolution. Also, it is capable of handling a large number of parameters and a large set of data with noise and yet achieves high accuracy.

- **Support Vector Machine (SVM):** It constructs a hyper-plane as shown in Figure 2.4 i.e. a plane in an infinite dimension to classify the training data points into clearly demarcated classes [36, 41].

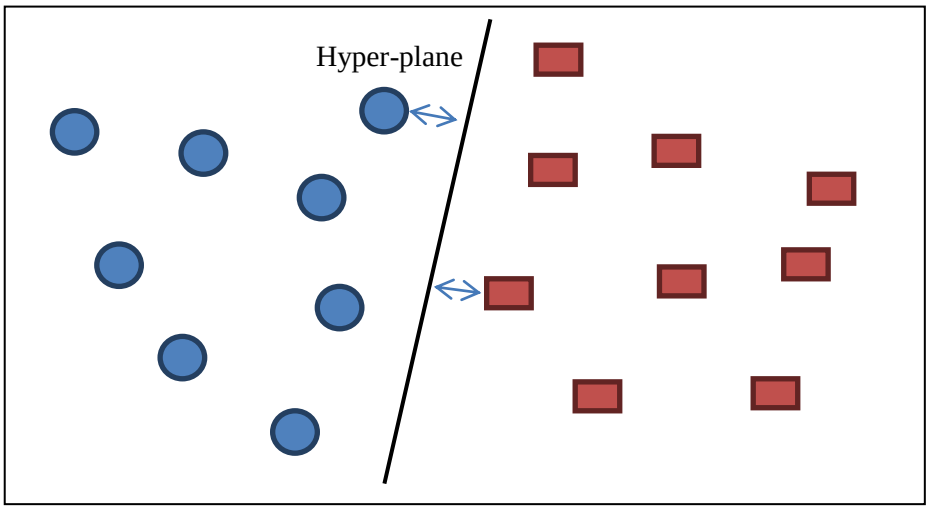

Figure 2.4: Hyper-plane in 2 dimensions

The construction of an optimal decision plane for classification requires minimizing the error function. The shape of the error function becomes the foundation for further classification of these algorithms in the broad categories of linear, polynomial, sigmoid and radial SVM. In simple terms, the philosophy of SVM is to obtain an optimal hyper-plane for data points which are linearly separable. Support vectors actually refer to the data points that are closest to the demarcating surface which are hence tricky to classify. The metric that alludes to the optimality of a hyper-plane is the margin around it. So the problem transitions into that of an optimization one. As established the maximum margin classifier learnt and derived from the training data would lead us to optimal hyper-plane. This is achieved by transforming the maximal margin classifier as the inner product (sum of multiplication of pair values) of two given data points rather than the data points. The general kernel function could then be defined as follows [20] where x is the new input vector and the coefficients $B_0$and a_i must be estimated from training data.

$$f(x) = B_0 + \sum_i \left(a_i \times (x, x_i)\right) \tag{2.9}$$

- **Decision Tree (DT):** In this technique, the whole data set or the whole collection of sample points is split into two or more homogeneous classes [24]. The split is established from the parameter or the factor which is determined to be the best splitter or a differentiator. There are various metrics that are used to determine a parameter's splitting power e.g. GINI index, Information gain, etc. GINI Index is given as [38]:

$$GINI(D) = 1 - \sum p_j^2 \tag{2.10}$$

Here p_j is the relative frequency of class j in D(data). If D is split into $D_1$and D_2 with $n_1$and n_2 tuples each.

$$GINI_{split} = \frac{n_1}{n} \times gini(D_1) + \frac{n_2}{n} \times gini(D_2) \tag{2.11}$$

- **Naïve Bayes (NB):** Rather than a single classifier, it actually is a combination of multiple classifiers, all working on the basic NB principle of independent features [24]. Hence, each feature is assumed to be independent and autonomous contributing individually to the training data point's probability of belonging to a particular class.

As per the Bayes theorem [38],

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$
$$P(c|X) = P(x_1|c)P(x_2|c) \times \ldots \times P(x_n|c)P(c) \tag{2.12}$$

Here $P(c|x)$ is the posterior probability of class given predictor, $P(c)$ is the prior probability of class, $P(x|c)$ is the likelihood probability of predictor given class, and $P(X)$is the prior probability of predictor.

- **Random Forest (RF)**: RF or random decision forests are an ensemble learning method for classification that operates by constructing a multitude of DTs at training time and outputting the class that is the mode of the classes (classification) of the individual trees [42]. RF fare better than DT by overcoming their shortcoming of over-fitting by pruning via the construction of multiple DT thus allowing for better prediction accuracies and performance results. The way for the ensemble of DT to perform better than any of the participating individual trees of the ensemble is to have the diversity among the participating trees. RF

ensemble techniques ensure this diversity by means of re-sampling over the data through the means of replacements and by changing over the predictive variable sets over each generation of individual DT. Major advantages of RF are owing to the fact that it is resistant to overtraining and even creating an exorbitantly large number of RF trees does not endanger the predictive performance by the risk of over-fitting due to the fact that each individual DT which is part of the ensemble is, in fact, an entirely independent random experiment [42].

2.4 Comparative Performance Metrics

For a comparative analysis of the proposed methods against the conventional benchmark models such as ANN, RF and SVM, it is imperative to define the performance statistics on the basis of which the relative performance of the models is judged and evaluated. In pursuance of contrasting the approaches for Hazard prediction against the most commonly used techniques, the parameters associated with ROCs are used. These include Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value, AUC and Accuracy. TP refers to the True Positives implying those locations in the testing dataset that witnessed the specific hazard (landslide, forest fire, flood etc.) and were predicted as Hazard location. TN refers to the True Negatives implying those locations in the testing dataset that did not witness hazard and were predicted as Non-Hazard location. FN refers to the False Negatives implying those locations in the testing dataset that did witness hazard and were predicted as Non-Hazard. FP refers to the False Positives implying those locations in the testing dataset that did not witness hazard and were predicted as Hazard. The Receiver Operating Characteristics (ROC) curves are generated using the plots of ratio values of true positives against false positives with various cutoff thresholds.

$$Sensitivity = \frac{TP}{TP + FN} = \frac{True\ Hazards}{True\ Hazards + False\ Non - Hazards} \tag{2.13}$$

Sensitivity represents the percentage of correctly classified Hazard points among all the actual Hazard points. Sensitivity or true positive rate or true hazard rate indicates the locations that have experienced hazard and also have been predicted as hazard-prone.

$$Specificity = \frac{TN}{TN + FP} = \frac{True\ Non - Hazards}{False\ Hazards + True\ Non - Hazards} \tag{2.14}$$

Specificity represents the percentage of correctly classified Non-Hazard points among all the actual Non-Hazard points. Also, specificity or the true negative rate or true non-hazard rate indicate the locations that have not experienced hazard and also have been predicted as non-hazard-prone.

$$Positive\ Predictive\ Value\ (PPV) = \frac{TP}{TP + FP} \tag{2.15}$$

$$Negative\ Predictive\ Value\ (NPV) = \frac{TN}{TN + FN} \tag{2.16}$$

The Positive Predictive Value (PPV) indicates the probability of points classified to Hazard whereas Negative Predictive Value (NPV) indicates the probability of points classified to Non-Hazard.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2.17}$$

The overall success rate or the accuracy is the total number of correctly predicted Hazard and Non-Hazard points divided by the number of all the points in the dataset. The gist of the ROC analysis is expressed in the metric of the area under the ROC curve (AUC). A predictive model having an AUC value nearing or equal to 1, being able to produce a sufficient number of correct predictions against the incorrect ones is considered informative. On the other hand, an AUC nearing or equal to 0, being unable to produce a sufficient number of correct predictions against the incorrect ones is considered non-informative. Other than the metrics mentioned above, the models are also compared on the basis of their Kappa index(κ). It is also used to measure and compare the predictive capability of a classifier. It is generally considered as a robust measure because of its inherent capability to take correct prediction that occurred by chance into account and exclude it from the predictor's overall ability. It is considered as a perfect agreement between observed hazard and predicted one if κ lies in the range of $0.8 - 1$, a substantial agreement for κ in the range $0.6 - 0.8$, moderate for κ in $0.4 - 0.6$, fair in $0.2 - 0.4$, slightly for $0 - 0.2$, and poor for range≤ 0.

CHAPTER 3

LITERATURE REVIEW

This chapter describes and summarizes the most relevant latest work that has been accomplished in the field of hazard susceptibility assessments.

Akgun et al. [37] conducted a study in one of the most landslide-prone regions in Turkey. The purpose of this study was to produce the LSZ maps for the eastern part of the Black Sea Region of Turkey. The landslide conditioning parameters considered include slope angle, land cover, lithology, and distance from roads and drainage lines, and slope aspect. The zonation maps were produced by employing the models of likelihood ratio model (LRM) and the weighted linear combination (WLC) model. The study proved that the WLC model fared better than LRM model.

Akgun and Turk [26] focused on dissecting the regions in Ayvalik, western Turkey on the basis of their landslide susceptibility using multi-criteria decision analysis. The parameters brought under consideration were SPI, slope gradient and aspect, lineament density, lithology, TWI, distance from drainage, weathering states of the rocks, and land cover and vegetation density. The degree of influence that each of the conditioning parameters has on the landslide susceptibility were computed by employing the Analytic Hierarchy Process method and the outcomes obtained signified the measure of the impact degree in terms of weights. The derived values were designated to their respective parameters and hence these suffused parameters gave the LSZ maps.

Ghosh and Bhattacharya [43] focused on developing an LSZ system comprising of feeding-in-data, connotation, expert and output modules. Here input module intakes the thematic images of conditioning factors. Expert module comprised of knowledge plinth and interpretation scheme categorizing regions into different zones depending on their respective susceptibilities. The output module produced the zonation map. The system was tested for a certain area in the state of Uttarakhand. The contributing factors considered here are lithology (rock and soil type), slope morphometric, LULC, rainfall etc.

Kundu et al. [5] focused on the generation of comprehensive LSZ map of the landslide prone area of Ganeshganga watershed. The factors of slope, aspect, lithology, tectonic

structures, relative relief, lineaments, LULC, etc. were considered. The binary LR model was used for completion of the work pertaining to the above tasks. Kumar and Annadurai [44] included an LSM in their study using the ordinal scale (qualitative) approach involving weighting-rating system or terrain parameters (ground-based knowledge). The stated study demarcated the susceptibility zones in the Kothagiri region of Tamil Nadu. The causative landslide factors selected in this study are Geology, Lineament Density, Drainage Density, Geomorphology, Soils, Drainage, Slope, and LULC.

Ahmed [34] in their work carried LSM for the Chittagong Metropolitan Area (CMA), Bangladesh. The factors considered are elevation, NDVI, slope, etc. The final result of the study involved the seven LSZ maps for the CMA area. The techniques used are Analytical Hierarchy Process, three variants of WLC, and three variants of Ordered Weighted Averages. The resulting maps were then compared on the basis of ROC Analysis.

Hong et al. [41] in their work prepared LSZ map for a city located in the Jiangxi province of China known as Luxi. The factors considered in this study are altitude, SPI, aspect, distance from river, slope, lithology, etc. The four kernel types of the SVM machine learning models, namely linear kernel, radial basis function, polynomial kernel, and sigmoid kernel were implemented to achieve the above-stated objective and compared using the prediction rates in the validation phase.

Bui et al. [36] in their work focused on the LSM for Son La hydropower basin (Vietnam). The landslide conditioning factors considered in here are slope, TWI, aspect, rainfall, altitude, relief, SPI, fault density, lithology amplitude and LULC. The Information Gain Ratio measure was used for the purpose of feature selection. The machine learning models employed in the mapping process were SVM, multilayer perceptron ANN, kernel LR, and logistic model trees. The models were then assessed using the ROC, Kappa index, and several statistical evaluation measures.

Hoang and Bui [45] in their work, the performance of the relevance vector machine classifier model have been compared against those of ANN and SVM for deciphering the landslide spatial prediction problem. This also involved cuckoo search optimization to ascertain the base model’s tuning parameter that is the basis function’s width. The

landslide causing factors employed here included slope, distance to roads, rivers and faults, relief amplitude, lithology, TWI, aspect, and rainfall.

Pham et al. [24] in their piece of study focused on the landslide susceptibility assessment for the area of the state of Uttarakhand, India. The landslide causative factors employed in the study for elucidating the spatial relations among these are slope angle, curvature, elevation, slope aspect, soil, LULC, distance to roads, lineaments, and rivers, lithology, and rainfall. The machine learning models compared for their performance on landslide susceptibility assessment are NB, MLP ANN, and Functional Trees. The linear SVM algorithm was employed for feature selection among the landslide conditioning factors on the basis of their predictive capability on landslide occurrence.

Pham et al. [21] in their work, used an ensemble model called Forest fuzzy rule based Classifier Ensemble for the spatial prediction of landslides by generating a landslide susceptibility map for certain areas in Uttarakhand state. The factors included are slope angle, elevation, curvature, profile curvature, soil type, plan curvature, slope aspect, LULC, etc. The results produced were then compared against the conventional machine learning models of Fuzzy Unordered Rules Induction Algorithm, SVM, AdaBoost, Bagging, MultiBoost. An important finding from this study was that road is the most crucial element responsible for landslides as it perturbs the natural environment.

Colkesen et al. [20] implemented the generation of LSZ maps for a district named Tonya in the region of Trabzon in Turkey. The conditioning factors considered responsible for a landslide that were selected for the study are slope angle, NDVI, lithology, LULC, profile curvature, slope aspect, elevation, and TWI. The methods employed for the LSZ were Gaussian Process Regression and Support Vector Regression. The results produced by the aforementioned methods were then against those of LR which has been the de facto method for the LSM right up till now.

Melchiorre et al. [40] performed LSM for the landslide-prone region Brembilla Municipality in the Southern Alps, Italy. LSM has been achieved by a culminated effort of ANN and Cluster Analysis. Cluster Analysis was used for a better utilization of landslide inventory by means of management of training and testing records while the actual mapping in the causative factors to inventory was effectuated by means of ANN. The study concludes with validating the enhancement in the predictive performance and

robustness of the black-box based approach of model building by the introduction of expert knowledge (Clustering in this case).

Moosavi and Niazi [22], applied a two-level approach to generate the LSM for the region of Ilam dam, Ilam Province, Iran. Firstly the LSZ maps for the study region were produced using approaches, namely ANN, SVM, maximum entropy, and generalized linear model. Then the generated maps were fed as inputs using different mother wavelets in different levels. Then, the hybrid models were developed using the wavelet-based preprocessed maps. The study signified that the wavelet packet, effectively enhanced the performance of individual classifiers with the blend of Wavelet-packet with SVM outperforming others.

Ding et al. [46] in their study implemented LSM for Taibai County (China). The results were obtained through the integration of landslide conditioning factors using the approaches of Frequency Ratio, WoE and Evidential Belief Function models and were compared, with Frequency Ratio having outperformed the other employed techniques.

Jaiswal et al. [11] in their study selected the forest fire prone area of Gorna Subwatershed in Madhya Pradesh for identifying the fire risk zones through the process of FFSM. These zones were delineated by assigning subjective weights to the classes of all the ignition factors according to their sensitivity to fire or their fire-inducing capability. The study identified annotated around 30% of the study area to be under very high and high-risk zones. Also, the original Wildfire inventory coincided with the obtained high and very high-risk zones.

Bui et al. [27] proposed and validated a new and unique hybrid approach of Particle Swarm Optimized Neural Fuzzy in employing the FFSM for the tropical forest at the province of Lam Dong (Central Highland of Vietnam). In the proposed approach the prediction was obtained using the classifier model of Neural Fuzzy whose parameters were optimized using the meta-heuristic optimization technique of Particle swarm. The proposed model was compared against RF and SVM and outperformed them on various performance parameters.

Adab et al. [28] in their study delineated the fire risk ones in the fire-prone areas of northeast Iran, using two of the widely used indexes of Structural fire index and Fire risk index along with a new index of Hybrid Fire index. These indices were set up by

assigning subjective weight values to the classes of the ignition factor layers based on their sensitivity ratio to fire. These indexes were validated against the hot spots data derived from Moderate-Resolution Imaging Spectroradiometer (MODIS) satellite sensor. The proposed index outperformed the other two indexes and predicted an area of around 57.5% of the study region to be under high-risk zones.

Gao et al. [29] in their work evaluated Huaguo Mountain Scenic Spot in Liangyungang under the FFSM. Using the geographical ignition factors and fuel type characteristics the study region was annotated as per the relative degree of fire risk via cluster analysis. The characteristics of fire source were modelled by Fuzzy multi-level synthetic evaluation. The study revealed a possibility for the realization of short time and high spatial resolution forecast in a similar region.

Pourghasemi [30] in their study produced FFSM for the region of Minudasht Forests, Golestan Province, Iran based on Evidential Belief Function and Binary LR models. The study inculcated the causative factors of slope, aspect, elevation, plan curvature, TWI, TPI, land cover, NDVI, and proximity to villages, roads, rivers and wind effect, temperature, rainfall etc. The Evidential Belief Function fared better than the other. Bui et al. [31] carried out the FFSM for the Cat Ba National Park area (Vietnam). It employed the Kernel-LR model for the mapping and compared the model's performance against SVM with the Kernel-LR outperforming SVM. The Model displayed good performance with a predictive capability of 92.2%.

Eugenio et al. [47] in their study developed a statistical model for FFSM using GIS for Espírito Santo State. The study classified the ignition factors into two categories of physical (slope, aspect, land cover, etc.) and climatic (rainfall, wind etc.) and accordingly assigned them weights. The study thus assumes a suitable adjustment of weights on the basis of the region where the model is being replicated. The study revealed 41.5% and 3.40% of the study area to be under high and very risk against forest fires.

Arpaci et al. [48] tested the applicability of two machine learning algorithms of Maximum Entropy and RF for their employability in the process of FFSM for the Alps fire-prone region of Tyrol in Austria. The results of the techniques were also compared and found to deliver coherent results. The study conclusively proves the dominance of climate and anthropogenic factors over other ignition factors. Satir et al. [49] discussed a

novel the FFSM approach for the Mediterranean forestland using multiple data assessment technique employing the Multi-layer perceptron approach based on back propagation algorithm of ANN. The fire-prone region of Seyhan Basin in Turkey was selected for the case study. The model was proved to have performed well, achieving an accuracy of 83%. Also, the Pearson's correlation coefficient was used to determine the highly correlated ignition factors which were found to be elevation, tree cover and temperature.

Suryabhagavan et al. [50] dealt with the identification of fire-prone areas in the Harenna forest using RS and GIS techniques. The study adopted the Multi-criteria decision-making technique for FFSM employing physiographic data and proximity factors, such as elevation, slope, aspect, vegetation type, proximity to settlements and distance from roads. The study classified an area of about 24% of the study region to be belonging to the high-risk category of Forest fire.

Tehrany et al. [32] applied ensemble technique for Flood susceptibility modeling in the region of Terengganu, Malaysia. An ensemble of weights-of-evidence (WoE) through bivariate statistical analysis for assessing the impact of flood conditioning factors and finally feeding these as weights to SVM model for prediction was employed. All the kernel types of SVM (linear kernel, polynomial kernel, radial basis function kernel, and sigmoid kernel) were employed for their comparison in performance while being ensembled. The study validated the superior performance of the ensemble against individual models signifying its scope and applicability in the field.

Khosravi et al. [33] implemented FSM by applying different bivariate statistical models, namely Shannon's entropy, statistical index, and weighting factor for Horace Watershed, Mazandaran Province, Iran. The factors employed are slope, plan curvature, etc. The study established the Statistical Index model as a better classifier than the others, having achieved an accuracy of 98%. Also, it demonstrated the distance from the rivers as the most dominating factor for the study region.

Lee at al. [51] conducted FSM for the flood-prone region of Busan in Korea. The FSM was realized through Frequency Ratio model, which allowed an in-depth assessment of flood-susceptible areas. It also assessed the flood-related factors for the study region. The model achieved an accuracy of 91.5% for the study region demonstrating promise for this

field of research. The FSZ map for the region was also generated as one of the final outcomes of this study.

Pradhan [52] presented the construction of an FSZ map for presumptive flood areas in and around in the Kelantan river basin in Malaysia using a statistical model and GIS. The study employed LR model for the determination of each conditioning factor's (topographical map, altitude, geological map, rainfall, hydrological map, Global Positioning System (GPS) data, and land cover map) rating, followed by overlaying the ratings for FSM. The results obtained delineated the high-risk areas that stand the highest chance of being inundated in the case of flooding conditions. Qualitatively, the model resembled reasonable results with an accuracy of 85%. Tehrany et al. [53] compared the prediction performances of two different approaches such as rule-based DT and the combination of Frequency Ratio and LR statistical methods for FSM at Kelantan, Malaysia. The study employed flood conditioning factors of altitude, TWI, SPI, etc. The ensemble of Frequency Ratio and LR (90%) outperformed DT (87%) in the prediction of flood-susceptible areas.

CHAPTER 4

PROBLEM STATEMENT

4.1 Problem Statement

Natural disasters are major inimical events which cause humongous loss of life and property on very wide scale spreading across a spectrum of topographies. In a developing economy like India, the human activities have not essentially been sustainable leading to circumstances that have increased the magnitude and frequency of such disasters that would have been easily subverted had it not been for the extensive damage caused to nature and its balance not been impaired. However, there is a unanimous agreement in the research community regarding the necessary efforts for the attenuation of the scale of the disaster-induced damage with adequate preparation and planning. Such contingency plans required for an in-depth assessment of hazard prone areas. This study aims to undertake a few of these hazard susceptibility assessments in identified hazard-prone regions on the Indian soils. The focus is to prepare high precision susceptibility maps for the regions selected by employing state-of-the-art machine learning techniques enhanced with certain novel tactics of optimizations and ensembling. It is aimed that such maps would empower the concerned agencies with information that could aid in their disaster mitigation and management efforts.

4.2 Research Gaps

Through the literature survey, the following gaps have been identified

- The general pattern observed in the studies indicates a void for the optimization techniques like those of the ACO; PSO, etc. which if employed would significantly improve the accuracy of the outcome.
- There is scope for a tailor made ensemble model for HSM, which might endeavor to achieve higher accuracies.
- The HSM scene in India has not really taken off due to the limited studies available on Indian lands, even with repeated occurrences of landslides, avalanches and floods in certain areas.

- There is no consensus in the Geo-research community on the best HSM technique [26]. While selecting the prediction method, the availability and the accuracy of the available data needs to be appraised.
- An all factor inclusive study is yet to be undertaken. As established different studies have been conducted considering different hazard conditioning factors. The factors employed generally are influenced by the availability of data for the region under study. Also the synergy existing among these factors varies from place to place. Hence there is an irrefutable absence of a universal scheme of things to be carried for a HSM. Each region would hence have to undergo a complete new HSM considering parameters and relations between them, without any scope for replication or reusability.
- All the HSM studies undertaken yet have taken conditioning factors like LULC and Rainfall without much consideration to the fact that these factors are in fact dynamic variables which even more so, of late have the tendency to undergo drastic changes in short span of time. So there arises the need to take the future prospective changes in these factors into account.

4.3 Aims and Objectives

The objectives of this thesis are

1. Study the existing hazard susceptibility assessment techniques.
2. Design and develop models for
    - LSM using Majority Based Voting ensemble of LR, Gradient Boosted Decision Trees (GBDT) and Voting Feature Interval (VFI).
    - FFSM using Evolutionary Optimized GBDT.
    - FSM using PSO-SVM.
3. Verification and validation of the proposed models
4. Comparison of proposed models with existing machine learning models such as SVM, RF, ANN etc. on the basis of their accuracy, specificity, sensitivity and AUC.

4.4 Proposed Methodology

The summarized general methodology followed for the case studies is as shown in Figure 4.1. The first stage of the study comprised of a thorough literature survey to identify the parameters which have been considered as the causative factors of the corresponding hazard. As per the findings, the data are collected and brought in an appropriate spatial format to be utilized. Here the thematic maps are prepared, corresponding to the necessary hazard conditioning factors and the hazard inventory.

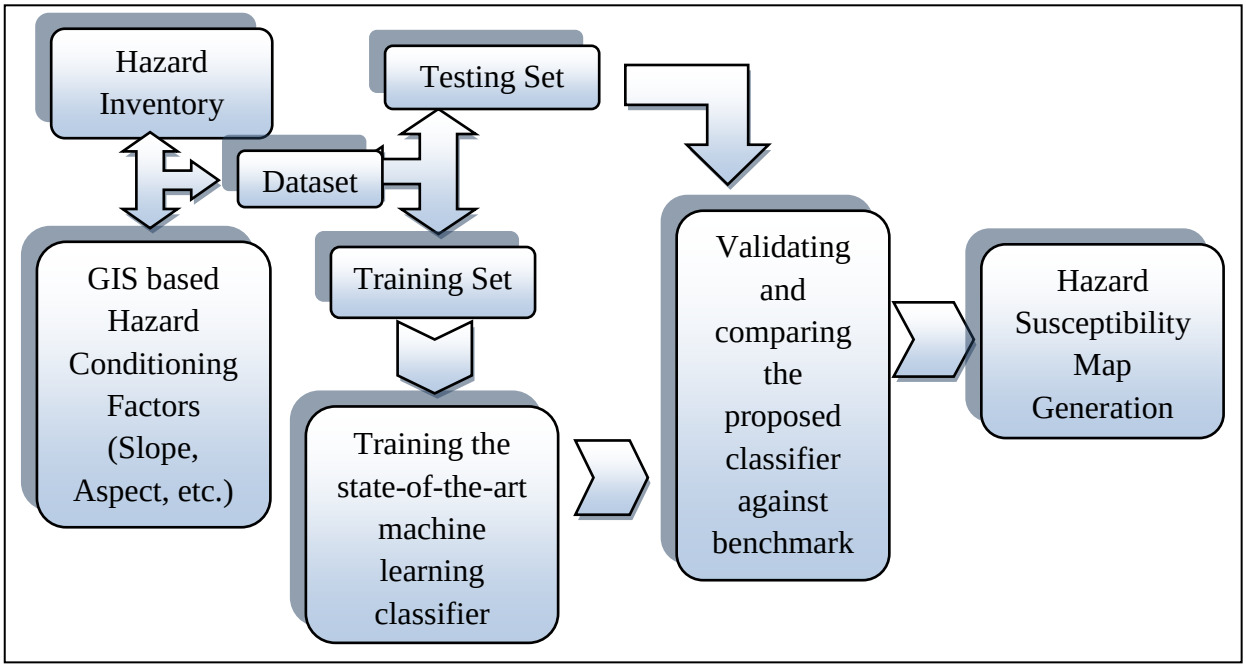

Figure 4.1: Proposed Methodology for HSM

In the next stage, these factor thematic maps are annexed over the inventory to obtain a dataset comprising of the hazard point coordinates (along with non-hazard points for balanced a dataset and avoiding bias) and their values for all the causative factors. Next, the generated dataset is randomly split into 70%-30% ratios as training and testing sets respectively. The training set is employed for training the classifier along with other benchmark models such as SVM, DT etc. The testing dataset is used to obtain predictions from the generated models. In the fourth stage, the performance of the built models was compared on the basis of the predictions in the testing dataset by employing prediction evaluation metrics, like those of accuracy, sensitivity, specificity etc. Finally, in the fifth stage, the validated model is used to build the HSZ map for the region.

CHAPTER 5

CASE STUDIES

The proposed methodology has been implemented in the following case studies in different regions of India on the basis of their proneness to different hazards. With North-Eastern region of India (parts of states Assam and Nagaland) being prone to landslides has thus been inculcated for landslide susceptibility assessment, while the state of Uttarakhand in general and NDBR in particular, having experienced multiple wildfires in 2016 has thus been considered for an FFSM and finally, the Chamoli district of Uttarakhand has been considered for an FSM due to the rampant floods that the region witnessed in 2013 and 2016.

5.1 Case Study I: LSM (North-East India)

The inherent complex topography and drastic weather patterns together have concocted various natural disasters worldwide. In difficult terrains like those prevalent in the North-Eastern regions of India, coupled with the population explosion and improper land use, intertwine rivetingly, leading them to witness some of the world's most drastic landslides with an astonishing frequency, reckoning landslide susceptibility assessment crucial in such regions. This case study focuses on exploring a promising machine learning ensemble technique of Majority-based voting, which has seldom been employed for landslide susceptibility assessment. The ensemble comprises of LR, GBDT and VFI to prepare LSZ map for the Brahmaputra valley region (Assam & Nagaland) and its close vicinity. The map indicated 6.8% (1360 km^2) of the area falls under very high landslide susceptibility while 7.2% (1440 km^2) under high landslide susceptibility zone.

5.1.1 Study Area

The area selected for this case study is located in the North-Eastern part of India enveloping parts of the Brahmaputra valley as shown in Figure 5.1. It is comprised of the north-eastern part of the state of Assam encompassing districts of Jorhat, North Lakhimpur, Dibrugarh, Tinsukia and Sivasagar along with adjoining Longleng and Mon districts of the neighboring state of Nagaland. The study area lies between 26°31′16″ N and 27°54′34″ N latitudes and 93°28′36″ N and 95°34′26″ N longitudes covering an area

of about 20, 000 km^2. The districts of Tuli, Tizit, Longleng and Mon have been included in the study area for this work due to recurring episodes of landslides being experienced here.

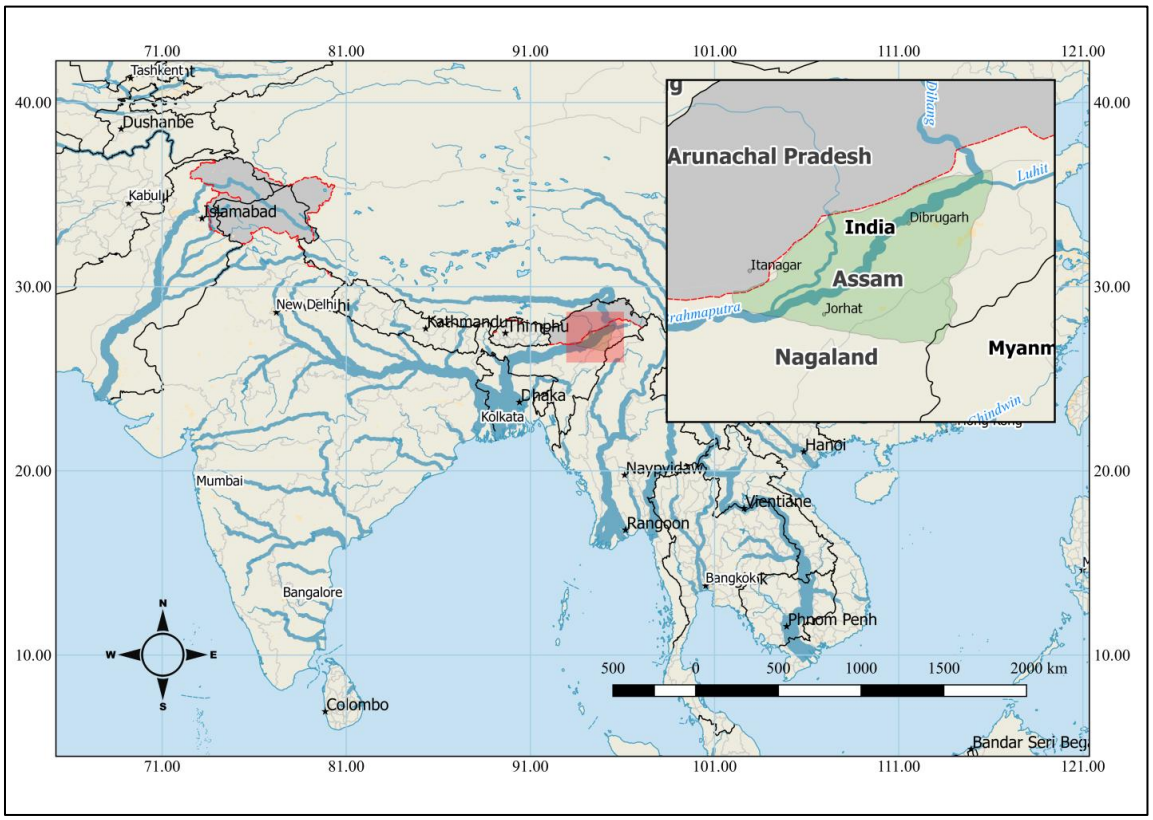

Figure 5.1: LSM Study Area

The study area endures monsoon climate with high humidity levels. The rainfalls are concentrated in the months of May to September, culminating the average rainfall levels to 70-100 inches. Also, it experiences temperatures ranging between 70° F and 104° F. In the cold winter months the temperatures do not generally drop below 39° F. Subsequently the summers send the mercury soaring to the levels of 61°F to 88°F. The summers are short, lasting only a few months and are promptly replaced by winters that make for an early arrival. The maximum average temperature experienced in these regions is 75°F. The majority of the area is covered by the semi-evergreen tropical forests that are typical of the climate in these areas.

5.1.2 Methodology

The summarized methodology for the case study is as shown in Figure 5.2.

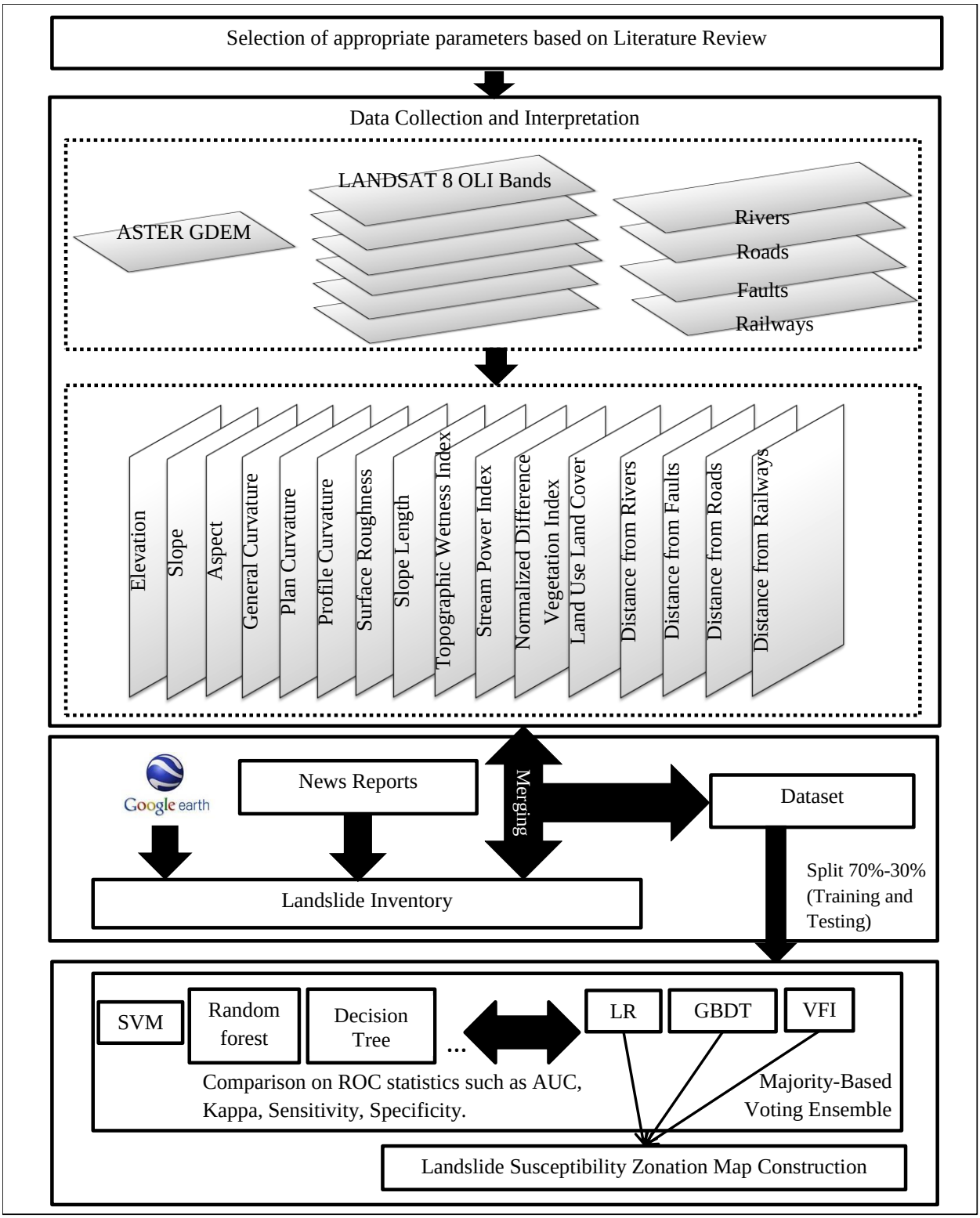

Figure 5.2: LSM Methodology

The first stage of the study comprised of a thorough literature survey to identify the parameters which have been considered as the landslide causative factors. As per the findings the data were collected and brought in an appropriate spatial format to be utilized. Here the thematic maps were prepared corresponding to the necessary landslide conditioning factors and the landslide inventory. Thus, a total of 16 thematic maps representing the elevation, slope, aspect, curvature, plan curvature, profile curvature, surface roughness, slope length, TWI, SPI, NDVI, land use, distance from roads, rivers, faults and railways were prepared. The landslide inventory accounting for the past landslide occurrences, was taken into consideration, and was validated by newspaper reports and historical articles. Next, the conditioning factor thematic maps were annexed over the inventory to obtain a dataset comprising of the location coordinates of the landslide points and the location's corresponding values for all the causative factors. In the third stage, the dataset thus generated, was randomly split into 70%-30% ratios as training and testing sets, respectively. The training set, was then employed for the purpose of training the proposed Majority-based Voting Ensemble LR-GBDT-VFI along with the other models like SVM, DT, etc. The remaining 30%, or the testing dataset was used to obtain predictions from the generated models. Next, the performance of the built models was compared on the basis of the predictions obtained in the previous stage against the original values in the testing dataset. This was accomplished by employing prediction evaluation metrics, like those of accuracy, sensitivity, specificity, etc. Finally, in the fifth stage the validated proposed model, after having proved its mettle was used to build the LSZ map for the region.

5.1.3 Spatial Data Generation and Handling

The study revolves around spatial data and hence the following sections emphasize on the two important facets namely the inventory and the conditioning factors that together comprise the dataset for analysis.

5.1.3.1 Landslide Inventory

A total of 436 landslide locations were identified by using the "zoom in" and "zoom out" features and "time slider" tool of Google Earth Pro for getting access to the historical imageries of the region and to take the past occurrences of landslides into account as

well. The visual interpretation focused on identifying locations bearing traits of massive erosion and of vegetation cover removal. Additionally, newspaper articles and historical reports were leveraged to validate the already identified landslide locations and for identifying new ones.

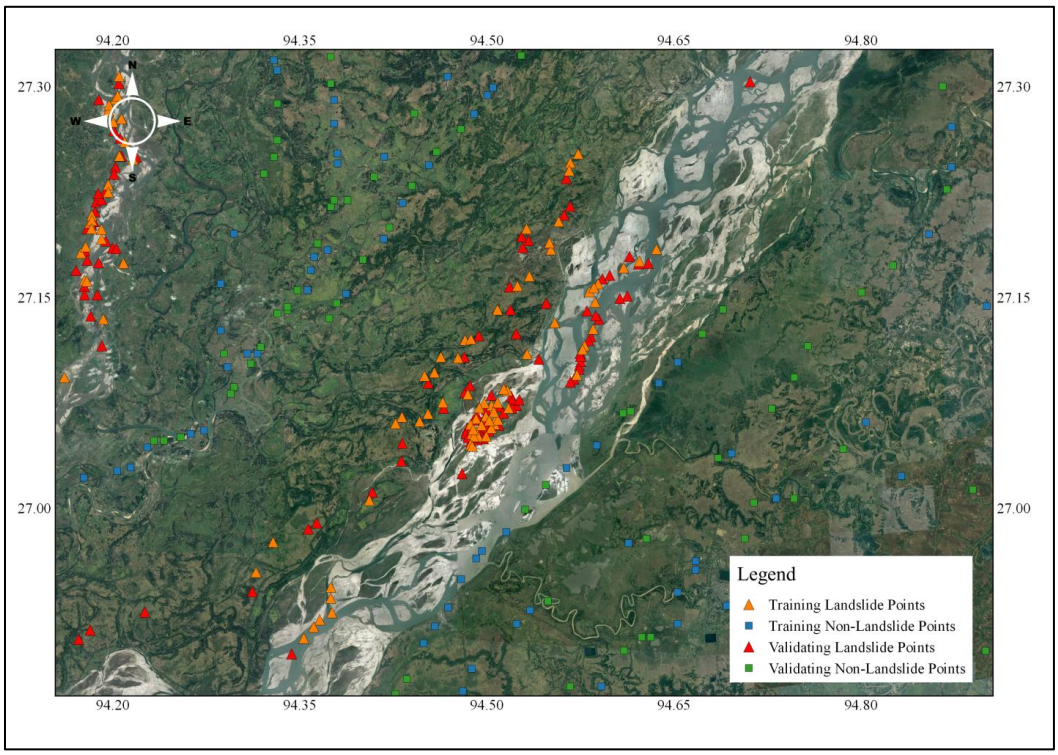

Figure 5.3: LSM Inventory

Fundamentally, LSM is a classic case of binary classification wherein the past occurrence of a landslide is signified by a variable bearing a value “1” and “0”, otherwise. So it is imperative to have adequate representation of non-landslide locations (locations that have not experienced landslides) in the dataset. In order to avoid any bias these locations were randomly picked in the same numbers as landslide locations. The landslide inventory was then randomly split at 70%-30% ratio to generate a training dataset of 327 landslide locations and 109 locations for testing dataset. A sample of training and testing landslide and non-landslide locations used in the study is as shown in Figure 5.3.

5.1.3.2 Landslide Conditioning Factors

Sixteen landslide conditioning factors, including elevation, slope, aspect, general curvature, plan curvature, profile curvature, surface roughness, TWI, SPI, slope length,

NDVI, land use, distance from roads, rivers, faults and railways were taken into consideration for this study. Distance from the railway is a factor which has never been covered in the literature and was found to play a vital role in affecting a region's landslide susceptibility. Thematic maps for the aforementioned factors were obtained using QGIS and SAGA GIS resources. A summarized description of the thematic maps generated, is mentioned in Table 5.1. All the thematic maps were prepared using QGIS 2.18.

Table 5.1: LSM: Description of Landslide Causative factors and their Thematic Maps.

S. No.	**Landslide Causative Factor**	**Source**	**Figure**
1.	Elevation	ASTER GDEM	5.4(a)
2.	Slope	DEM	5.4(b)
3.	Aspect	DEM	5.4(c)
4.	General Curvature	DEM	5.4(d)
5.	Plan Curvature	DEM	5.4(e)
6.	Profile Curvature	DEM	5.4(f)
7.	Surface Roughness	DEM	5.4(g)
8.	Slope length	DEM	5.4(h)
9.	TWI	DEM	5.4(i)
10.	SPI	DEM	5.4(j)
11.	NDVI	Landsat 8 OLI	5.4(k)
12.	Land use	Landsat 8 OLI	5.4(l)
13.	Distance from Faults	Topographic Map	5.4(m)
14.	Distance from Roads	Topographic Map	5.4(n)
15.	Distance from Rivers	Topographic Map	5.4(o)
16.	Distance from Railways	Topographic Map	5.4(p)

A new factor of Distance from Railways has also been taken into account in this study and was found to hold significance in predicting the target region's landslide susceptibility as can be proved from the weights assigned to the parameters on the basis of GINI Index, as it obtained a weight as high as 0.097 as shown in Table 5.2.

Table 5.2: LSM: Relevance of Landslide Conditioning Factors as per GINI Index

No.	**Causative Factor**	**Weight by GINI**	**No.**	**Causative Factor**	**Weight by GINI**
1.	ASPECT	0.001	9.	SLOPE	0.011
2.	SPI	0.001	10.	GENERAL CURVATURE	0.013
3.	TWI	0.003	11.	DISTANCE FROM ROADS	0.027
4.	PLAN CURVATURE	0.003	12.	ELEVATION	0.045
5.	SLOPE LENGTH	0.005	13.	DISTANCE FROM RAILWAYS	0.097
6.	PROFILE CURVATURE	0.006	14.	DISTANCE FROM FAULTS	0.108
7.	DISTANCE FROM RIVERS	0.008	15.	LULC	0.221
8.	SURFACE ROUGHNESS	0.010	16.	NDVI	0.345

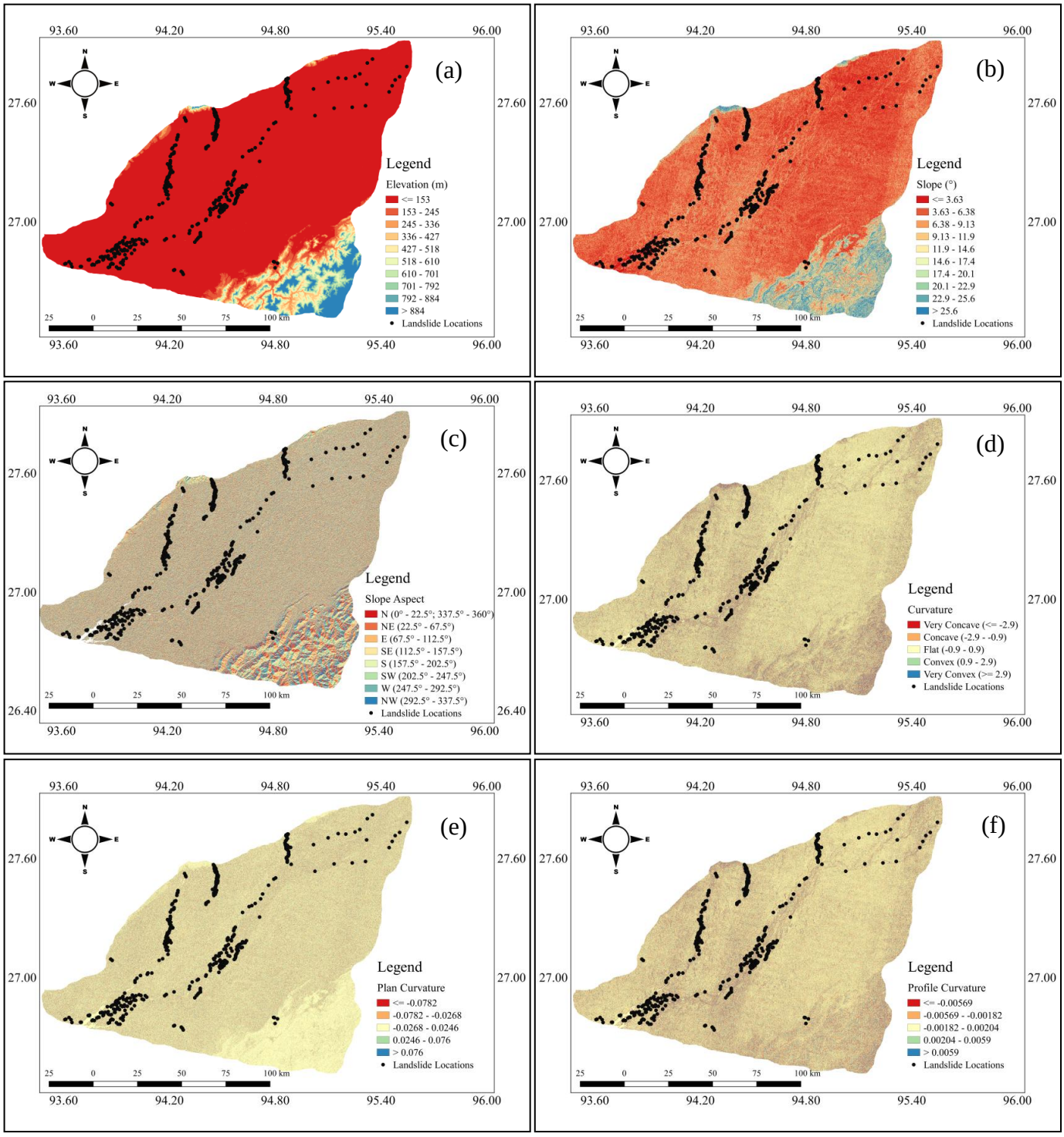

Figure 5.4: LSM Conditioning Factors (a)-(f)

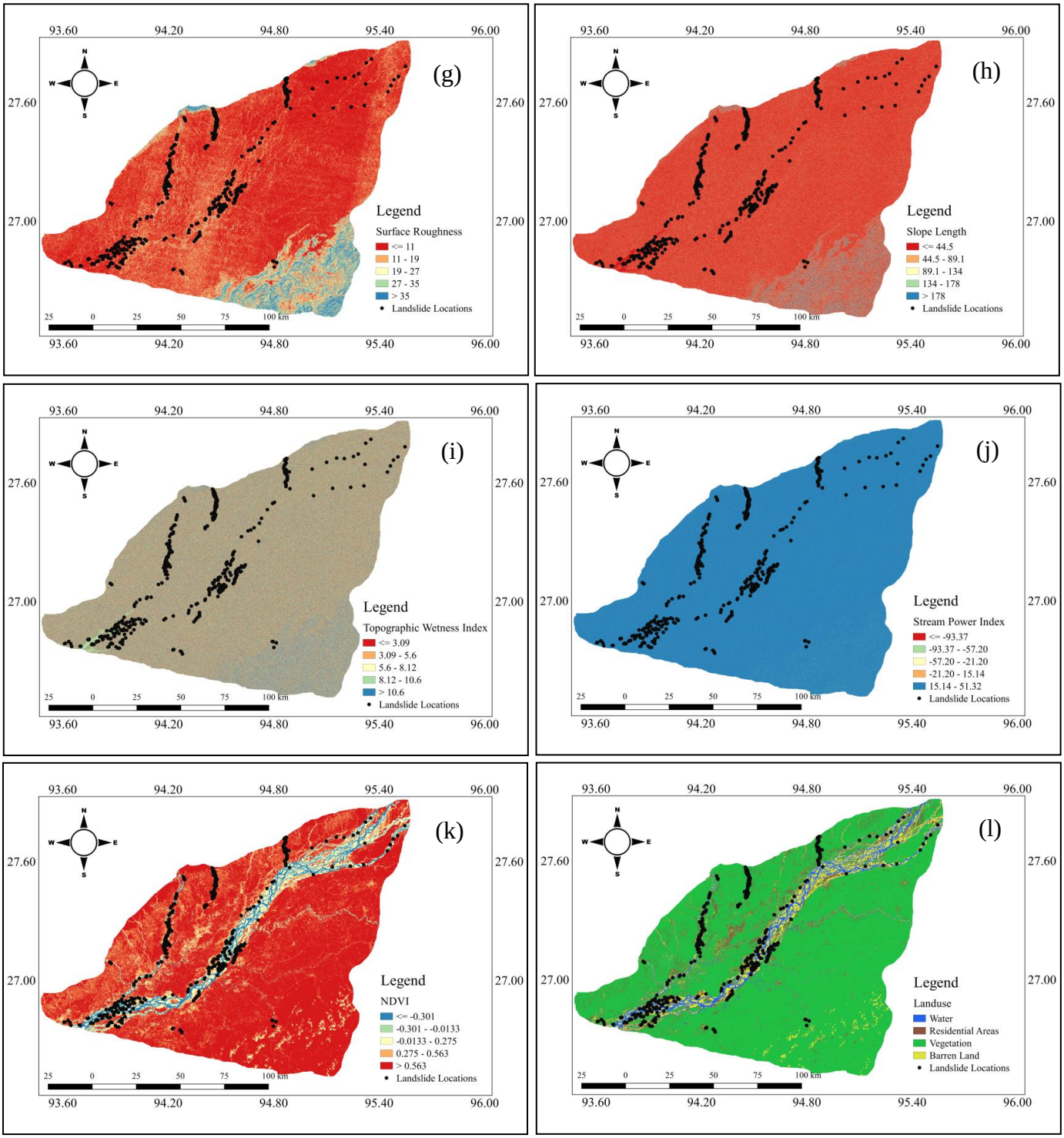

Figure 5.4 (cont.): LSM Conditioning Factors (g)-(l)

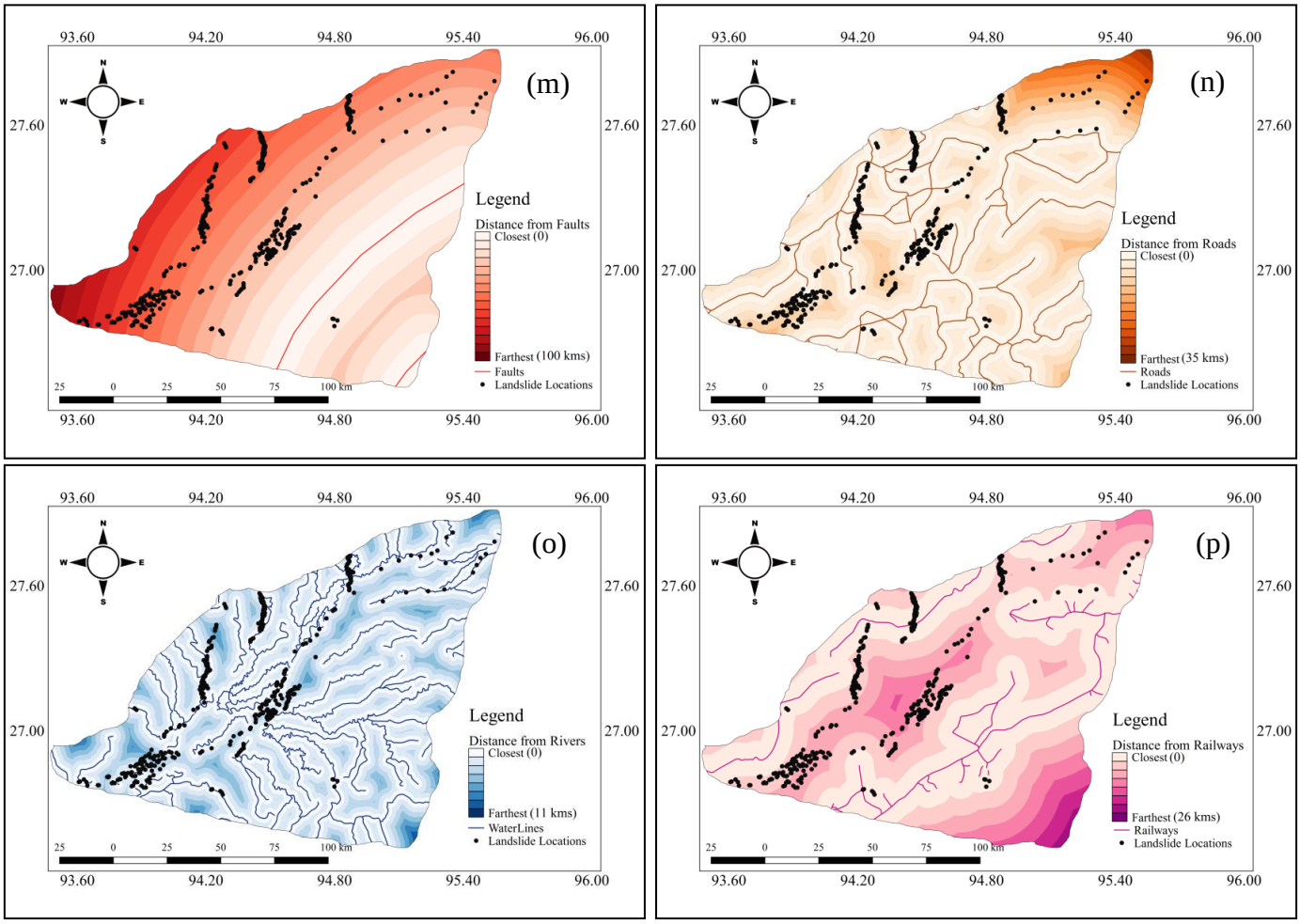

Figure 5.4 (cont.): LSM Conditioning Factors (m)-(p)

5.1.4 Proposed Approach

In the proposed model, LR, GBDT, and VFI have been combined using a majority-based -voting ensemble for landslide susceptibility analysis. All the constituent predictors of the proposed model, including the technique of the ensemble have been briefly described next.

LR: It is one of the conventionally used classifiers that universally have been acknowledged for its prowess in the field of landslide susceptibility analysis. It is one amongst the classifiers which have been borrowed from the field of statistics. This approach predicts the value of the dependent variable using a logistic function defined in Equation (5.1) which takes independent variables as arguments [20]. In this case, the dependent variable defines the presence or absence of landslides while the independent variables are the landslide conditioning/causative factors.

$$Logit(p_i) = \log\left[\frac{p_i}{1-p_i}\right] \tag{5.1}$$

Here $Logit$ is the natural logarithm of the odds of event i which are equivalent to the ratio of probability of event i occurring, i.e. p_i, to the probability of the event i not occurring, i.e. $1 - p_i$. Equation (5.2) demonstrates the linear logistic classifier [20].

$$Logit(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{5.2}$$

Here $\beta_0$is the intercept of the model and $X_1, X_2, \ldots, X_n$are the input landslide predictors with $\beta_1, \beta_2, \ldots., \beta_n$ being their coefficients signifying their weightage in prediction and n is the number of factors considered (16 here). Comparing Equations (5.1) and (5.2) and taking antilog both sides, Equation (5.3) is obtained.

$$p_i = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}} \tag{5.3}$$

The Maximum Likelihood Estimation method is adopted to find an optimal fit for the logistic function and hence the coefficients are obtained. Therefore, the model predicts the output dependent variable using the input independent variables and their corresponding coefficients. One of the advantages associated with this classifier is that it has no pre-requisites from predictors in terms of their type (nominal, ratio scale, ordinal, and interval) or their distributions [54].

GBDT: GBDT is an effective method for classification tasks. It combines classification and regression trees (CART) with gradient boosting [54]. CART analyses all the observations of our landslide dataset and horizontally divides it into groups, each containing all the independent predictors and the dependent binary variable indicating presence/absence of landslides. Each of the groups is then binary spilt into homogenous partitions on the basis of its decision rules. This could be represented as the basis function shown in Equation (5.4) [55].

$$h(x) = \sum_{j=1}^{J} \gamma_j \, I\left(x \in R_j\right), where \; I = 1, if \; x \in R_j; I = 0 \; otherwise \tag{5.4}$$

Here x are the predictors, $R_j's$ represent the disjoint groups and γ_j represent the constant value predicted by the tree for each group [55]. In order to overcome the problem of over fitting CART adopts the technique of pruning to eliminate any unnecessary splits,

obtaining an optimal fitting tree, best describing the data. In contrast, gradient boosting iteratively generates trees of a fixed size, rather than generating a complex tree requiring simplification or pruning. In here, the second tree would be an improvement over the first by eliminating its errors and third being an enhancement over the first and second, so on and so forth. The approach is represented as an approximation function as in the Equations (5.4) and (5.5).

$$g(x) = \sum_{m=1}^{M} g_m(x) = \sum_{m=1}^{M} \beta_m h(x; a_m) \; with \; h(x; a_m) = \sum_{j=1}^{J} \gamma_{jm} \, I\left(x \in R_{jm}\right) \qquad (5.5)$$

The parameters β_m are indicative of the weights allotted to each tree's node and the way predictions are derived from each of the trees whereas a_m are the mean value of splits and leaf nodes for all independent variables. These are estimated by minimization of the loss function. This complex minimization is simplified by gradient boosting which provides an approximation, by means of the application of steepest descent to forward stage estimation. The GBDT has three main optimizing parameters to avoid under-fitting or over-fitting i.e. Number of trees, Learning rate (weightage given to each tree), and Complexity of tree (Splits and Depth of tree).

VFI: VFI is a non-incremental classification algorithm implying a single analysis through the records of the dataset. The primary approach for prediction relies on the individual classification by all the independent variables. These individualistic classifications are then amalgamated together through a voting procedure among them for obtaining the final classification. Primarily, every observation of the dataset is treated as a vector of features (independent variables) plus a label (dependent variable/ binary class value). Using this, the algorithm constructs VFI for all the features. An interval refers to the set of values for a feature for which the class variable exhibits the same value so that no two adjacent intervals have the same value for a class. An interval for a linear feature corresponding to a class value comprises of a range represented by lower and upper bounds while for a nominal feature it comprises of all observed values leading to the particular class value for the interval [56]. However, only the lower bound for ranges is stored as the ending point for one range becomes the beginning of next. In the case of landslide susceptibility analysis the conditioning

factors are first divided into ranges of values or subset of values depending on the type of factor: linear (TWI, SPI) or nominal (Aspect, LULC). Each interval for each of the features would then be represented by $\langle lower\_bound,\ count_{Landslides}, count_{Non-Landslides}\rangle$, where $count_{Landslides}$ represents the number of observations having value for this feature in this interval and are labeled as "Landslides" indicating presence of landslides and $count_{Non-Landslides}$ represents the number of observations having value for this feature in this interval and are labeled as "Non-Landslides" indicating absence of landslides. The creation of all intervals for all features ends the training phase of the algorithm. In the prediction phase, the instance I is analyzed. The votes would be assigned to the probable two classes of $L(Landslide)$or $(NL)Non-Landslide$ depending on the interval of the value for all the features of the instance I. The vote assigned to the $L(landslide)$class on the basis of the interval h of value I_f of feature f, for instance I, is given as in Equation (5.6).

$$FeatureVote(f, L) = \frac{LandslideCountForInterval(f, h)}{LandslideCount} \tag{5.6}$$

Similarly the vote for $NL(Non-Landslide)$class is given as in Equation (5.7)

$$FeatureVote(f, NL) = \frac{Non-LandslideCountForInterval(f, h)}{Non-LandslideCount} \tag{5.7}$$

In case the value I_f falls between adjacent intervals h and $h+1$ then a mean of the counts of the two intervals is considered. The above 2 vote metrics are calculated for all the features and a final vote for both classes$(L\ or\ NL)$ is calculated as described in Equations (5.8) and (5.9).

$$Vote(I, L) = \sum_{i=1}^{N} FeatureVote(I, f_i, L) \tag{5.8}$$

$$Vote(I, NL) = \sum_{i=1}^{N} FeatureVote(I, f_i, NL) \tag{5.9}$$

Here N is the number or causative factors. Depending on the larger of the values the instance is predicted into the corresponding class.

Majority Voting Ensemble: This ensemble technique is also known as plurality vote method [57]. In this technique, the final predicted class of the target variable is the one that would have been estimated by the majority of classifiers. Hence the class which obtained the highest votes (by all classifiers) amongst all the classes is the final prediction of the ensemble model. Mathematically, the proposed approach is expressed in Equation (5.10).

$$class(I) = \max_{c \in \{L, NL\}} (LR(I, c) + GBDT(I, c) + VFI(I, c)) \quad (5.10)$$

Here I is the unlabeled instance whose class needs to be predicted. If LR predicted the presence of landslide for the given instance then $LR(I, L) = 1$ otherwise, $LR(I, NL) = 1$ indicating the absence of landslide. Similarly, the above equation applies for other 2 predictors of GBDT and VFI. In simple terms, the predicted class for unlabeled instance would be "$Landslide\ (L)$" if 2 of the 3 classifiers (LR, GBDT and VFI) predicted "$Landslide\ (L)$", and "$Non - Landslide\ (NL)$" otherwise.

5.1.5 Results and Analysis

All the models mentioned in Table 5.3 were trained on the training set and their performance was then compared by, feeding the testing set as input into the trained models and evaluating the value of their predictive statistics.

Table 5.3: LSM: Model Comparison of LR-GBDT-VFI against benchmark models

S. No	Model	Accuracy	AUC	Kappa	Sensitivity	Specificity
1.	Majority Based Voting Ensemble LR-GBDT-VFI	96.66	.984	.932	96.79	96.48
2.	LR	92.71	.97	.849	94.33	90.37
3.	GBDT	95.44	.981	.905	95.94	94.7
4.	VFI	65.65	.917	.366	44.90	96.24
5.	DT	91.79	.901	.833	94.57	88.28
6.	SVM	91.19	.97	.796	89.81	91.06
7.	ANN	90.27	.968	.796	89.81	91.06
8.	NB	68.39	.955	.418	49.25	98.44
9.	RF	91.49	.94	.824	92.82	89.55
10.	PSO-SVM	92.14	.95	.782	87.72	89.11
11.	EO-GBDT	95.54	.901	.933	95.57	94.78

For formal definitions of the metrics employed refer Section 2.4. As can be observed from Table 5.3, our proposed Majority Voting ensemble (LR-GBDT-VFI) model performs well for the prediction of both Landslides and Non-Landslides locations

attaining values of 96.79 and 96.48 for sensitivity and specificity, respectively, which is a significant improvement from Logistics Regression (94.33, 90.37), GBDT (95.94, 94.7), VFI (44.9, 96.24) and various other classifiers.

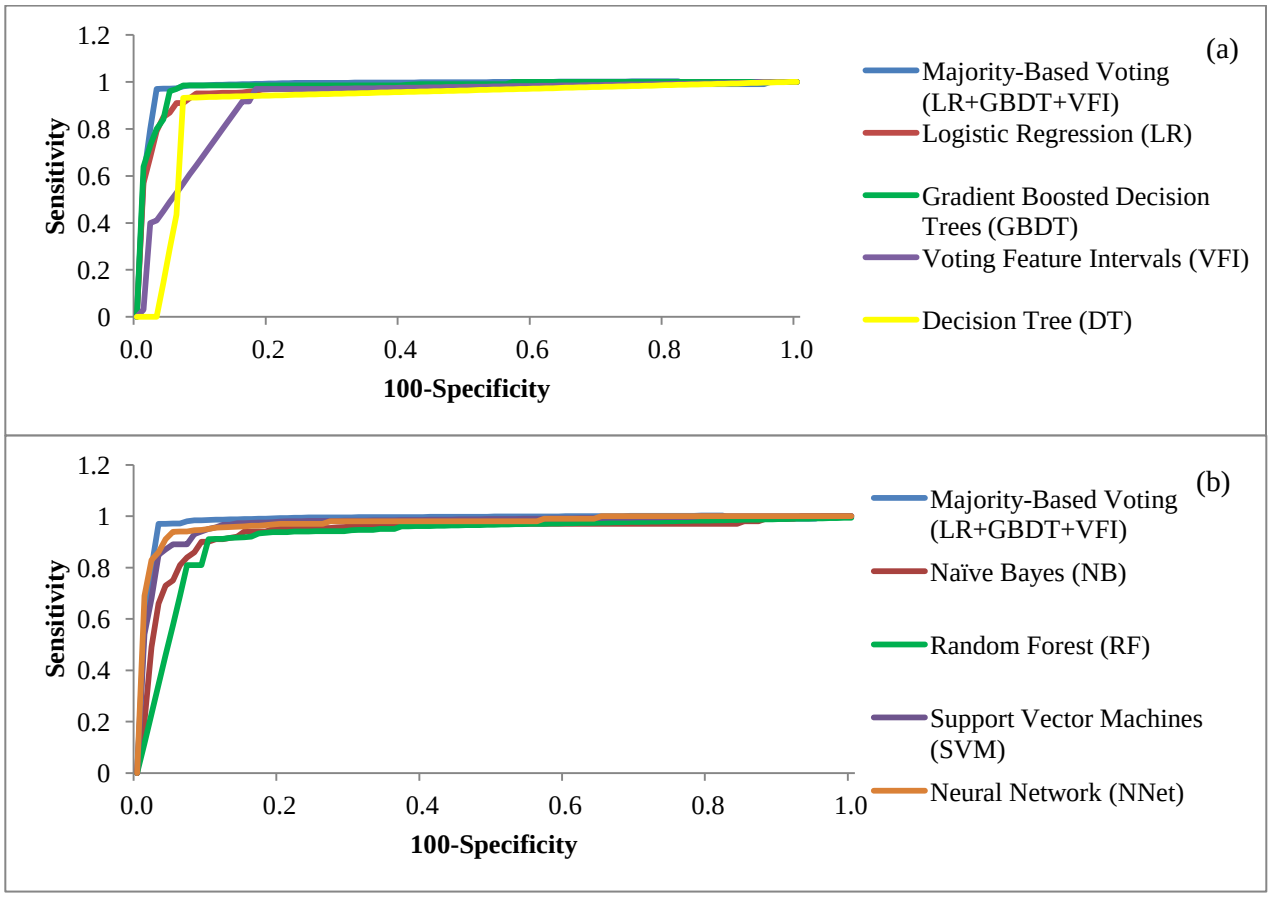

Figure 5.5: LSM ROC curve comparison

Also the ROCs have also been plotted as shown in Figure 5.5 to get a better perspective of the performances displayed by the models. Pair of $sensitivity$ and $100 - specificity$ values have been utilized to generate the ROC curves. AUC values have been deduced from these curves and depicted in Table 5.3. It can be inferred from the Figure 5.5 that the Majority-based Voting Ensemble LR-GBDT-VFI (AUC: 0.984) performs substantially better than the rest of the models and has performance eminently better than each of its individual constituent models. As is evident from Figure 5.5 and Table 5.3, the proposed model (AUC: 0.984) outperformed models such as DT (AUC: 0.90), SVM (AUC: 0.97), ANN (AUC: 0.968) and RF (AUC: 0.94) etc. An important milestone in the

landslide susceptibility analysis is the generation of the LSZ map of the study area ascertaining the future chances of being struck by a landslide as shown in Figure 5.6.

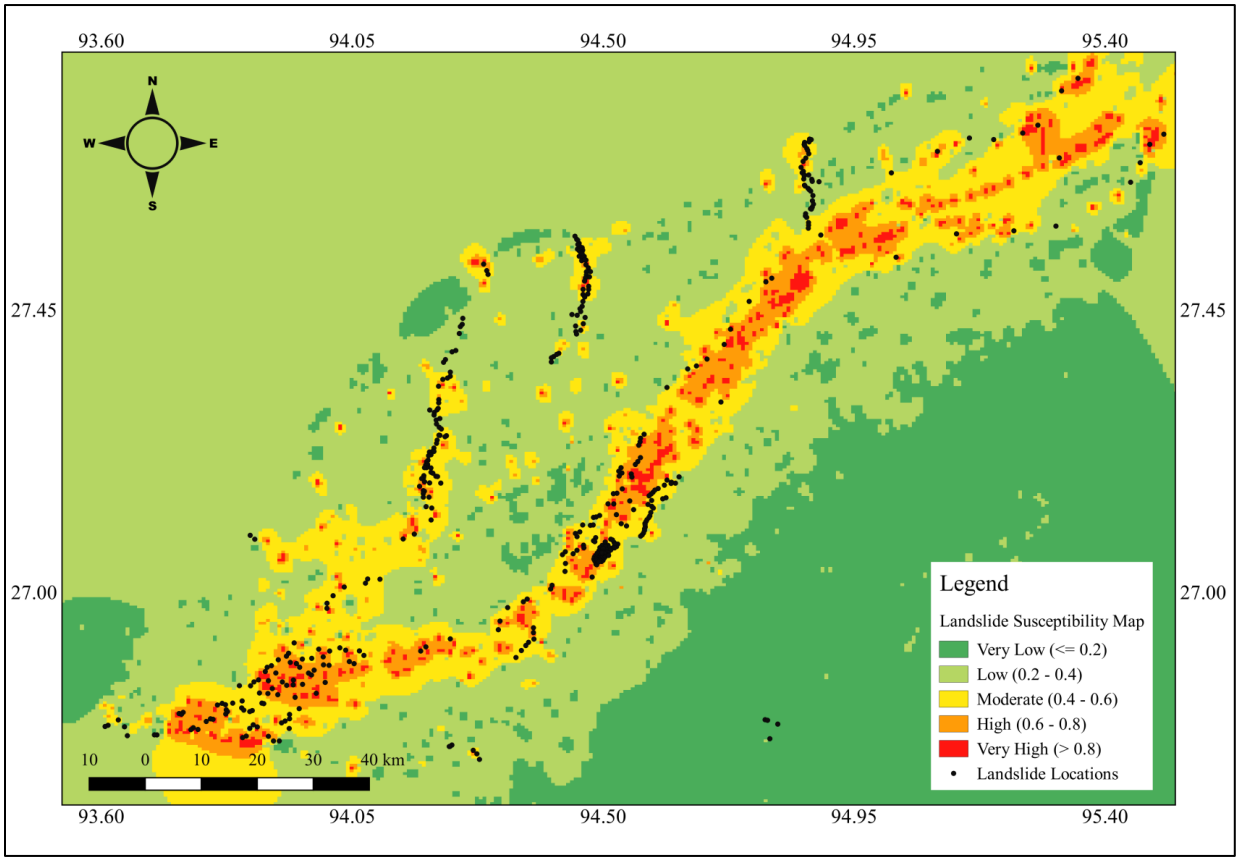

Figure 5.6: LSZ Map

After successfully training, validating and comparing the proposed prediction model, the trained model is then used to generate this map that would provide a visual clarity for the region's susceptibility by distinguishing areas that are highly prone to those which are relatively un-susceptible. In this study, this task was accomplished by randomly selecting 6000 locations spread wide over the study region and obtaining the values for all the causative factors at these locations. This constituted our new unlabeled dataset. The values of the susceptibility indexes were then estimated by using our trained model. The map was then constructed in QGIS by an Inverse Distance Weighted (IDW) Interpolation on these indexes. The map was then classified on these indexes into intervals ranging in susceptibility as Very Low (0-0.2), Low (0.2-0.4), Moderate (0.4-0.6), High (0.6-0.8) and Very High (0.8-1). The map indicated 6.8% (1360 km^2) of the area falls under very high landslide susceptibility while 7.2% (1440 km^2) under high landslide susceptibility zone.

Also, 22.7% (4540 km^2), 38.7% (7740 km^2) and 24.6% (4920 km^2) of the study area fell under the categories of moderate, low and very low susceptibility respectively.

5.1.6 Summary

Landslides are among the most impactful hazards to have made their presence felt predominately in the North-East of India. Ensemble is an area of machine learning which remains largely unexplored in landslide hazard assessment. The Majority-based voting ensemble of classifiers which gives the final prediction as one which has been predicted by the majority of models has proved to perform well in a diversity of fields. So this paradigm of the ensemble has been utilized in this study to combine the predictions from LR, GBDT, and VFI classifiers. The ensemble model proved its prowess over the conventional standalone classifiers, including its own constituent models achieving an AUC of 0.984. Other important statistical measures of Kappa (0.932), Sensitivity (96.79) and Specificity (96.48) also outperformed those of conventional machine learning models such as DT, SVM, ANN and RF. Also, the model achieved an accuracy of 96.66%. The validated model was then put to use to generate a zonation map that could pose as a blueprint for future development and infrastructure projects in this region. Also, the map indicated that approximately 6.8% of study area falls under very high susceptibility. This could aid in government's initiatives of disaster handling and management.

5.2 Case Study II: FFSM (NDBR)

Rampant pasture burning has lead to various forest fires taking their toll over the health of many forests. NDBR, located in the northern part of India witnessed a majority of these incidents in the recent past, though it remains comprehensively untouched from research studies. The scale of these wildfires has led to an immense requirement for preventive measures to be taken for recuperating from such events. This requires an in-depth analysis of the study area, its history of wildfires and their causes. These efforts would assist in laying a blueprint for a contingency plan in the event of a wildfire. This case study explores an Evolutionary Optimized GBDT for preparing wildfire susceptibility map for the study area that would aid in the government's forest preservation and disaster management activities. The study took 18 ignition factors of elevation, slope, aspect, plan curvature, TPI, TWI, NDVI, soil texture, temperature,

rainfall, aridity index, potential evapotranspiration, relative humidity, wind speed, land cover and distance from roads, rivers and habitations into consideration. The proposed model was compared against various machine learning models such as RF, ANN and SVM and it outperformed them by achieving an overall accuracy of 95.5%. The study revealed that approximately 1,432.025 km^2 of the area was very highly susceptible to forest fires while 1,202.356 km^2 was highly susceptible to forest fires. The proposed model demonstrated good prospects for application in the field of HSMs.

5.2.1 Study Area

The NDBR comprises of the three districts of Chamoli, Bageshwar and Pithoragarh as shown in Figure 5.7. The reserve is located in the Northern part of India and shares a border with China and Nepal.

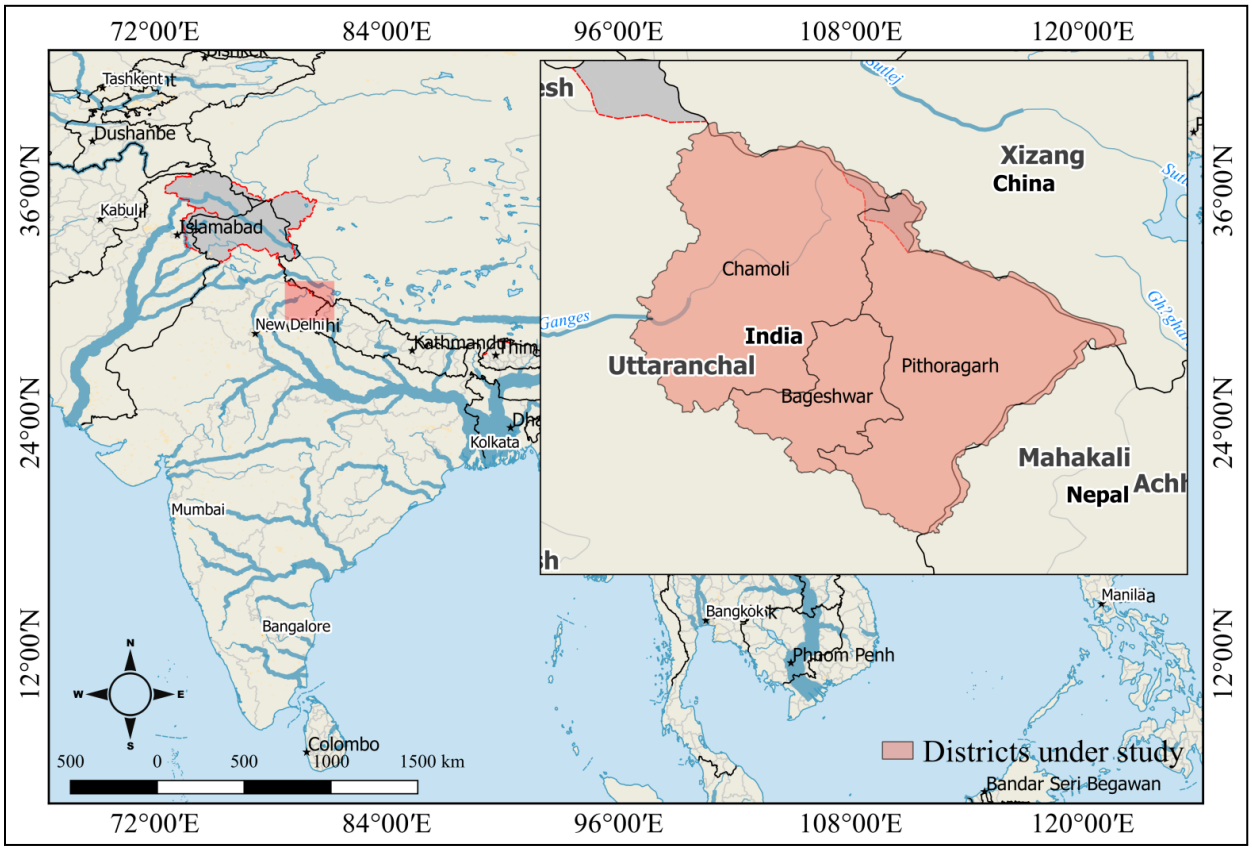

Figure 5.7: FFSM Study Area

It is situated around the peak of Nanda Devi (a part of the Himalayan mountain range) which is the second highest peak in India at an elevation of greater than 3,500 m above the sea level. The study area covering the districts of Chamoli, Bageshwar and

Pithoragarh encompasses about 20,166.329 km^2 of land area located between 29°26'33" N and 31°4'29" N latitudes and 79°4'47" E and 81°2'34" E longitudes. The study area experiences a micro climate with it being dry with low yearly precipitation. It receives heavy rainfall during the monsoon, which lasts from the months of June till September. The upper ranges of the area remain snowbound for a period of about seven months lasting from October to March. The lower altitudes in the south experience deeper snow as compared to valleys on the north. The region encounters temperatures in the range of 10°- 23° C in the months of April to June while temperatures in the range of 7°- 22° C are experienced in the months of July to October.

5.2.2 Methodology

The detailed procedure followed through the study is depicted in Figure 5.8. It comprised of identifying the locations of historic forest fires in the area as Forest Fire Inventory. Along with the coordinates of previous forest fires, 18 ignition factors were considered to obtain the information about the conditions prevailing at the aforementioned locations which led to wildfires. These aforementioned ignition factors included Elevation, Slope, Aspect, Plan Curvature, TPI, TWI, NDVI, Land Cover, Annual Temperature, Annual Rainfall, Relative Humidity, Wind Speed, Aridity Index, Potential Evapotranspiration, Soil Texture and Distance from Roads, Rivers and Habitations. The merger of the ignition factors and the forest fire inventory together clubbed to form the dataset. The values of the ignition factors for an equal number of locations, which did not experience forest fires, were also included in the dataset to avoid bias. Thus the points which experienced wildfires were depicted to as belonging to the class having value ‘1’ while the non-forest fire locations belonged to the class with value ‘0’. The dataset was then split at 70%-30% ratio for training and testing the model respectively. The training set which was 70% of the original dataset was employed to train the GBDT model using a set of initial values of the parameters (Number of trees, Learning Rate and Maximum Depth). The accuracy of the trained model was then computed on the Testing dataset which comprised of remaining 30% of the original dataset. The Evolutionary optimization (EO) was then adopted to obtain a new set of parameters of the GBDT model and the model was then trained again on the training set.

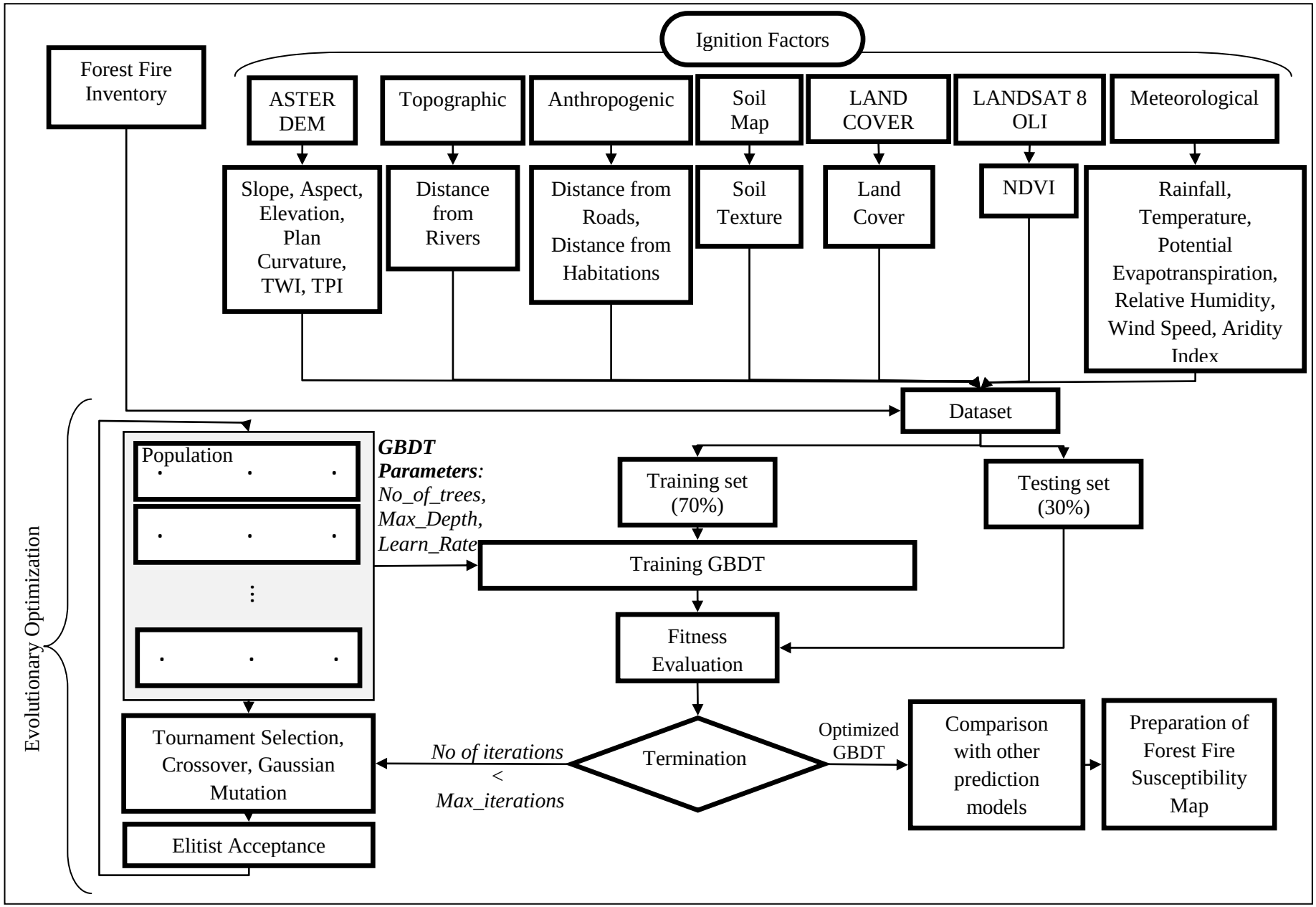

Figure 5.8: FFSM Methodology

The performance of the trained model was again evaluated on the testing dataset. The procedure is repeated till the termination condition is reached, i.e. the number of iterations performed exceed the maximum allowed iterations. The performance of the final optimized model with the optimized set of parameters was then compared against the performance of conventional models of RF, ANN and SVM. The proposed model having outperformed its competitors, was employed to construct the susceptibility map for the region, on the basis of the susceptibilities of the locations indicating the probability of occurrence of a wildfire on a given region's traits.

5.2.3 Spatial Data Generation and Handling

This section deals with the mechanisms employed to generate the data and the generated data's corresponding analysis.

5.2.3.1 Forest Fire Inventory

The foundation of FFSM is based on the relationship between the spatial coordinates of the historical forest fires and the ignition factors at those locations, hence the role of the forest fire inventory is indispensable [27]. The inventory for this study was prepared using the wildfire locations recorded by the Forest Survey of India (available at http://fsi.nic.in/forest-fire.php). A total of 702 forest fire locations were spotted in the districts of Chamoli, Bageshwar and Pithoragarh in the period from January 2004 to December 2016. Out of the 702 fires, 70% (491 forest fires) were exercised to train the proposed model while the remaining 30% (211 forest fires) was utilized for validating its performance against those of other predictors. The forest fires used for training and testing the proposed models and its contemporaries, are as depicted in Figure 5.9. Along with these locations, the locations which did not experience any wildfires were randomly selected for the study area in an equal number (i.e. 702 non-forest fire locations) and were made part of the dataset for training and testing the proposed model.

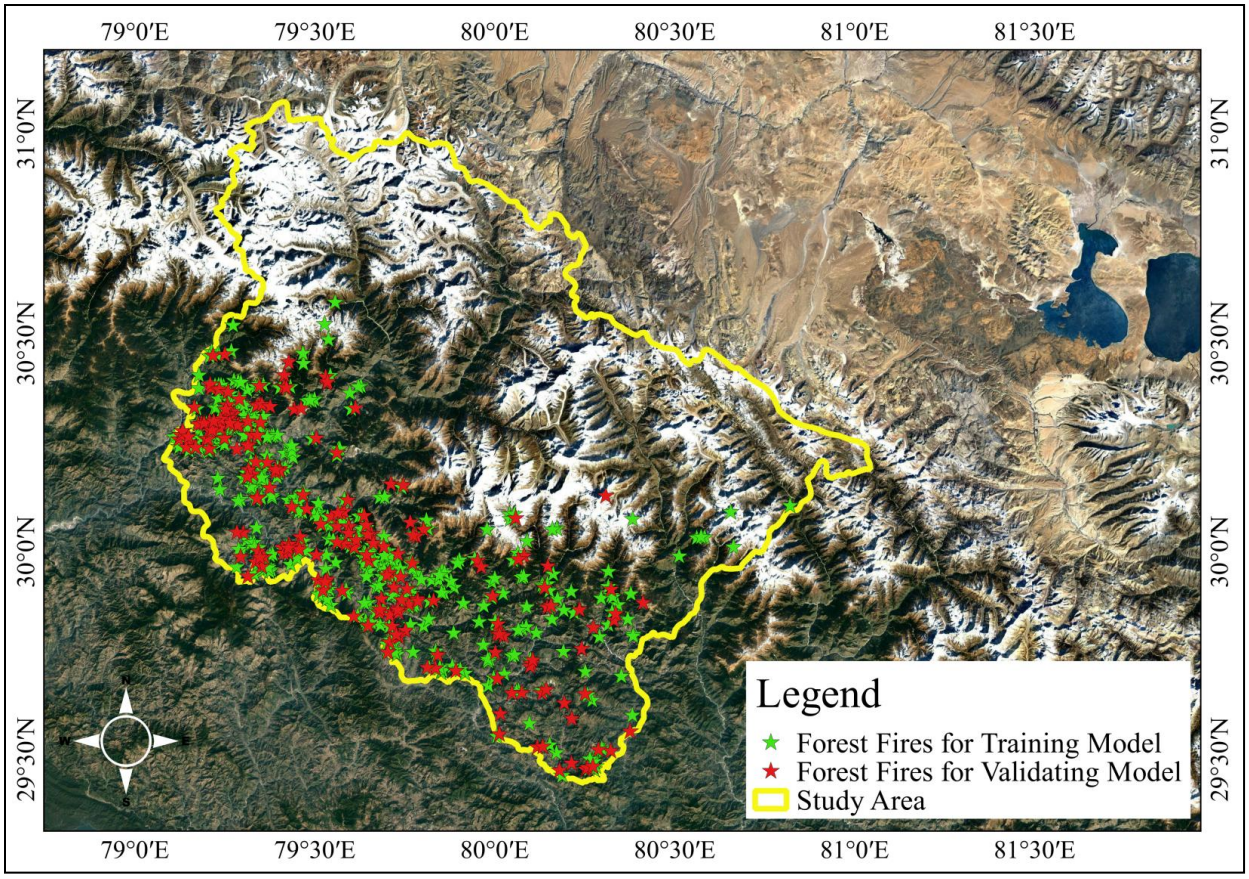

Figure 5.9: FFSM Inventory

5.2.3.2 Ignition Factors

The quality of the prediction depends vastly on the quality of the parameters selected for prediction. As the causes of wildfires are varied the same variety is reflected in the ignition factors being considered for predicting forest fires. Also, the scale of the wildfire does not only depend on its sources, but also the factors that would be responsible for its spread and aggravation. This research included eighteen ignition factors for predicting the wildfires namely Elevation (m), Slope (°), Aspect, Plan Curvature, TPI, TWI, NDVI, Land Cover, Annual Temperature (°C), Annual Rainfall (mm), Relative Humidity, Wind Speed (m/s), Potential Evapotranspiration (mm), Aridity Index, Soil Texture, and Distance from Roads (km), Rivers (km) and Habitations (km). A summary of the maps generated and their respective sources with corresponding scale/resolution is specified in Table 5.4.

Table 5.4: FFSM: Description of Ignition Factors and their Thematic Maps

Data	Source	Scale/ Resolution	Data Format	Unit	Fig.
Elevation	Advanced Space-borne Thermal Emission and Reflection Digital Elevation Model (ASTER DEM)	30 arc seconds at equator	Raster	Meters (m)	5.10(a)
Slope	Advanced Space-borne Thermal Emission and Reflection Digital Elevation Model (ASTER DEM)	30 arc seconds at equator	Raster	Degree (°)	5.10(b)
Aspect	Advanced Space-borne Thermal Emission and Reflection Digital Elevation Model (ASTER DEM)	30 arc seconds at equator	Raster	-	5.10(c)
TPI	Advanced Space-borne Thermal Emission and Reflection Digital Elevation Model (ASTER DEM)	30 arc seconds at equator	Raster	-	5.10(d)
Plan Curvature	Advanced Space-borne Thermal Emission and Reflection Digital Elevation Model (ASTER DEM)	30 arc seconds at equator	Raster	-	5.10(e)
TWI	Advanced Space-borne Thermal Emission and Reflection Digital Elevation Model (ASTER DEM)	30 arc seconds at equator	Raster	Ratio	5.10(f)
Land Cover	USGS Land Cover Institute (USGS LCI)	15 arc seconds at equator	Raster	-	5.10(g)
Soil Texture	Food and Agriculture Organization (FAO)	1:5.000.000	Vector	-	5.10(h)
NDVI	LANDSAT 8	15 arc seconds at equator	Raster	Ratio	5.10(i)
Potential Evapo-transpiration	Consortium for Spatial Information (CGIAR-CSI)	30 arc second at equator	Raster	(mm)	5.10(j)
Aridity Index	Consortium for Spatial Information (CGIAR-CSI)	30 arc second at equator	Raster		5.10(k)
Wind Speed	Global Weather Data	-	Excel	(m/s)	5.10(l)
Relative Humidity	Global Weather Data	-	Excel	Ratio	5.10(m)
Annual Temperature	WorldClim-Global Climate Data	30 arc second at equator	Raster	Degree Celsius (° C)	5.10(n)
Annual Rainfall	WorldClim-Global Climate Data	30 arc second at equator	Raster	(mm)	5.10(o)
Distance from Rivers	DIVA-GIS	-	Vector	(km)	5.10(p)
Distance from Roads	DIVA-GIS	-	Vector	(km)	5.10(q)
Distance from Habitations	India Biodiversity Portal		Vector	(km)	5.10(r)

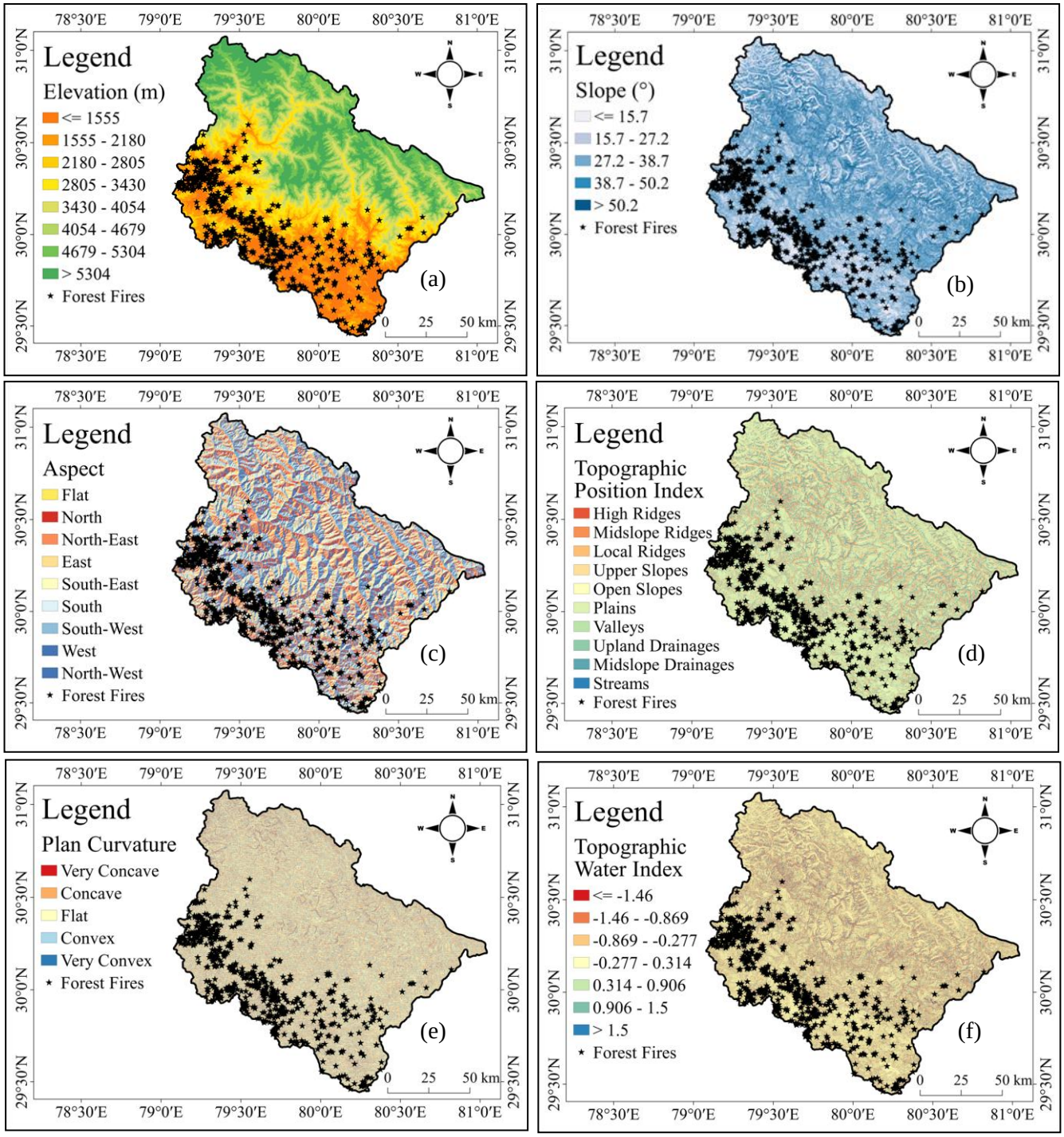

Figure 5.10: FFSM Ignition Factors (a)-(f)

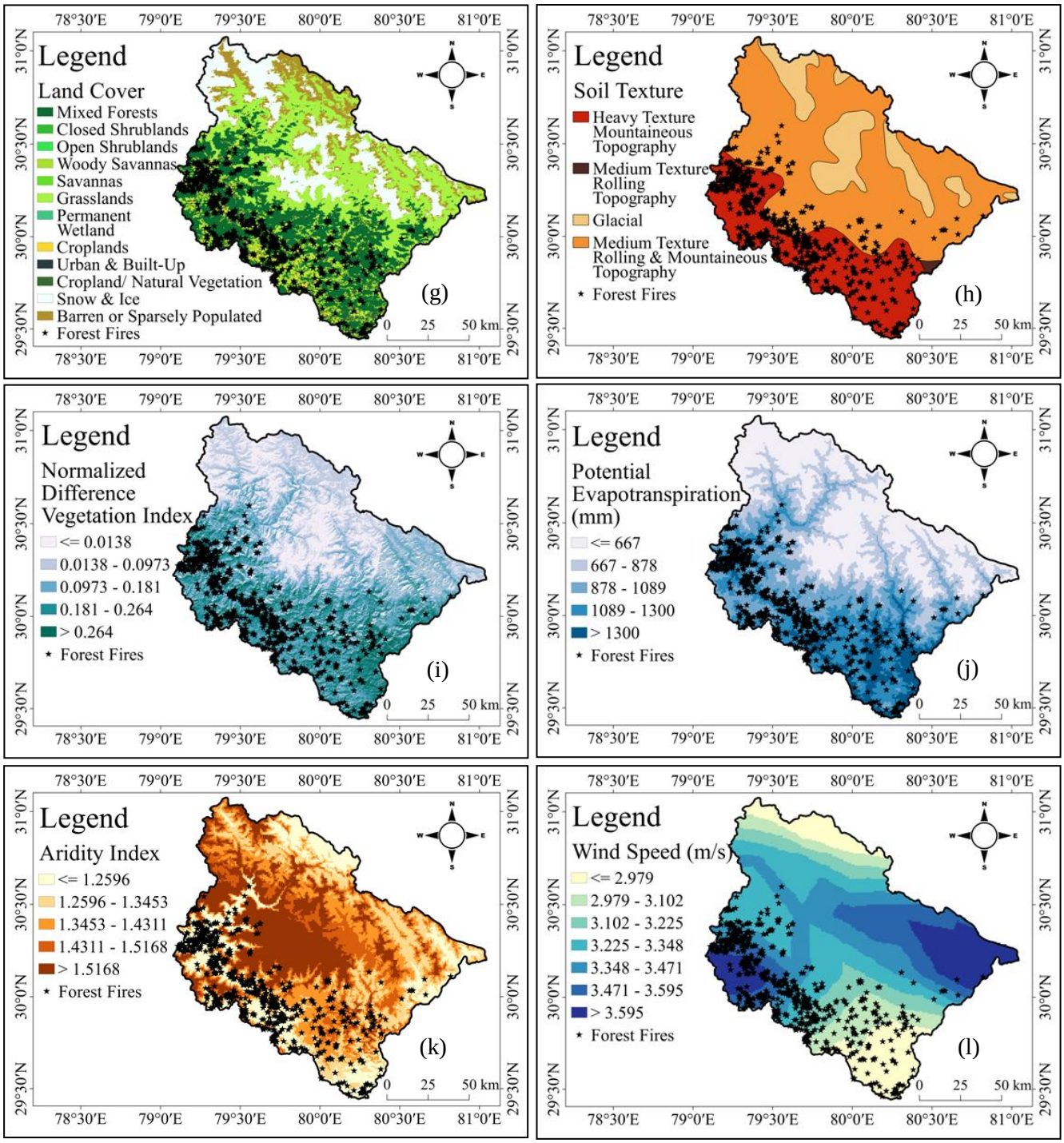

Figure 5.10 (cont.): FFSM Ignition Factors (g)-(l)

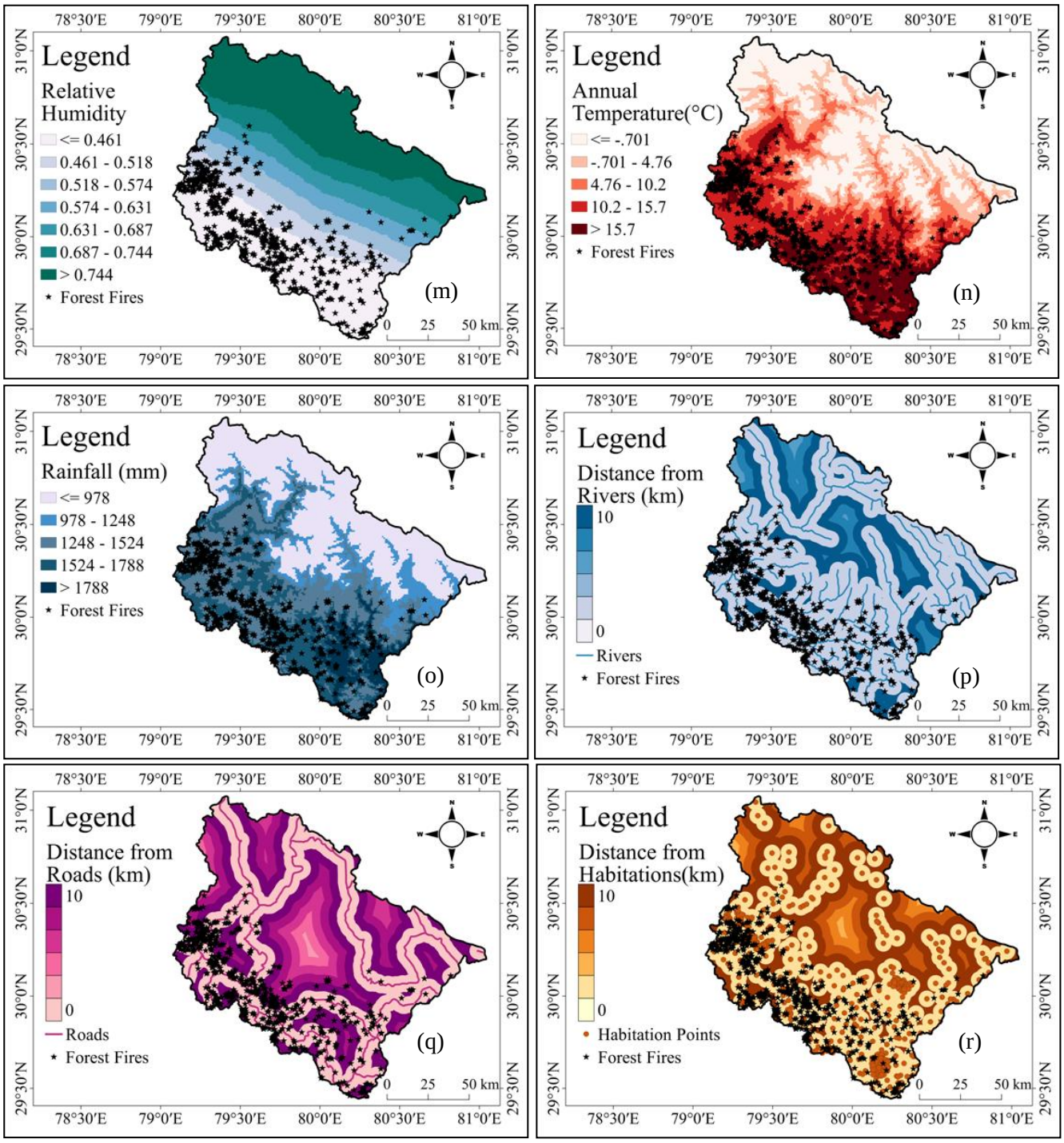

Figure 5.10 (cont.): FFSM Ignition Factors (m)-(r)

5.2.4 Proposed Approach

The proposed model applies an Evolutionary Optimized Gradient Boosted Decision Tree (EO-GBDT) for the spatial prediction of Forest Fires.

GBDT: The formal definition is given in Section 5.1.4. The GBDT has three main optimizing parameters to avoid under-fitting or over-fitting i.e. Number of Trees, Learning rate (weightage given to each Tree), and Complexity of tree (Splits and Depth of Tree) which have been optimized to obtain maximum accuracy of prediction using the EO as described in the next section.

EO: It is referred to as a randomized heuristic search method founded on Darwin's theory of Survival of Fittest. The optimization starts with a population of potential parameter values. Iteratively new potential parameter values are generated (reproduced) and subsequently the good values (those parameters giving high predictive accuracy on testing dataset) are selected to form the new population. By continual selection of good population, and subsequent good offspring the parameter gradually progresses towards best-optimized value giving best possible accuracy [58]. The parameters of GBDT being optimized include the No. of Trees, Learning Rate and the Maximum Depth of Tree.

Table 5.5: FFSM: Optimized Parameters and Accuracy of GBDT

GBDT Parameters	Optimized Value
No. of Trees	9366
Learning Rate	0.189
Maximum Depth of Tree	6
Accuracy	95.5
EO-Iteration	8

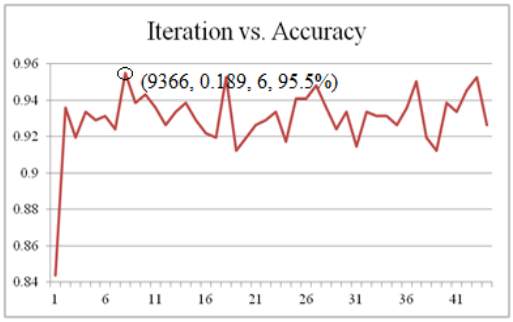

Figure 5.11: FFSM Optimized Accuracy

In the first iteration, the initial population of parameter values is randomly generated. The model is then trained once on the training dataset. The trained model is then used for prediction on the testing dataset. The predictive accuracy is the evaluation function whose value needs to be optimized. The population record giving the best value for the evaluation function is selected as the best fitness known as the elitist selection. Next, a new population is generated using the Tournament selection, Crossover and Gaussian

mutation. The tournament selection refers to the selection of population records (parents) for the next crossover. In here, several tournaments are executed among the few individuals chosen randomly from the population. The tournament fraction specifying the percentage of the current population to be used as participants of the tournament was set at 0.25. The winner is decided on the basis of the output of evaluation function. The subsequent winners undergo crossover operation for generating the next possible candidates. Crossover refers to the process of taking up two of these winner parameter sets and generating a new parameter value set [58]. The process of Mutation follows next wherein a new parameter value set is generated from a single initial population candidate by adding a unit Gaussian distributed random value to each initial parameter value. The population size remains constant through Elitist acceptance where only the best candidates are maintained with the rest being spared. The process iterates for a fixed number of times (45 here). The optimized parameter values are depicted in Table 5.5 and Figure 5.11.

5.2.5 Results and Analysis

The performance of the proposed EO-GBDT model was compared against those of ANN, RF and SVM, etc. The results have been summarized in the Table 5.6. Also the ROC curve for the proposed model on the testing dataset has been depicted in Figure 5.12.

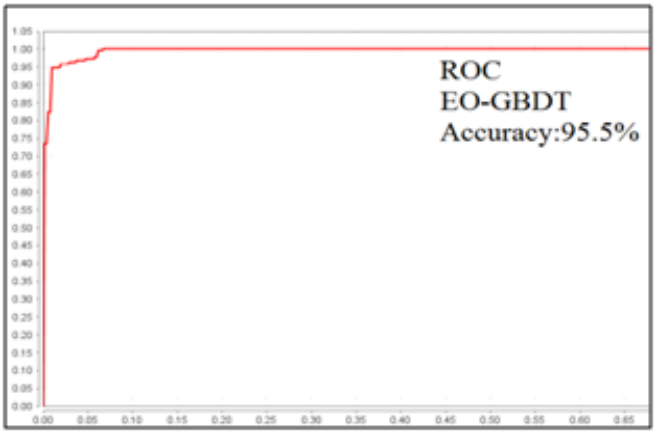

Figure 5.12: FFSM: ROC Curve for EO-GBDT

For a formal definition of statistical performance measures refer to Section 2.4. Observing from Table 5.6 the proposed model outperformed the likes of RF, ANN and SVM and achieved an accuracy of 95.5% , hence was utilized for constructing the

region's susceptibility map as shown in Figure 5.13. This task was accomplished by calculating the susceptibility indexes using the proposed model for 6,000 points in the study area distributed randomly and interpolating the susceptibilities using the Inverse Distance Weighted (IDW) interpolation in QGIS. The areas were then categorized on the basis of the range of susceptibilities as Very Low (susceptibility in the range less than 0.173), Low (between 0.173 and 0.341), Moderate (between 0.341 and 0.51), High (between 0.51 and 0.678) and Very High (greater than 0.678). It revealed that approximately 1,432.025 (7 %) km^2 of the area was very highly susceptible to forest fires while 1,202.356 km^2 (5 %) was highly susceptible to forest fires. Moderately susceptible areas covered about 1,671.133 km^2 (10%) of the region with susceptibility in the range of (0.341-0.51) and areas with low susceptibility comprised 2,317.143 km^2 (11%) with susceptibility in the range of (0.173-0.341). 13,543.673 km^2 (67%) of the region was found to have a very low susceptibility to wildfires (less than 0.173).

Table 5.6: FFSM: Model Comparison of EO-GBDT against benchmark models

Sr. No.	Model/ Metrics	Kappa	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Accuracy
1.	EO-GBDT	0.91	99.05	91.94	92.48	91.94	95.5
2.	Majority Based LR-GBDT-VFI	.886	95.26	93.36	93.49	93.36	94.31
3.	PSO-SVM	0.848	92.42	91.47	91.55	91.47	91.94
4.	LR	0.872	88.63	98.58	98.42	98.58	93.6
5.	GBDT	0.834	98.58	84.83	86.67	84.83	91.71
6.	VFI	0.706	98.1	72.51	78.11	72.51	85.31
7.	DT	0.806	92.89	87.86	88.29	87.86	90.28
8.	NB	0.711	96.68	74.41	79.07	74.41	85.55
9.	ANN	0.839	91.94	91.94	91.94	91.94	91.94
10.	RF	0.815	97.16	84.36	86.13	84.36	90.76
11.	SVM	0.848	92.42	91.47	91.55	91.47	91.94

5.2.6 Summary

India being an agrarian society, has a high dependency on lands and forests. Due to the limited facilities and resources the people directly associated often tend to take the health of these natural resources for granted. The practice of stubble burning is one of those that have taken a toll over these crucial reservoirs. These exercises deemed as harmless are poorly planned and even more so, executed, often escalate into wildfires causing losses of

lives, property and biodiversities. One such location that has become a victim of these unfortunate encounters is the NDBR, Uttarakhand located in the northern part of India. This work explored a novel technique of predicting the region's levels for the Uttarakhand's districts of Chamoli, Bageshwar and Pithoragarh for their susceptibility towards forest fires. The proposed method, EO-GBDT was used for predicting the forest fires on a dataset that composed of the forest fire inventory annexed with the value of ignition factors in the region.

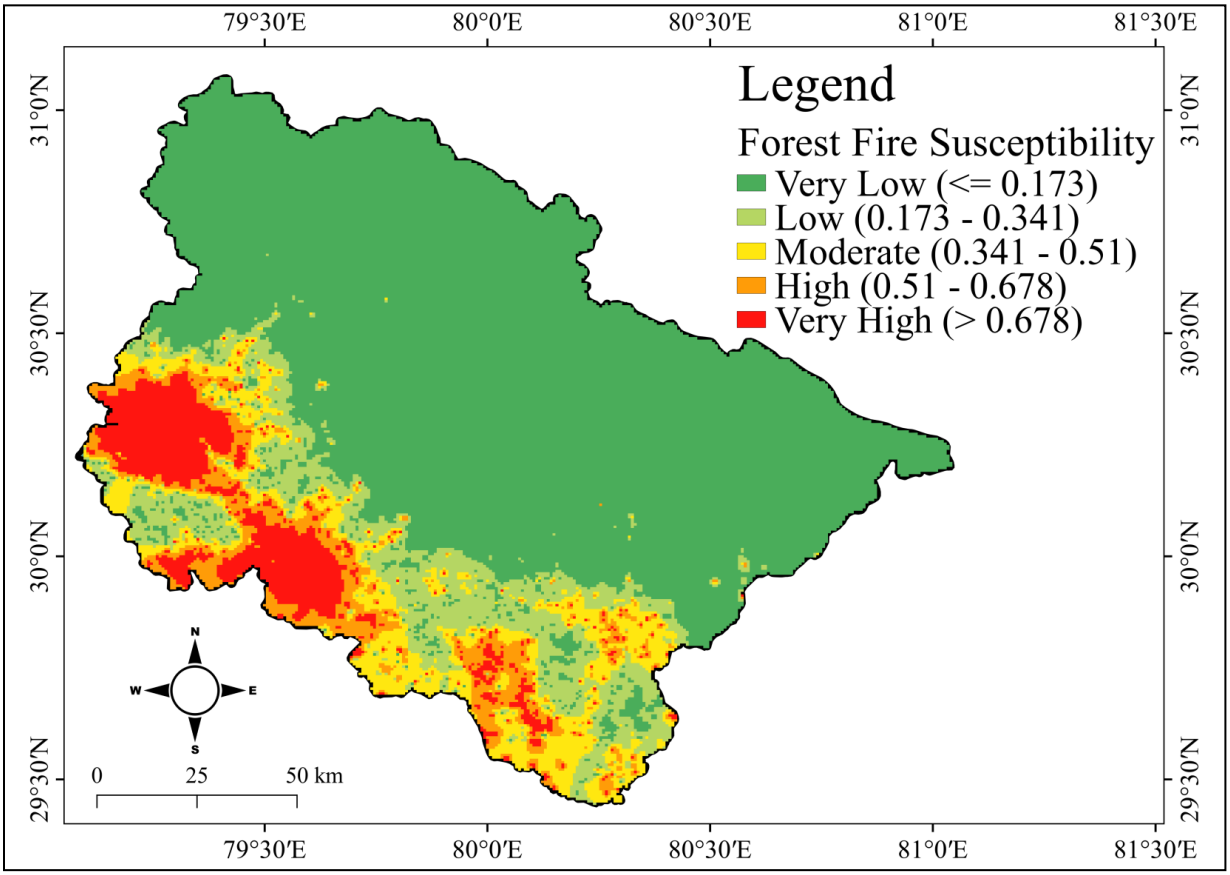

Figure 5.13: FFSZ Map

The proposed model was compared on various statistical measures of Accuracy, Sensitivity, and Specificity etc. and outperformed the likes of RF, SVM and ANN which have on numerous occasions established themselves in the field of susceptibility mapping for various hazards like landslides and floods. The proposed model achieved an accuracy of 95.5%, which justified it being contemplated for generating a susceptibility map for the region that categorized the regions on the basis of their susceptibility to wildfires. The map revealed that 1,432.025 km^2 (7%) of land in the region was very highly susceptible

to wildfires having susceptibilities above 0.678, and 1,202.356 km^2 (5%) was highly susceptible with susceptibilities in the range of (0.51-0.678). This information could aid in the government initiatives of wildlife protection and management.

5.3 Case Study III: FSM (Uttarakhand)

July 2013 unleashed one of the worst floods in the history of mankind for the North Indian state of Uttarakhand. Four years later and still little can be said about the state of affairs in the region that is still recuperating from the loss of lives, property and biodiversity. This case study attempts to utilize the state-of-the-art machine learning techniques for flood susceptibility assessment in the flood ravaged district of Chamoli. The predictive accuracy of the proposed particle swarm optimized-Support Vector Machine (PSO-SVM) model was compared against various conventional models like RF, ANN. Outperforming them; it achieved an accuracy of 96.55%. An FSZ map for the study area was generated indicating that approximately 1,832.3 km^2 (20 % of the region) area is very highly susceptible to flooding.

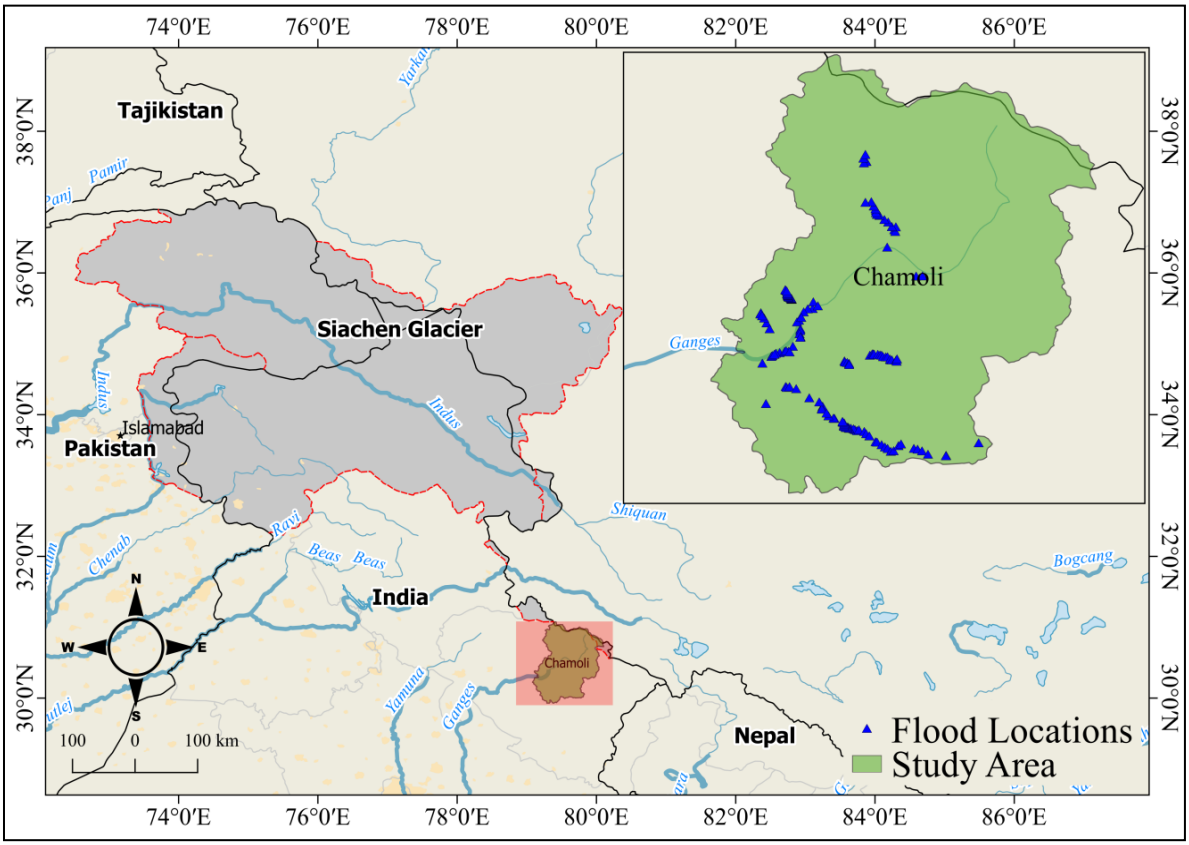

Figure 5.14: FSM Study Area

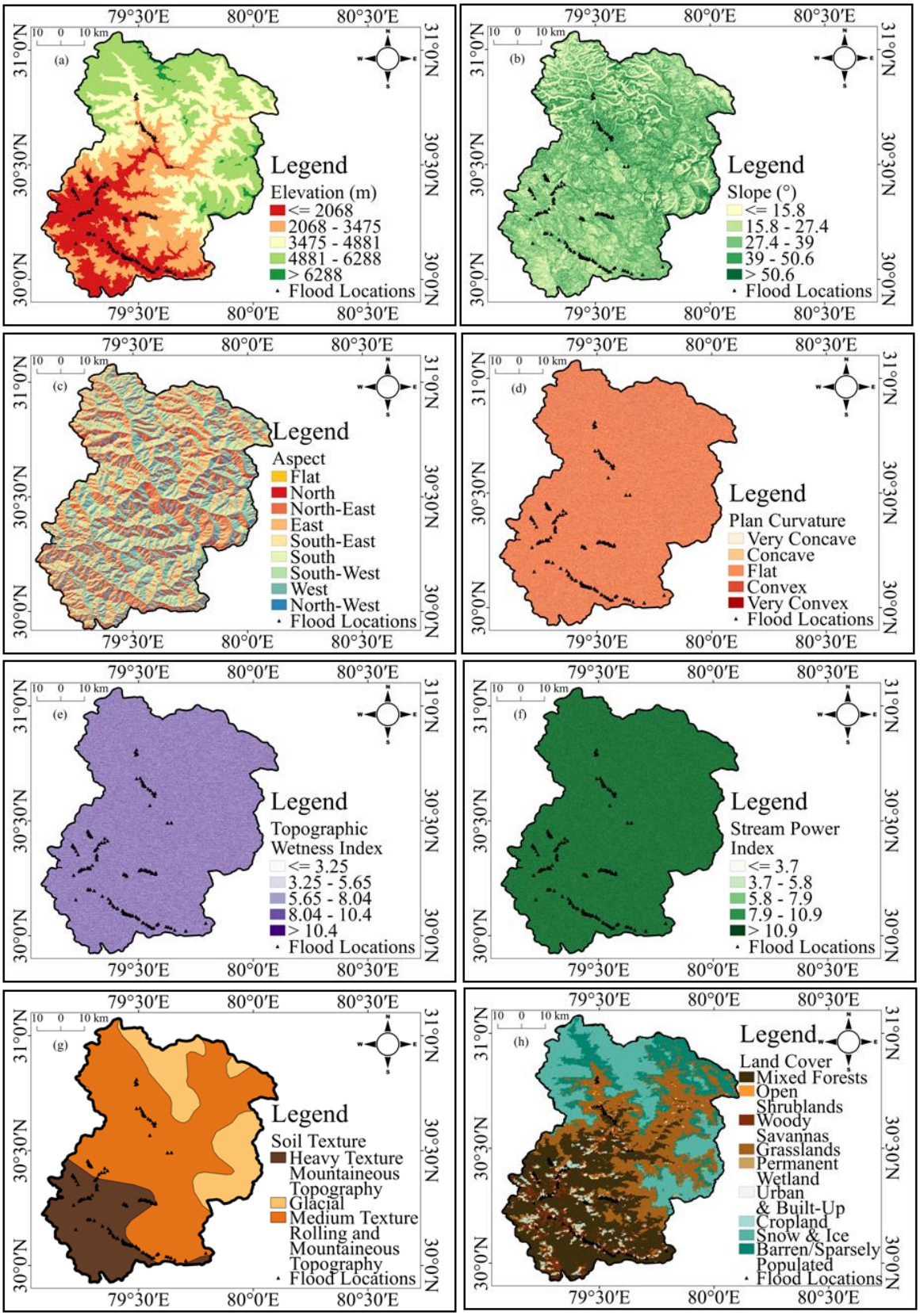

Figure 5.15: FSM Conditioning Factors (a)-(h)

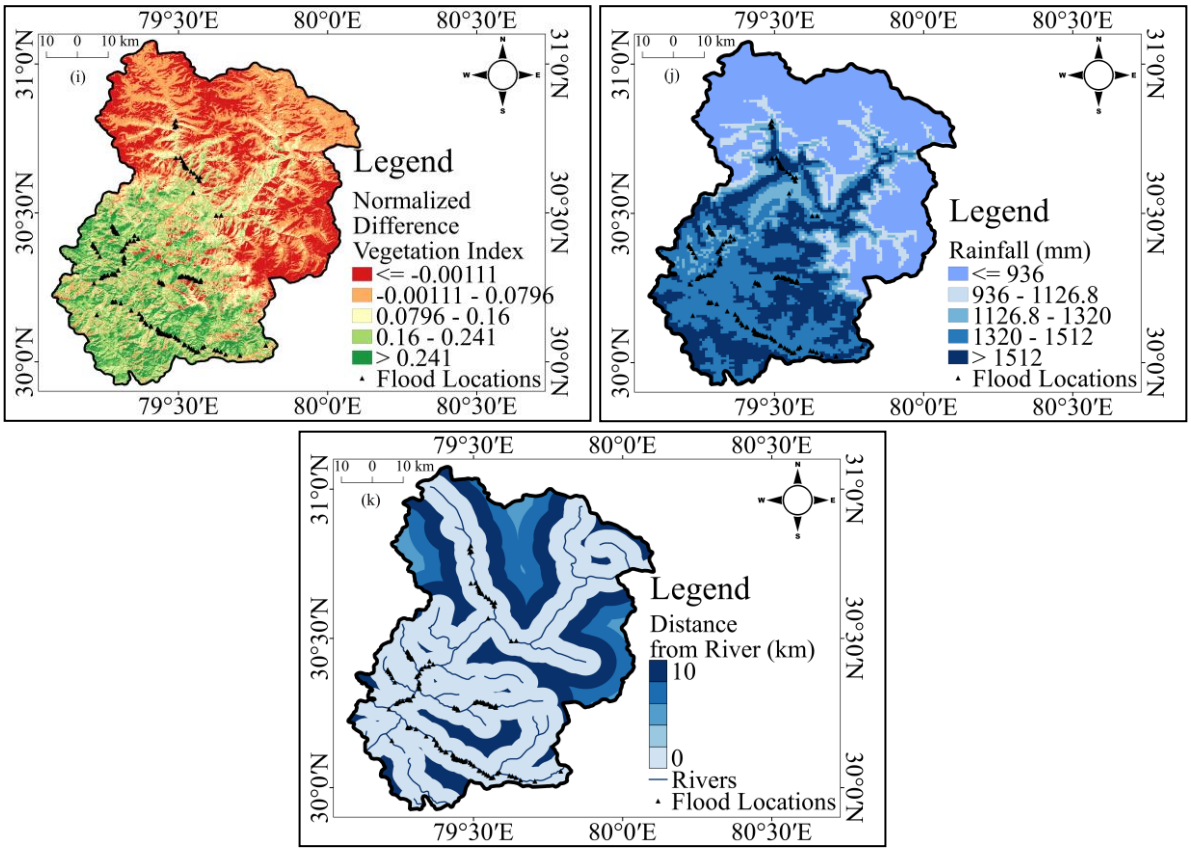

Figure 5.15 (cont.): FSM Conditioning Factors (i)-(k)

5.3.1 Spatial Data Generation and Handling

The study has been undertaken in the district of Chamoli and its surrounding areas in Uttarakhand which has been a witness of various floods in the past, like those of 2013 and 2016. The location of the study area is depicted in Figure 5.14. The study area spans over 9,120.129 km^2 of land between longitudes 79°4'47"E and 80°6'17"E and between latitudes 29°55'34"N and 31°4'29"N. The dataset used for susceptibility assessment comprises of the location coordinates (both Flood and Non-Flood) and their characteristic trait values called flood conditioning factors. The Non-Flood points are randomly obtained from the entire expanse of the study region while the flood points have identified through the observation of satellite images in Google Earth Pro using the time slider and zoom-in tools. A total of 143 flood points identified as aforementioned comprises the flood inventory. An equal number, i.e. 143 non flood points were randomly selected for the dataset to avoid bias in prediction. Of these, 70% (107 points) were used

for training while 30% (36 points) were used for testing the model. A total of 11 conditioning factors depicted in Figure 5.15 have been considered.

5.3.2 Proposed Approach

A summarized description of the proposed approach is depicted in Figure 5.16.

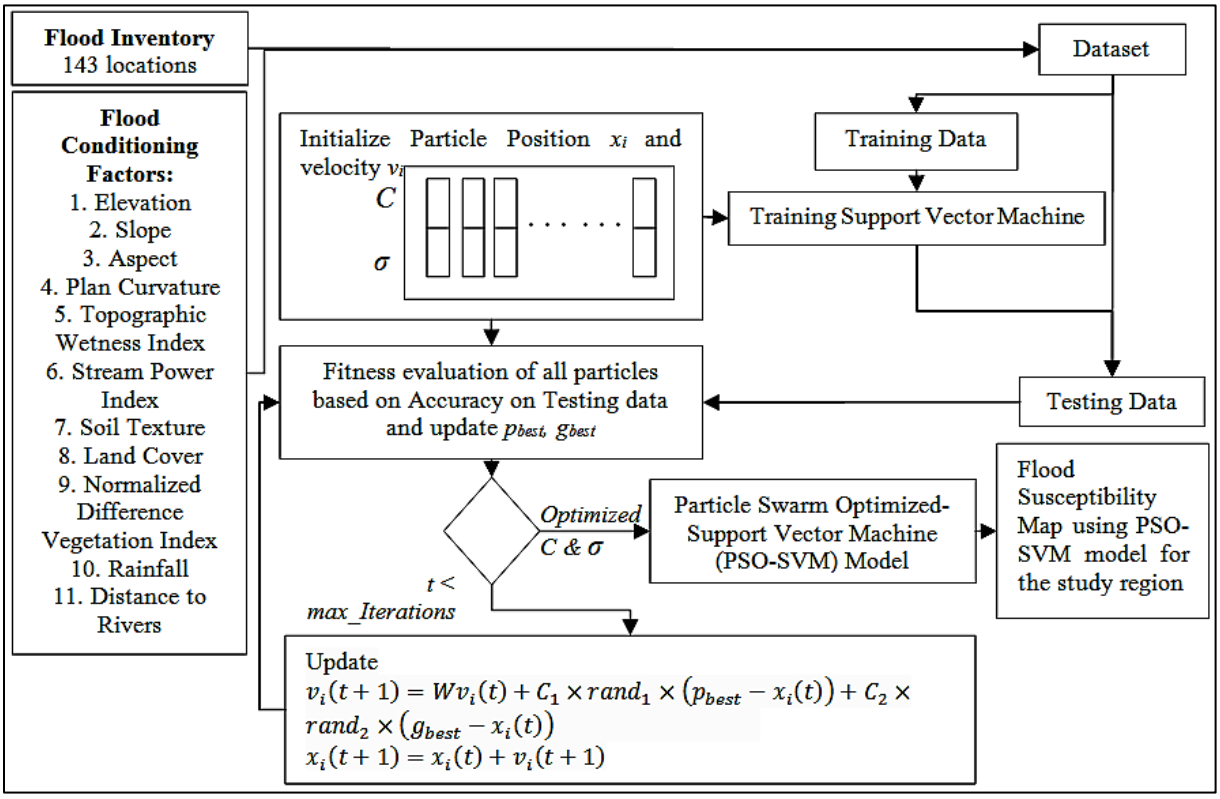

Figure 5.16: FSM Methodology

The dataset was generated using the coordinates of the location that have experienced floods called “Flood Locations” as well as those coordinates that have not experienced flooding in the past called “Non-Flood Locations” with their respective values of aforementioned 11 flood conditioning factors of elevation, slope, aspect, plan curvature, TWI, SPI, soil texture, land cover, NDVI, rainfall, and distance from rivers. To avoid bias, Flood points (represented by class value ‘1’) and Non–Flood points (represented by class value ‘0’) were taken in equal proportion of 1:1 making it a binary classification problem. 70% of this dataset were used for training the proposed model of PSO-SVM and other similar models, while 30% of the dataset were used for testing and comparing their performance on the basis of various performance statistics like Sensitivity, Specificity etc. The proposed model having outperformed the rest of the models, was

then employed to construct the FSZ map for the region demarcating the regions on the basis of their probability of flooding from very highly susceptible to very low susceptibility.

SVM: The model is based on the structural risk minimization principle [20] that focuses on searching for a hyper-plane formation in the training dataset that would be capable of separating the population/dataset-records into separate classes. In the case of FSM, each record is represented as a coordinate in the n dimensional space where n is the number of flood conditioning factors (11 here). Each record i is represented as (x_i, y_i) where x_i refers to the values of conditioning factors and y_i is the class, i.e. Flood "1" or Non-Flood "0". So the aim here is to find a hyper plane that would segregate all the coordinates into 2 separate classes of "Flood" having the class label "1" and "Non-Flood" having a class label "0" on the basis of their independent variables i.e. flood conditioning factors, with a fair amount of precision. For a good amount of accuracy, the hyper plane should be such that, there exists a maximum amount of margin between the 2 classes. If the point lies above the plane then that vector/record of conditioning factors is predicted to be a flood "1", otherwise Non-flood "0". The vectors/coordinates closest to the hyper-plane are called Support Vectors. In this study, Radial basis function has been used as the kernel function for SVM model as given in Equation (5.11) [20].

$$K(x_i, y_i) = \exp\left(-\|x_i - y_i\|^2 / 2\sigma^2\right) \tag{5.11}$$

The predictive performance here depends on the regularization parameter *C* and the kernel width σ. In order to obtain the optimized value of accuracy these parameters are optimized using particle swarm optimization (PSO).

PSO: It is a nature inspired algorithm as it takes inspiration from the swarm/collective motion of biological organisms [27]. It has been used in this study to determine the parameters (C, σ) for the SVM with Radial Basis Kernel function. It comprises of a swarm of particles that coherently search for the best position/solution. The movement of the particles in each iteration *t + 1* is defined by Equations (5.12) and (5.13) [27].

$$v_i(t+1) = Wv_i(t) + C_1 \times rand_1 \times \left(p_{best} - x_i(t)\right) + C_2 \times rand_2 \times \left(g_{best} - x_i(t)\right) \tag{5.12}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (5.13)$$

Here $x_i$is the position of the ith particle in the population with v_i being its velocity. W is the inertial weight coefficient and $rand_1$ and $rand_2$ are positive random numbers between 0 and 1 under normal distribution with C_1being the personal learning factor and C_2 is the social learning factor. p_{best} is the personal best position of ith particle and $g_{best}$the best particle among all particles. The optimization is carried iteratively, beginning with the initial population of particles in the context of PSO. The particles with their characteristic position and velocity are initialized. For each particle in the population, the SVM model is trained using the parameters associated with the corresponding particle. The trained model is then tested on the testing dataset and the predictive accuracy obtained on this dataset constitutes the corresponding particle's fitness. Within an iteration, the highest accuracy is termed as p_{best} while the highest accuracy throughout all the iterations till now, is termed g_{best}. If the requisite number of iterations *max_Iterations* has not been executed, then the particles are updated. Again the fitness for all the particles is estimated and p_{best} and g_{best} updated. The iterations are continued till the stopping criterion of a requisite number is reached. After this condition is reached the optimized parameters are obtained with the g_{best} being the highest accuracy obtained.

5.3.3 Results and Analysis

The proposed model was compared against the likes of RF, ANN, and LR. The performance statistics of Sensitivity, Specificity, etc. were used for comparing the predictive performance of trained models on the testing dataset. Their formal definitions can be referred to in Section 2.4. The results have been summarized in Table 5.7. The proposed model outperformed the conventional algorithms, thus is used to generate the FSZ map for the region as shown in Figure 5.17. For this, 4000 points spread over the study region were randomly selected and the value of their conditioning factors derived and then the trained model is used to calculate the susceptibility indexes for these points. These indexes were then interpolated using Inverse Distance Weighted interpolation thus providing the final FSZ map. The generated map indicates that an area of 1,808.226 km^2 falls under the category of very low susceptibility (<=0.127), about 1,820.886 km^2 under low susceptibility (0.127-0.212), 1,839.623 km^2 falls in the category of moderate

susceptibility (.212- .361) and 1,819.029 km^2 in the high susceptibility range (.361-.476). A total of 1,832.365 km^2 of land that is approximately 20.09% of the study area belong to the very high susceptibility category (>.476).

Table 5.7: FSM: Model Comparison of PSO-SVM against benchmark models

Sr. No.	Model Performance Metric	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Accuracy
1.	PSO-SVM	100	93.75	92.86	93.75	96.55
2.	Majority Based LR-GBDT-VFI	95.37	90.7	91.11	90.7	93.02
3.	EO-GBDT	95.37	90.7	91.11	90.7	93.02
4.	SVM	97.67	90.7	91.3	90.7	94.19
5.	GBDT	97.67	90.7	91.3	90.7	94.19
6.	ANN	100	88.37	89.58	88.37	94.19
7.	RF	97.67	90.7	91.3	90.7	94.19
8.	LR	95.35	88.37	89.13	88.37	91.86
9.	NB	97.67	88.37	89.36	88.37	94.19
10.	DT	93.02	95.35	95.24	95.35	94.19
11.	VFI	95.35	86.05	87.23	86.05	90.7

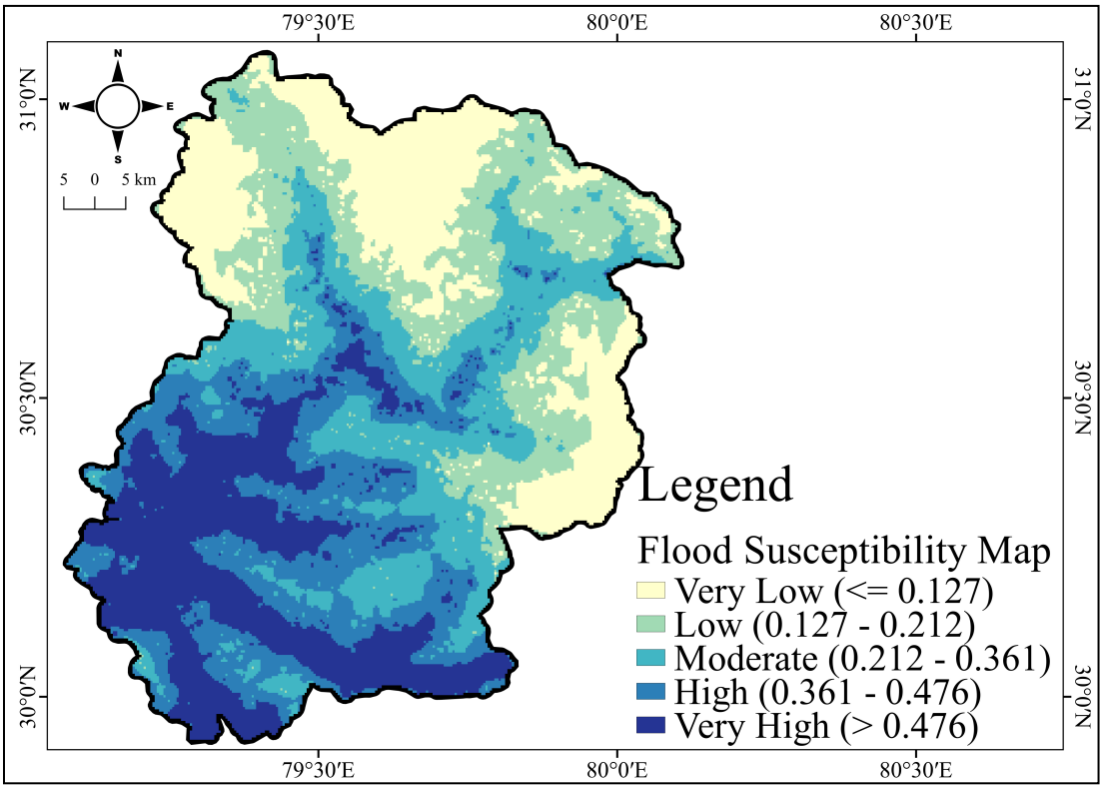

Figure 5.17: FSZ Map

5.3.4 Summary

Floods being one of the most widespread and fatal disasters, so there is an immediate need to minimize its impact. This study is an attempt in this direction. It explored a novel technique of PSO-SVM for predicting floods using the flood conditioning factors and prepared the susceptibility map for the flood-prone district of Chamoli which has in the past encountered various episodes of devastating floods. The proposed technique PSO-SVM (96.55%) outperformed various other techniques like RF (94.19%), LR (91.86%) etc. in terms of accuracy. The FSZ map revealed that nearly 20% of the study region are very highly susceptible to flooding. Such information can come in handy in mitigating disasters in the future with appropriate preparation and by placing strong precautionary measures in place.

CHAPTER 6

CONCLUSION & FUTURE SCOPE

6.1 Conclusion

One of the most hazard affected countries in the Asian continent is India. A particular region's susceptibility towards a natural hazard varies from other regions depending on the region's traits such as geology, morphology, climate, etc. An automated analysis of such traits can be used in predicting the level of susceptibility a particular region falls into. This process is termed as HSM. This work undertook three such hazard susceptibility assessment studies with the means of geospatial data generated from aerial images and thematic maps using GIS. The spatial dataset generated was then processed by employing state-of-the-art machine learning classifiers with performance enhancements through optimization and ensembling for delineating the region on the basis of its susceptibility to the corresponding hazard.

The Case Study I undertook an LSM focused on Majority-based voting, which has seldom been employed for landslide susceptibility assessment. The ensemble comprised of LR, GBDT and VFI to prepare LSZ map for the Brahmaputra valley region (Assam & Nagaland) and its close vicinity. The ensemble model proved its prowess over the conventional standalone classifiers, including its own constituent models achieving an AUC of 0.984. Also, the model achieved an accuracy of 96.66% outperforming SVM (91.19%), LR (92.71%) and RF (91.49%). The model was thus employed for generating the region's LSZ map. The map indicated that 6.8% (1360 km^2) of the area fall under very high susceptibility while 7.2% (1440 km^2) under high susceptibility zone.

Case Study II involved an FFSM for NDBR and explored EO-GBDT for preparing an FFSZ map for the study area. The proposed model outperformed the likes of RF, SVM and ANN. The proposed model achieved an accuracy of 95.5%, which justified it being contemplated for generating an FFSZ map for the region that categorized the regions on the basis of their susceptibility to wildfires. The map revealed that 1,432.025 km^2 (7%) of land in the region was very highly susceptible to wildfires having susceptibilities above 0.678, and 1,202.356 km^2 (5%) was highly susceptible with susceptibilities in the range of (0.51-0.678).

Finally, an FSM was undertaken in Case Study III for the flood-ravaged district of Chamoli. It explored a novel technique of PSO-SVM for predicting floods and prepared the FSZ map. The proposed technique PSO-SVM outperformed various other techniques like RF (94.19%), LR (91.86%) etc. in terms of accuracy (96.55%). The FSZ map revealed that nearly 20% of the study region are very highly susceptible to flooding.

6.2 Future Scope

There is an imminent need in India for various such studies to be undertaken as there are various hazard-prone areas that lie in the immediate threat of wide scale hazard-induced devastation. The proposed approach holds solutions for dealing with the cruel vagaries of nature. An all factor inclusive study holds a promise for generating highly accurate susceptibility maps which could lay the foundations for immaculate and strongly footed hazard mitigation efforts. Also, a due accreditation of dynamic factors of LULC, Rainfall extends the limitless scope in this field. The conditioning factors of LULC and rainfall have traditionally been employed as static factors. However, the factors have metamorphosed into dynamic ones due to climate change and environment degradation, driven by the fast pace of development and evolution, rendering a new perspective to the hazard assessment studies. The dynamicity of these factors could be accounted for by extending the original information extracted from their formats into their possible future status, predicted via the techniques of machine learning, pattern recognition, etc. The statistical approaches of Analytical Hierarchy Process and WoE have rendered themselves useful in a number of similar studies. So a possible hybrid approach amalgamating machine learning with these aforementioned techniques, could contribute to this emerging field of HSM. Comparing such a hybrid approach against the individual models of WoE, ANN, SVM, etc. could lead the way for future HSM studies and empower the mankind with indigenous ways of handling hazards.

REFERENCES

[1] D. Alexander, *Natural disasters*. London: Routledge, 2002.

[2] M. Kahn, "The Death Toll from Natural Disasters: The Role of Income, Geography, and Institutions", *Review of Economics and Statistics*, vol. 87, no. 2, pp. 271-284, 2005.

[3] S. Goswami, S. Chakraborty, S. Ghosh, A. Chakrabarti and B. Chakraborty, "A review on application of data mining techniques to combat natural disasters", *Ain Shams Engineering Journal*, 2016.

[4] J. Yin, A. Lampert, M. Cameron, B. Robinson and R. Power, "Using Social Media to Enhance Emergency Situation Awareness", *IEEE Intelligent Systems*, vol. 27, no. 6, pp. 52-59, 2012.

[5] S. Kundu, A. Saha, D. Sharma and C. Pant, "Remote Sensing and GIS Based Landslide Susceptibility Assessment using Binary Logistic Regression Model: A Case Study in the Ganeshganga Watershed, Himalayas", *Journal of the Indian Society of Remote Sensing*, vol. 41, no. 3, pp. 697-709, 2013.

[6] Home - National Disaster Management Authority", *Ndma.gov.in*, 2017. [Online]. Available: http://www.ndma.gov.in/. [Accessed: 5- Jan- 2017].

[7] Nagaland State Disaster Management Authority", *Nsdma.gov.in*, 2017. [Online]. Available: http://nsdma.gov.in. [Accessed: 5- Jan- 2017].

[8] Assam Disaster Management Authority", *Sdmassam.nic.in*, 2017. [Online]. Available: http://sdmassam.nic.in. [Accessed: 5- Jan- 2017].

[9] Earth Policy Institute – Building a Sustainable Future | Home", *Earth-policy.org*, 2017. [Online]. Available: http://www.earth-policy.org. [Accessed: 20- Feb- 2017].

[10] FIRE GLOBE: The Global Fire Monitoring Center (GFMC)", *Fire.uni-freiburg.de*, 2017. [Online]. Available: http://www.fire.uni-freiburg.de. [Accessed: 20- Feb- 2017].

[11] R. Jaiswal, S. Mukherjee, K. Raju and R. Saxena, "Forest fire risk zone mapping from satellite imagery and GIS", *International Journal of Applied Earth Observation and Geoinformation*, vol. 4, no. 1, pp. 1-10, 2002.

[12] P.M. Sangal. “Suggested classification of forest fires in India by types and causes”, in *National Seminar on Forest Fire Control*, Kulamavu, Trivandrum, 1981.

[13] National Aeronautics and Space Administration", *NASA*, 2017. [Online]. Available: https://www.nasa.gov. [Accessed: 20- Feb- 2017].

[14] Latest News Online | The Indian Express", *Indianexpress.com*, 2017. [Online]. Available: http://indianexpress.com. [Accessed: 20- Feb- 2017].

[15] UNISDR", *Unisdr.org*, 2017. [Online]. Available: https://www.unisdr.org. [Accessed: 19- Mar- 2017].

[16] IFRC.org - IFRC", *Ifrc.org*, 2017. [Online]. Available: http://www.ifrc.org. [Accessed: 19- Mar- 2017].

[17] C. Kala, "Deluge, disaster and development in Uttarakhand Himalayan region of India: Challenges and lessons for disaster management", *International Journal of Disaster Risk Reduction*, vol. 8, pp. 143-152, 2014.

[18] News: National & International News | Times of India", *The Times of India*, 2017. [Online]. Available: http://timesofindia.indiatimes.com/. [Accessed: 19- Mar- 2017].

[19] D. Cruden, "A simple definition of a landslide", *Bulletin of the International Association of Engineering Geology*, vol. 43, no. 1, pp. 27-29, 1991.

[20] I. Colkesen, E. Sahin and T. Kavzoglu, "Susceptibility mapping of shallow landslides using kernel-based Gaussian process, support vector machines and logistic regression", *Journal of African Earth Sciences*, vol. 118, pp. 53-64, 2016.

[21] B. Pham, D. Tien Bui, P. Indra and M. Dholakia, "Rotation forest fuzzy rule-based classifier ensemble for spatial prediction of landslides using GIS", *Natural Hazards*, vol. 83, no. 1, pp. 97-127, 2016.

[22] V. Moosavi and Y. Niazi, "Development of hybrid wavelet packet-statistical models (WP-SM) for landslide susceptibility mapping", *Landslides*, vol. 13, no. 1, pp. 97-114, 2015.

[23] B. Pham, D. Tien Bui, P. Indra and M. Dholakia, "Landslide Susceptibility Assessment at a Part of Uttarakhand Himalaya, India using GIS – based Statistical Approach of Frequency Ratio Method", *International Journal of Engineering Research and*, vol. 4, no. 11, 2015.

[24] B. Pham, D. Tien Bui, H. Pourghasemi, P. Indra and M. Dholakia, "Landslide susceptibility assesssment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and

functional trees methods", *Theoretical and Applied Climatology*, vol. 128, no. 1-2, pp. 255-273, 2015.

[25] H. Nefeslioglu, T. Duman and S. Durmaz, "Landslide susceptibility mapping for a part of tectonic Kelkit Valley (Eastern Black Sea region of Turkey)", *Geomorphology*, vol. 94, no. 3-4, pp. 401-418, 2008.

[26] A. Akgun and N. Türk, "Landslide susceptibility mapping for Ayvalik (Western Turkey) and its vicinity by multicriteria decision analysis", *Environmental Earth Sciences*, vol. 61, no. 3, pp. 595-611, 2009.

[27] D. Tien Bui, Q. Bui, Q. Nguyen, B. Pradhan, H. Nampak and P. Trinh, "A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area", *Agricultural and Forest Meteorology*, vol. 233, pp. 32-44, 2017.

[28] H. Adab, K. Kanniah and K. Solaimani, "Modeling forest fire risk in the northeast of Iran using remote sensing and GIS techniques", *Natural Hazards*, vol. 65, no. 3, pp. 1723-1743, 2012.

[29] X. Gao, X. Fei and H. Xie, "Forest fire risk zone evaluation based on high spatial resolution RS image in Liangyungang Huaguo Mountain Scenic Spot", in *Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2011 IEEE International Conference on*, Fuzhou, China, 2011, pp. 593-596.

[30] H. Pourghasemi, "GIS-based forest fire susceptibility mapping in Iran: a comparison between evidential belief function and binary logistic regression models", *Scandinavian Journal of Forest Research*, vol. 31, no. 1, pp. 80-98, 2015.

[31] D. Tien Bui, K. Le, V. Nguyen, H. Le and I. Revhaug, "Tropical Forest Fire Susceptibility Mapping at the Cat Ba National Park Area, Hai Phong City, Vietnam, Using GIS-Based Kernel Logistic Regression", *Remote Sensing*, vol. 8, no. 4, p. 347, 2016.

[32] M. Tehrany, B. Pradhan and M. Jebur, "Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS", *Journal of Hydrology*, vol. 512, pp. 332-343, 2014.

[33] K. Khosravi, H. Pourghasemi, K. Chapi and M. Bahri, "Flash flood susceptibility analysis and its mapping using different bivariate models in Iran: a comparison between

Shannon’s entropy, statistical index, and weighting factor models", *Environmental Monitoring and Assessment*, vol. 188, no. 12, 2016.

[34] B. Ahmed, "Landslide susceptibility mapping using multi-criteria evaluation techniques in Chittagong Metropolitan Area, Bangladesh", *Landslides*, vol. 12, no. 6, pp. 1077-1095, 2014.

[35] S.L. Bhutia et al. “A Survey on Landslide Susceptibility Mapping Using Soft Computing Techniques” *IOSR Journal of Applied Geology and Geophysics, v*ol. 3, no. 1, pp. 16-20. 2015

[36] D. Tien Bui, T. Tuan, H. Klempe, B. Pradhan and I. Revhaug, "Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree", *Landslides*, vol. 13, no. 2, pp. 361-378, 2015.

[37] A. Akgun, S. Dag and F. Bulut, "Landslide susceptibility mapping for a landslide-prone area (Findikli, NE of Turkey) by likelihood-frequency ratio and weighted linear combination models", *Environmental Geology*, vol. 54, no. 6, pp. 1127-1143, 2007.

[38] D. Tien Bui, B. Pradhan, O. Lofman and I. Revhaug, "Landslide Susceptibility Assessment in Vietnam Using Support Vector Machines, Decision Tree, and Naïve Bayes Models", *Mathematical Problems in Engineering*, vol. 2012, pp. 1-26, 2012.

[39] L. Ayalew and H. Yamagishi, "The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan", *Geomorphology*, vol. 65, no. 1-2, pp. 15-31, 2005.

[40] C. Melchiorre, M. Matteucci, A. Azzoni and A. Zanchi, "Artificial neural networks and cluster analysis in landslide susceptibility zonation", *Geomorphology*, vol. 94, no. 3-4, pp. 379-400, 2008.

[41] H. Hong, B. Pradhan, M. Jebur, D. Bui, C. Xu and A. Akgun, "Spatial prediction of landslide hazard at the Luxi area (China) using support vector machines", *Environmental Earth Sciences*, vol. 75, no. 1, 2015.

[42] O. Rahmati, H. Pourghasemi and A. Melesse, "Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran", *CATENA*, vol. 137, pp. 360-372, 2016.

[43] J. Ghosh and D. Bhattacharya, "Knowledge-Based Landslide Susceptibility Zonation System", *Journal of Computing in Civil Engineering*, vol. 24, no. 4, pp. 325-334, 2010.

[44] M.K. Kumar and R. Annadurai. “Mapping of Landslide Susceptibility Using Geospatial Technique-A Case Study in Kothagiri Region, Western Ghats, Tamil Nadu, India.”, *International Journal of Engineering Research and Technology,* vol. 2, no. 12, 2013.

[45] N. Hoang and D. Tien Bui, "A Novel Relevance Vector Machine Classifier with Cuckoo Search Optimization for Spatial Prediction of Landslides", *Journal of Computing in Civil Engineering*, vol. 30, no. 5, p. 04016001, 2016.

[46] Q. Ding, W. Chen and H. Hong, "Application of frequency ratio, weights of evidence and evidential belief function models in landslide susceptibility mapping", *Geocarto International*, pp. 1-21, 2016.

[47] F.C. Eugenio, A.R. Dos Santos, N.C. Fiedler, G.A. Ribeiro, A.G. da Silva, A.B. Dos Santos, V.R. Schettino. “Applying GIS to develop a model for forest fire risk: a case study in Espírito Santo, Brazil”, *Journal of environmental management*, pp.65-71, 2016

[48] A. Arpaci, B. Malowerschnig, O. Sass and H. Vacik, "Using multi variate data mining techniques for estimating fire susceptibility of Tyrolean forests", *Applied Geography*, vol. 53, pp. 258-270, 2014.

[49] O. Satir, S. Berberoglu and C. Donmez, "Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem", *Geomatics, Natural Hazards and Risk*, vol. 7, no. 5, pp. 1645-1658, 2015.

[50] K.V. Suryabhagavan, M. Alemu, M. Balakrishnan. “GIS-based multi-criteria decision analysis for forest fire susceptibility mapping: a case study in Harenna forest, southwestern Ethiopia”, *Tropical Ecology*, vol. 57, no. 1, pp. 33-43, 2016.

[51] M. Lee, J. Kang and S. Jeon, "Application of frequency ratio model and validation for predictive flooded area susceptibility mapping using GIS", in *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, Munich, Germany, 2012, pp. 895-898.

[52] B. Pradhan. “Flood susceptible mapping and risk area delineation using logistic regression, GIS and remote sensing”, *Journal of Spatial Hydrology,* vol. 9, no. 2, 2010.

[53] M. Tehrany, B. Pradhan and M. Jebur, "Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS", *Journal of Hydrology*, vol. 504, pp. 69-79, 2013.

[54] L. Lombardo, M. Cama, C. Conoscenti, M. Märker and E. Rotigliano, "Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy)", *Natural Hazards*, vol. 79, no. 3, pp. 1621-1648, 2015.

[55] C. Ding, D. Wang, X. Ma and H. Li, "Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees", *Sustainability*, vol. 8, no. 12, p. 1100, 2016.

[56] G. Demiröz and H. A. Güvenir, "Classification by Voting Feature Intervals", in *9th European Conference on Machine Learning*, Prague, Czech Republic, 1997, pp. 85-92.

[57] L. Rokach, "Ensemble-based classifiers", *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1-39, 2009.

[58] J. Branke, *Evolutionary optimization in dynamic environments*. Springer Science & Business Media, vol. 3, 2013.

LIST OF PUBLICATIONS

1. S Sachdeva, T Bhatia, AK Verma, "Flood Susceptibility Mapping using GIS-based Support Vector Machine and Particle Swarm Optimization: A case study in Uttarakhand (India)", In Proceedings of The Eight International Conference On Computing, Communication And Networking Technologies (ICCCNT 2017) IEEE, *IIT Delhi*, 3-5th July, 2017. (Accepted and Presented)
2. S Sachdeva, T Bhatia, AK Verma, "GIS-based Evolutionary Optimized Gradient Boosted Decision Trees for Forest Fire Susceptibility Mapping". (Communicated)
3. S Sachdeva, T Bhatia, AK Verma, "A voting ensemble logistic regression with gradient boosted decision trees and voting feature interval models for spatial prediction of landslides using GIS". (Communicated)

PLAGIARISM REPORT

ORIGINALITY REPORT

%7	%3	%5	%3
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Tien Bui, Dieu, Quang-Thanh Bui, Quoc-Phi Nguyen, Biswajeet Pradhan, Haleh Nampak, and Phan Trong Trinh. "A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area", Agricultural and Forest Meteorology, 2017. Publication	<%1
2	Submitted to Multimedia University Student Paper	<%1
3	Submitted to Birla Institute of Technology Student Paper	<%1
4	Submitted to Sanger High School Student Paper	<%1
5	Pourghasemi, Hamid Reza. "GIS-based forest fire susceptibility mapping in Iran: a comparison between evidential belief function and binary logistic regression models", Scandinavian Journal of Forest Research,	<%1