A Bayesian Approach to Lightning Ground-Strike Points Analysis

Wandile Lesejane
Student Number: 929538

Supervisors:
Dr. Ritesh Ajoodha
Dr. Hugh Hunt

A research report submitted in partial fulfillment of the requirements for the degree of Master of Science in the field of e-Science

in the

School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg

14 November 2022

Declaration

I, Wandile Lesejane, declare that this research report is my own, unaided work. It is being submitted for the degree of Master of Science in the field of e-Science at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.

wk. lesejane

Wandile Lesejane

14 November 2022

Abstract

Studying cloud-to-ground lightning strokes and ground strike points provides an alternative method of lightning mapping for lightning risk assessment. Various k-means algorithms have been used to verify the ground strike points from lightning locating systems. These algorithms produce results but have the potential to be improved.
This research report proposes using Bayesian Network which is a model that has not been used before to verify lightning ground strike points. A Bayesian Network is a probabilistic graphical model that makes use of Bayes Theorem to represent the conditional dependencies of variables.
The network created for this research were learned from the data using a score-based structure learning and the Bayesian Information Criterion score function was used. The models were evaluated using a confusion matrix and a kappa index. They produced accuracies ranging from 86% to 94% with a kappa index of up to 0.76.
The results from the Bayesian Network models are within the range of the available algorithms used currently to analyse lightning ground strike points but have an advantage of not needing a predetermined distance, easy to interpret and as well as being suitable for small data sets. The use of a Bayesian network is a good candidate for an alternative method to analyse lightning ground strike points.

Acknowledgements

A special thanks to my supervisors, Dr. Ritesh Ajoodha and Dr. Hugh Hunt for their guidance with this project.
Thank you Dr. Carina Schumann for the supporting material and comments on the research report.
The support of the DST-CSIR Nationl e-Science Postgraduate Teaching and Training Platform (NEPTTP) towards this research is hereby acknowledged.
Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NEPTTP.

Contents

List of Figures

List of Tables

List of Abbreviations

LLS	**L**ightning **L**ocating **S**ystem
IEC	**I**nternational **E**lectrotechnical **C**ommission
GSP	**G**round **S**trike **P**oint
BN	**B**ayesian **N**etwork
PSP	**P**revious **S**trike **P**oint
NGC	**N**ew **G**round **C**hannel
PEC	**P**re-**e**xisting **C**hannel
CC	**C**loud-to-**C**loud
CG	**C**loud-to-**G**round
ELF	**E**xtremely **L**ow **F**requency
UHF	**U**ltra **H**igh **F**requency
TOA	**T**ime-**o**f-**a**rrival
MDF	**M**agnetic **D**irection **F**inding
SMA	**S**emi-**M**ajor **A**xes
DAG	**D**irected **A**cyclic **G**raph
MLE	**Maximum** **L**ikelihood **E**stimation
AIC	**A**kaike **I**nformation **C**riterion
BIC	**B**ayesian **I**nformation **C**riterion
TP	**T**rue **P**ositive
TN	**T**rue **N**egative
FP	**F**alse **P**ositive
FN	**F**alse **N**egative

Chapter 1

Introduction

Lightning is an electrostatic phenomenon that occurs naturally in the atmosphere [1]. Lightning occurs mostly within the clouds in the atmosphere but often at times it reaches the ground. The energy that produced during a lightning stroke has been estimated to more than 5×10^7 J, this amount of energy has detrimental effects on our livelihoods when it comes in contact with grounded objects or human beings [2, 3, 4, 5]. It is, therefore, important that we thoroughly understand the risks associated with lightning strokes and use that information to preserve lives and infrastructure.

FIGURE 1.1: A video frame of lighting striking a high rise building in Sandton, Johannesburg. Video captured by Lark and Owl Productions.

Figure 1.1 is a still frame from a video captured by Lark and Owl Productions. The video is a time lapse of a thunderstorm occurring in Sandton, Johannesburg, which means the image is a stacked image of multiple frames and it has a high resolution. On the image, lightning flash can be seen striking one of the high rise building in the area with multiple lightning strokes.

Lightning Locating Systems (LLSs) are multi-sensors used to collect data from lightning return strokes. The strokes are grouped into light flashes according to the International Electrotechnical Commission (IEC) 62585 to create lightning density maps for lightning risk assessment [6, 7]. Location of the first stroke is taken as the location of the flash which becomes an underestimation problem when the flash consists of multiple strokes with different Ground Strike Points (GSPs) [6], thus using stroke data may result in more accurate lightning density maps [8]. LLSs use various techniques and rely on triangulation to locate lightning strokes which bring about uncertainties due technical errors [9].

There are various machine learning algorithms used to analyse the lightning GSPs [8, 10] to identify the locations of strokes in each lightning flash. These algorithms are all variations of k-means method [10]. The results from the algorithms were compared with ground truth data which is observed using a high speed cameras [8]. The results obtained from the algorithms output accuracies ranging between 79% and 95% [10].

This research report introduces the use of Bayesian Networks (BN) as another algorithm for GSP analysis. Bayesian statistics has been proposed and used to validate GSPs and to assess the performance of LLSs [11, 12] but in the form of Bayes theorem whereby the prior probability and likelihood were estimated from current literature. The method of Bayesian network for (GSP) analysis will use the variables to determine their conditional dependencies and independencies with respect to the target variable.

A target variable, Previous Strike Point (PSP) is introduced from the ground truth data by matching it to New Ground Channel (NGC) or Pre-existing Channel (PEC). There are five BN models created using the data as it is in one of the models and by setting conditional dependencies between the features of the data. The models are learned using a score-based structure learning procedure to obtain the structure of the networks. The models are then cross validated to obtain the results.

The BN models output accuracies ranging between 86% and 94% with mean classification errors between 11% and 15%. The kappa statistics of the models were good with the lowest at 0.53 and the highest at 0.76. The models are good contenders for an alternative to the current k-means methods as they output similar outputs with low errors. The BN models are able to handle class imbalance as observed from their high kappa coefficients.

The contributions that this research report will make are as follows: *(a)* introducing a new application of BN to the analysis of GSPs, *(b)* alleviate uncertainties associated with predetermined distance threshold which is a requirement in all the k-means algorithms, and *(c)* the set conditional dependencies between the variables gives information on which features are the most significant to identify lightning strokes.

1.1 Research Aims and Objectives

The proposed method of this research report is using Bayesian networks to validate the location of GSPs from the LLSs. This method has not been used before and could prove as an improvement or an alternative to the current methods used for GSP analyses. Bayesian networks make use of statistics and correlation between variables to obtain the desired results.

1.1.1 Research Aims

The aim of this research is to classify lightning strokes that are part of a single flash using Bayesian Network models. The provided lightning data was captured during thunderstorms in Johannesburg by Lightning Locating Systems. The use of Bayesian networks for ground-strike point analysis serves as an approach that differs from the current K-means methods and could lead to improved GSP classification results.

1.1.2 Objectives

The objectives of the research are as follows:

- To obtain the data of lightning strokes
- To select the relevant features
- To categorise the data set into relevant groups according to the IEC 62858
- Develop a Bayesian network model to fit the data
- Analyse the performance of the model

1.2 Limitations

The research report has some limitations that are due to the lightning data and resources for data modelling.

- The data used might be too small which could lead to model underfitting or overfitting.
- There is a class imbalance which affects the model classification performance.
- The necessary resources for the research such as architecture and literature of Bayesian networks are limited.

1.3 Overview

This research report will proceed as follows; background of the research will be discussed in Chapter 2, then literature review will be discussed in Chapter 3. Chapter 4 consists of the discussion of the methodology which also includes the preprocessing steps. Chapter 5 highlights and discuses the results of the research. Chapter 6 is the conclusion of the research report.

Chapter 2

Background

This chapter focuses on the background of lightning. It first discusses the phenomenon of lightning which is a description of how lightning is formed and other aspects of lightning. Then it discusses lightning locating system and the detection methods used by these systems. Then it concludes by discussing lightning risk assessment.

2.1 Lightning Phenomenon

Lightning develops in the atmosphere due to storms, volcanic eruptions and dust storms [13]. These changes in the atmosphere result in a pressure and temperature change, wind and friction, and clouds are formed. During cloud formation, there is an electric charge separation due to various possible reasons such as ion charging, convection, induction and drop break-up [1, 14].

The charge separation makes the cloud a bipole region with a net charge of zero [15]. Charges are then driven by an electric force towards a region with opposite polarity. The propagation of the charges in motion, called leaders [13], leaves a trail of light until they reach an opposite charge and make a connection [1, 15]. Once the connection between the two opposite charges is made, electrons begin to flow to the positive region through the channel created by leaders, a lightning return stroke is thus achieved [13].

A leader connection that occurs within the clouds is called cloud-to-cloud (CC) and a leader that connects to the ground or objects on the surface of the ground is called cloud-to-ground (CG). Leaders with negative charges are called negative

CG stroke or negative CC stroke and leaders with positive charges are called positive CG stroke or positive CC stroke [1]. CG lightning strokes are very significant as they have an effect on us [3, 4, 5], they can damage infrastructure and cause death.

From previous observations, most CGs are identified as negative flashes whereby all the strokes of a flash have negative leaders [16]. A single flash can have one or more strokes, these strokes occur within a short interval [16, 17]. The International Lightning Protection Association Symposium came up with the IEC 62858 to standardize flash data and modern LLSs use that standard to categorize strokes into flashes[6]. According to the IEC 62858, a flash has strokes that occur within 10 km of each other, have a 500 ms inter-stroke delay and all occur within 1 second [6].

The strokes in a single flash follow the same channel, the initial stroke creates a conductive path as electrons propagate along the channel [13]. This in turn makes the succeeding strokes propagate at a faster rate [17, 18, 19], making the strokes occur within an interval of 40 ms and 50 ms on average but the maximum flash duration observed by with high speed camera was about 1400 ms [16].

A flash connects to the ground at a ground-strike point (GSP). On average, a flash consists of about 4 or 5 strokes, these strokes do not necessarily connect to the same GSP, and a flash with multiple strokes is more likely to have multiple GSPs [16, 20, 21]. Each stroke has large amounts of current flowing continuously for its duration, thus CG flashes with longer continuous current release large amounts of energy onto the surface of the ground or object on the ground.

Figure 2.1 is a schematic of a flash hierarchy. A single flash with four strokes, three of the strokes hit the ground at the same point while one hit the ground at another point, therefore there are 2 GSPs. Figure 2.2 is a schematic of the occurrence of the flash, 2.2a depicts the locations of the of the strokes while 2.2b depicts the occurrence of the lightning flash over a period time.

The energy produced by CG flashes may reach up to 5×10^7 J, such high energies produce air temperatures that are high enough to spark bush fires. The energy is sufficient to damage infrastructure such as electricity cables and buildings, and

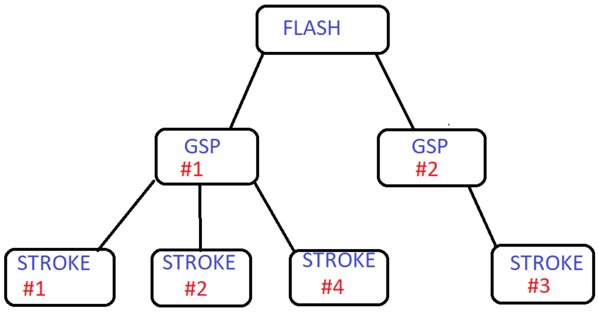
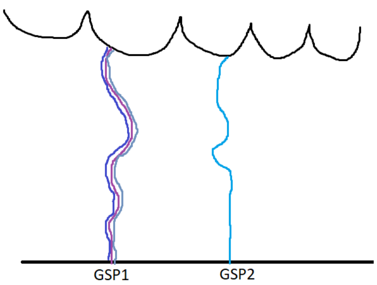
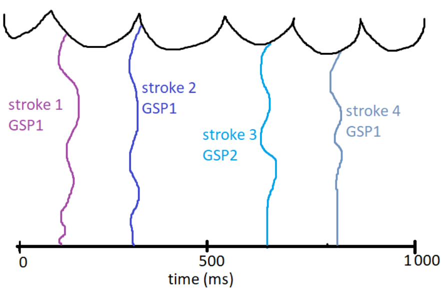


FIGURE 2.1: A diagram which is a schematic of a flash hierarchy adapted from Pédeboy [6]. It shows a single flash with two GSPs and four strokes, stroke 1,2 and 4 are at GSP1 and stroke 3 is at GSP 2.



(A) Illustration of lightning flash with two GSPs.



(B) Illustration of lightning occurrence over a period of time.

FIGURE 2.2: A diagram which shows a flash with two GSPs and four strokes occurring over time, adapted from Valine and Krider [22]. Stroke 1,2 and 4 are at GSP1 and stroke 3 is at GSP 2

also result in fatalities when they are in contact with human beings [3, 4, 23]. Thus, monitoring lightning is significant as it helps with the mitigation of risks brought about by lightning strikes.

2.2 Lightning Locating

There are various systems in place that monitor lightning. Each system is created differently to observe the different aspects of lightning. These systems are places on the surface of the Earth and above the atmosphere to detect both cloud-to-cloud lightning and cloud-ground lightning. These systems are used in various industries such as meteorology, geophysics, insurance and electric power to locate lightning strokes for weather casting and lightning safety and warning purposes [24].

Lightning Locating Systems (LLSs) are automatic multi-sensors that have been used since the 1920s [9]. Lightning flashes radiate electromagnetic waves at various frequencies of the spectrum [9, 25]. They range from Extremely Low Frequency (ELF) to Ultra High Frequency (UHF). Events of X-ray [25] and Gamma ray [26] have been observed during lightning flashes. Low Frequencies occur due to continuous lightning strokes and High Frequencies are the result of short pulses of lightning discharge [9].

2.3 Detection Methods

There are various techniques used to detect lightning, each with its own advantages and disadvantages. There are three popular techniques currently used for lightning detection which are Time-of-arrival (TOA) which uses time specific occurrences of electromagnetic waves [24], Magnetic Direction Finding (MDF) which uses interferometry [24, 9] and optical imaging which relies on photography of geometric projections of lightning flashes [24].

TOA detectors are sensors that receive data of electromagnetic pulses from the atmosphere called *sferics* [9, 27]. These sensors function by timing the arrival of sferics. For a more accurate measure of location there needs to be more than one sensor observing the same wave [9, 24]. Each sensor outputs a hyperbola of time differences between feature occurrences of an event and the intersection of these hyperbolae from these sensors give location of lightning origin [9, 27]. This technique solves the problem of polarisation errors [27].

MDF is also referred to as Magnetic Direction Finding as it employs sensors that receive electromagnetic fields from produced by lightning [28]. It consists of a sensor which has two looped antennae which are orthogonal to each other that detect the magnetic wave and a flat plate that detects the electric wave [28]. The data from these sensors can be used to calculate time and 3 dimensional location, latitude, longitude and altitude of the lightning flash with more than one sensor working together [9]. To alleviate the polarisation error, the sensor are calibrated to receiving waves that are from lightning return strokes.

Optical Imaging technique uses CCD imaging array to observe waveform that are within the near infrared and optical light spectrum range [29]. These sensors output data that can be used to calculate time, latitude and longitude of a lightning flash. Such sensors are usually space-based unlike MDF and TOA and they cannot be used to calculate the altitude of a flash [9] but they are calibrated to account for polarisation error [29].

2.4 Lightning Risk Assessment

The standard method of obtaining the lightning density assess lightning risk has been using flash data, which is an underestimate given that a flash may have multiple GSPs as seen on Figure 2.3, an image of a lightning flash with more than one stroke. The image was captured using a camera with low frame rate but a high resolution, the strokes contacts the ground at different points and not one as the flash data would suggest.

LLSs capture individual strokes but use frequency detection and triangulation technique to locate the strokes [6, 9, 24]. The data that they output comes with some uncertainty due to technical issues thus GSP analysis is performed to optimise the true

FIGURE 2.3: A low resolution image of a lightning flash with multiple strokes, the strokes contacts the ground at different point.

location of lightning GSPs. One method is using high speed cameras to verify the GSPs however it is not efficient, therefore other methods are necessary as solutions.

Chapter 3

Related Work

There are various researches that have studied and analysed lightning ground-strike points. These papers proposed and used various methods and algorithms to achieve their objective. A commonality in these papers are the features used for the research but they all used data from various sources. They all obtained various results from their chosen methodology.

A research conducted on the mechanisms of multiple lightning ground contacts used data from the U.S National Lightning Detection Network across multiple states [30] while another research on the characteristics of negative flashes with multiple GSPs made use of data from the Japanese Lightning Detection Network [31]. A research on validating GSP identification utilised data from ALDIS which is an Austrian LLS and Météorage which is a French LLS [32]. A research on LLSs made use of data from various sources around the world [24].

3.1 Feature Selection

The researches made use of similar features to study ground strike points but some of the features were specific and more suitable for the chosen methods. The research on LLSs used various features for their ground-strike point analysis. They used the number of reporting sensors, the time reported by sensors and the estimated distance from each sensor and the GSP. They also used peak current of the lightning discharge Nag et al. [24].

Research on the mechanisms of multiple lightning ground contacts proposed using Semi-Major Axes (SMA) of the lightning return strokes for GSP analysis [30]. For the analysis of the characteristics of negative flashes with multiple GSPs, various lightning were used characteristics as their chosen features [31]. Latitude, longitude and peak current of lightning strokes were the selected features. Research on validating GSP identification used time and direction of signal [32] as the chosen features.

3.2 Methods

Research on validating GSP identification used LLS techniques to geo-locate ground-strike points. The sensors used by ALDIS and Météorage combine time of arrival of each signal and the direction in which it came from. The signals from the sensors are then grouped according the corresponding values of the parameters. Then the grouped signals are put through a triangulation algorithm to located the point of contact of each lightning stroke [32].

Meteorage developed a GSP K-means clustering algorithm to assist with the verification of lightning detection [32]. The K-means algorithm is a two step process. The first step is to randomly pick k number of centres from a group of points, where k is a predetermined fixed number. The second step is find the nearest center to each of the other points in the cluster using Euclidean distances [33]. Once all the points are included in some clusters, a recalculation of the average of the clusters is done iteratively until an optimum is reached [33].

Research on the mechanisms of multiple lightning ground contacts proposed a then new method, *groupGCP* of which Campos later used for a thesis. The algorithm begins by sorting the strokes according to their SMA and a predetermined maximum value. It then applies the K-means algorithm to output strokes with SMAs lower than the maximum value by first obtaining the ground contacts points (GCPs) of each flash. The next step calculates the error ellipse scale of each output SMA with respect to the strokes with SMAs above the maximum value. The algorithm then sorts the SMAs into GCP groups according to their SMA rankings [30].

Research on the characteristics of negative flashes with multiple GSPs employed the use of JLDN in their methodology of GSP analysis. The network had 31 sensors in northern and southern parts of Japan by 2018. The geo-locations of the lightning strokes were recalculated using one of the new techniques at the time, Propagation Delay Correction (PDC), that had been use to improve locating algorithms that better select the sensor message of a return stroke [31].

The research on LLSs used various methods to validate ground-strike points in their paper. One of the methods was LLS self-reference which is statistical analysis of the parameters output by the LLS such as standard deviation of sensor timing error and semi major axis length of the 50% confidence ellipse. The second method uses data, such as M-component and peak current from tall object and rocket as ground truth data. The third method makes use of video camera data to validate ground strike points and the last method compares the performances of the LLSs against each other [24].

3.3 Evaluation

With the improvement made to the JLDN, characteristics of negative flashes with multiple GSPs research obtained a 3.5 arithmetic mean of number of strokes per flash, this was found to be consistent with another study that focused on the thunderstorms in Brazil [16]. The ratio of the peak current to that of the first return stroke was also larger in the New Ground Contact (NGC) at 0.82 than in the Pre-Existing Channel (PEC) at 0.74 [31].

The network of the model in the research was able to detect an increased number of strokes with the new technique with 50% of the flashes having multiple GSPs and the mean number of GSP per flash was 2.0. 21% of the points received multiple strokes and the mean number of strokes per GSP was found to be 1.6 [31].

The data that was obtained from Austria and France in 2012 and 2013 respectively on the research on validating GSP identification, consisted of 227 negative lightning flashes and 767 strokes. 221 flashes and 691 strokes we matched with lightning data which yielded DEs of 97% and 90% respectively. Out of 397 GSP, only 379 were matched with lightning data, that is a DE of about 95%. 636 strokes were classified as either NGC or PEC, bringing the discriminatory efficiency to 83% [32].

The research on the mechanisms of multiple lightning ground contacts used a data set that consisted of 25 negative flashes with 144 return strokes and their optimal parameters were input values of the validation algorithm, *groupGCP*. Using more than one of the parameters, the performance efficiency of the model ranged between 96.5% and 98.2% for individual return strokes and 88% and 92% for the full lightning flashes [30].

Another research on the GSP characteristics in negative downward lightning flashes applied the K-means algorithms used in the other researches to the same lightning data for model comparisons [10]. The results obtained from this research are summarised in Table 3.1. The first algorithm was the research on validating GSP identification, the second algorithm was from the research on mechanisms of multiple lightning contacts and the third algorithm was from characteristics of negative flashes with multiple GSPs research.

TABLE 3.1: A table of summarised results obtained using three algorithms used in the current literature [10].

Algorithm	NGC (%)	PEC (%)	All strokes (%)
Meteorage K-means	63.8 – 92.0	84.4 – 99.4	79.9 – 90.6
GroupGCP	64.6 – 95.3	87.6 – 99.4	80.3 – 94.4
Matsui K-means	98.3 – 99.1	63.8 – 79.3	83.1 – 89.7

The first algorithm was able to classify New Ground Contact lightning strokes with an accuracy range of 60.8% – 92.0%, while the second and third algorithms had accuracy ranges of 64.6% – 95.3% and 98.3% – 99.1% respectively. The Pre-Existing Contacts of lightning strokes were classified with an accuracy range of 63.8% – 79.3% by the third algorithm and the first and second algorithms had accuracy ranges of 84.4% – 99.4% and 87.6% – 99.4% respectively. Overall the three algorithms had accuracy range of 79.9% – 94.4%.

Chapter 4

Research Methodology

The current state-of-the-art makes use of K-means model to analyse and identify ground-strike points. As observed in the current literature, research on validating GSP identification employed a K-means algorithm to validate their ground-strike points data [32]. Research on mechanisms of multiple lightning ground contacts suggested a new approach to the K-means model and applied it to the U.S National Lightning Detection Network [30, 34]. The K-means models rely on a user defined distance threshold which bring about an uncertainty as it is unknown variable [30, 31, 32].

This research proposes using Bayesian Network (BN) to classify lightning ground-strike points (GSPs). Bayesian statistics have been proposed and applied to assess the performance of LLSs and to verify GSPs on other researches lightning locating systems [11, 12] and produced good results, therefore paving a way for BN to be used in this field. Unlike K-means models BN do not required a user defined distance threshold.

In this chapter the research design will be discussed in the first section and the data in the second section. The methods of the research are discussed in the third section, the analysis in the fourth section and the possible limitations of the research in the fifth section of the chapter. The final section discusses the significance of the research.

4.1 Research Design

For this research report, we propose using Bayesian networks for ground-strike point analysis. A Bayesian network is a type of a probabilistic directed acyclic graph (DAG) [35, 36]. A DAG is a graphical model that consists of nodes and directed edges, the nodes are variables and the directed edges serve as connections between adjacent nodes [36, 37]. The directed edges follow a path in which they do not cycle back to the initial node [37]. A Bayesian network is DAG that incorporates Bayesian statistics in its framework to achieve its objective [36].

Bayesian networks afford simplicity by being compact because the conditional probability of each variable is dependent on the joint probability of the parent variables, which are the variable directly linked to them [35]. The joint probability of Bayesian network is described by;

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i|px_i), \tag{4.1}$$

where x_i are variables and px_i are the parent variables [36, 38].

Figure 4.1 is diagram that depicts a Bayesian network with six variables that is not fully connected because not all the nodes in the DAG have edges between them. The joint probability if network is $P(x_1, x_2, x_3, x_4, x_5, x_6)$. Expanding joint probability by using the Bayesian theorem chain rule and then applying Equation 4.1 we find;

$$P(x_1, x_2, x_3, x_4, x_5, x_6) = P(x_1)P(x_2)P(x_6)P(x_3|x_1, x_2)P(x_4|x_3)P(x_5|x_3, x_6). \tag{4.2}$$

Learning of a Bayesian network aims to obtain a model that can be described the joint probability distribution of the variables using a data sample. It is an important step as it allows for density estimation which helps to make inference queries, specific predictions which is suitable for training models that can be used to predict values in future data, and knowledge discovery which helps with the analysis of the data [40]. Bayesian network models can be learned through different methods, structure learning and parameter learning.

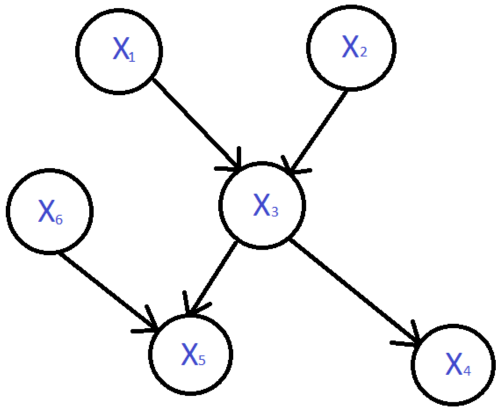


FIGURE 4.1: A Bayesian Network schematic diagram, adapted from Zhang and Poole [39]. x_1 and x_2 are parents of x_3, x_3 and x_6 are parents of x_5 and x_4 is a descendant of x_3.

Parameter learning works by estimating the parameters of a fixed Bayesian network model and a fully data set with the variables in the model. Parameter learning can be achieved by applying the Maximum Likelihood Estimation (MLE) or Bayesian statistics [40]. Inference is the reduction of a global probability distribution to a conditional probability of the observed variables [41]. Inference is used in Bayesian networks to make probabilistic queries of a variable given a certain event [40, 41]. It can also be used to fill in missing data [40].

Structure learning is a key factor to knowledge discovery, as it allows one to observe the correlation of the variables. There are three approaches one can do to perform structure learning from data. There is score-based structure learning, which employs statistics in its learning procedure. Each potential model is seen as a statistical problem, then a scoring function is used to fit each model to the data. The model with the highest score is the most suitable one [40].

Constraint-based structure learning assumes a Bayesian Network as a representation of variable independence. A conditional dependence and independence is run for all the variables in the data and then a model that is suitable is created [40]. Bayesian model averaging is a method that creates multiple models and combines the models to obtain the the average [40].

Bayesian Networks make use of various algorithm for structure learning such as Hill Climbing and Tabu search. Tabu search is a heuristic algorithm with the ability to optimise graphical models. Tabu search iteratively searches for the minimum of a given objective function [42]. The algorithm sets a random minimum from the set of variables of the objective function and continuously compares it to the next value until the minimum is found [42].

Bayesian Network,like other probabilistic models, rely on probabilistic model selection to identify a model that best fits the particular data by scoring the models [43]. There are various scoring functions such as Gaussian log-likelihood, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Gaussian log-likelihood is defined as [44]:

$$L_k = \sum_{i=1}^{n} \log f_k(x_i),$$

where by f_k is the Gaussian function:

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}.$$

AIC function estimates the variance between the data use to generate the model and the fitted potential model [45]. The function is defined as;

$$AIC = -2\ln L(\Theta_i|x) + 2k$$

where by L is the likelihood function and k is the dimension of the parameters of the model [45].

BIC is similar to AIC has in computes the variance between the data and the potential model but the BIC aims to transform the posterior probability of the model [46, 47]. The BIC scoring function is defined as:

$$BIC = -2 \ln L(\Theta_i | x_j) + 2k \ln{(n)}$$

where k is the number of model parameters, n is the number of data records and L is the the likelihood function which is the probability density function [46]. Θ_i are vectors of parameters in the model and x_j is the observed data [48].

Derivation of BIC

BIC was introduced for independent, identically distributed observations, and linear models [49], with the assumption that the likelihood is from the regular exponential family [40, 46]. The criterion selects a model by maximising the posterior probability of a potential model from a data set [46, 48, 49]. The posterior probability is described by:

$$P(M_i | x_j) = \frac{P(x_j | M_i) P(M_i)}{P(x_j)} \tag{4.3}$$

where $P(x_j)$ is the marginal probability distribution of the data and $P(M_i | x_j)$ is the marginal likelihood of the model [46, 48].

To maximise the the posterior probability of a potential model, Equation 4.3 is considered a continuous function given that there are unlimited number of potential model and thus it described as:

$$P(M_i | x_j) = \frac{P(M_i)}{P(x_j)} \int_{\Theta_i} L(\Theta_i | x_{j]}) f(\Theta_i | M_i) d\Theta_i \tag{4.4}$$

where the integral is the continuous function of the likelihood of a model given the vector parameters of the model and the the prior distribution of the parameters of the model as $L(\Theta_i | x_{j]})$ and $f(\Theta_i | M_i)$ respectively [46, 48].

From Equation 4.4

$$P(M_i|x_j) = \frac{P(M_i)}{P(x_j)} \int_{\Theta_i} L(\Theta_i|x_{j]})f(\Theta_i|M_i)d\Theta_i \tag{4.5}$$

$$-2\ln P(M_i|x_j) = 2\ln P(x_j) - 2\ln P(M_i) - 2\ln \int_{\Theta_i} L(\Theta_i|x_{j]})f(\Theta_i|M_i)d\Theta_i \tag{4.6}$$

$2\ln P(x_j)$ is a constant and thus cannot be maximised, $-2\ln P(M_i|x_j)$ is defined as the model selection $S(M_i|x_j)$ [46]. The first part of the integrand can be expanded into a Taylor expression,

$$\begin{aligned} \ln L(\Theta_i|x_{j]}) &\approx \ln L(\Theta_i|x_{j]}) + (\Theta_i - \Theta_i')\frac{\partial \ln L(\Theta_i'|x_{j]})}{\partial(\Theta_i} \\ &+\frac{1}{2}(\Theta_i - \Theta_i')\left[\frac{\partial^2 \ln L(\Theta_i'|x_{j]})^2}{\partial(\Theta_i\Theta_i')}\right](\Theta_i - \Theta_i') \\ L(\Theta_i|x_{j]}) &\approx L(\Theta_i|x_{j]}) \exp\frac{1}{2}(\Theta_i - \Theta_i')[I](\Theta_i - \Theta_i') \end{aligned}$$

where $I = \left[-\frac{1}{n}\frac{\partial^2 \ln L(\Theta_i'|x_{j]})^2}{\partial(\Theta_i\Theta_i')} \right]$ thus the integral becomes [46, 48]:

$$\begin{aligned} \int_{\Theta_i} L(\Theta_i|x_{j]})f(\Theta_i|M_i)d\Theta_i = L(\Theta_i'|x_j) \int_{\Theta_i} \exp\left(-\frac{1}{2}(\Theta_i - \Theta_i')[I](\Theta_i - \Theta_i')\right) \\ \times f(\Theta_i|M_i)d\Theta_i \end{aligned}$$

The integral on can be approximated by its symmetry to obtain:

$$\int_{\Theta_i} \exp\left(-\frac{1}{2}(\Theta_i - \Theta_i')[I](\Theta_i - \Theta_i')\right) f(\Theta_i|M_i)d\Theta_i = (2\pi)^{i/2}|I|^{-\frac{1}{2}} \tag{4.7}$$

and now replacing Equation 4.7 into the into Equation 4.6 to obtain [46, 48]:

$$\begin{aligned} S(M_i|x_j) &= -2\ln P(M_i) - 2\ln\left(L(\Theta_i'|x_j)(2\pi)^{i/2}|I|^{-\frac{1}{2}}\right) &(4.8) \\ &\approx -2\ln L(\Theta_i'|x_j) - 2k\ln n &(4.9) \end{aligned}$$

Bayesian Networks are advantageous since the are no assumptions that are needed to be made about the data prior to creating models and they are suitable for small data. The networks created can show which features have a direct effect on the target value and how the features are connected to each other from the data. The BN models can be used to predict outcomes of target by means of inference.

4.2 Data

4.2.1 Data Acquisition and Description

The data that is used was provided by Dr. Hugh Hunt and was recently used in a research about global GSP characteristics in negative downward lightning flashes which he contributed to [8]. The data that is used for this analysis was acquired in the city of Johannesburg, South Africa (SA) by SALDN between 2017 and 2019 [8]. The data was from SALDN consist of 15 features and had 1311 entries with no missing data. The 13 features described in Table 4.1 are directly measured by LLSs.

Another feature, Flash Number, is an added feature which is used to categorize the lightning strokes according to the IEC 62585. Strokes that have inter-stroke delay of 500 milliseconds and have a maximum distance of 10 kilometers between them are categorized as a single flash [6]. The ground truth data set is represented by the feature strike point. The ground truth data was obtained using high speed cameras which are able to capture occurring events to the millisecond to validate GSP locations of each strike point.

Figure 4.2 is an image representation of a lightning flash number 74 from the data over a map of Johannesburg. The red bold crosses on the map are GSP locations of strokes captured by the LLSs. From the figure, there are multiple strokes in the same region and a few scattered further from the cluster. Table 4.2 shows the raw data of flash 74. The fourth column, strike point, is the confirmed location of each stroke and is observed with a high speed camera.

High speed cameras are have been used to capture lighting strokes and lightning flashes for observation and to learn about the characteristics of various types of

TABLE 4.1: A table of descriptions of features of the data.

Feature	Description
Year	The year which the lightning stroke was captured by the LLSs.
Month	The month which the lightning stroke was captured by the LLSs.
Day	The day which the lightning stroke was captured by the LLSs.
Time	The time which the lightning stroke was captured by the LLSs in hours:minutes:seconds.
Latitude (Lat)	The latitude of the lightning GSP in degrees.
Longitude (Long)	The longitude of the lightning GSP in degrees.
Peak current (Peak_kA)	The peak current measured during a lightning stroke in kilo amperes
Chi-square	chi square measurement of the GSP location.
Semi-major axis (ell_smaj)	The distance from the GSP to the edge of the semi-major axis of the uncertainty ellipse in kilometres.
Semi-minor axis (ell_smin)	The distance from the GSP to the edge of the semi-minor axis of the uncertainty ellipse in kilometres.
Ellipse angle (ell_angle)	The orientation angle of the uncertainty angle.
Degrees of freedom (freedom)	The number of independent variables.
Number of detectors (num_dfrs)	The number of detectors that captured a particular lightning stroke.

flashes [50, 51, 52] such as in a research about high speed observation of lighting over Johannesburg, South Africa [53]. The images are can also be used to validate the locations of the GSPs and thus be used ground truth data.

Strokes that are at strike point 1 are enclosed in green triangles and those that are at strike point 2 are enclosed in the blue rectangle on Figure 4.2. The images of the two strike points captured by a high speed camera can be seen on Figure 4.3. Strike point 1 on 4.3a appears to be on on the left while strike point 2, 4.3b, is on the right and slightly farther than strike point 1. The time stamps on the two images show that the strokes occurred within a tenth of a second apart from each other and thus

TABLE 4.2: A table of Flash 74 data, strike point column that is highlighted is the confirmed locations of the strokes by high speed camera.

Year	Month	Day	Strike point	Time	Lat	Long	Peak_kA	chi_square	ell_smaj	ell_smin	ell_angle	freedom	num_dfrs
2017	11	14	1	20:5:13.54309	-26.207	28.081	-53	1.2	0.1	0.1	9.8	29	18
2017	11	14	1	20:5:13.56192	-26.206	28.080	-13	0.2	0.2	0.1	96.8	7	5
2017	11	14	1	20:5:13.57486	-26.207	28.080	-16	0.7	0.2	0.1	4.4	11	7
2017	11	14	1	20:5:13.58613	-26.243	28.109	-35	3.1	0.3	0.2	95.4	16	15
2017	11	14	1	20:5:13.60296	-26.206	28.080	-19	0.5	0.2	0.1	17.6	19	11
2017	11	14	2	20:5:13.64314	-26.230	28.095	-15	0.8	0.3	0.2	26.4	11	9
2017	11	14	1	20:5:13.68081	-26.206	28.081	-59	0.7	0.2	0.2	116.9	9	8
2017	11	14	1	20:5:13.73531	-26.206	28.080	-27	0.6	0.1	0.1	166.0	26	15
2017	11	14	2	20:5:13.76021	-26.219	28.092	-14	0.7	0.2	0.1	52.4	9	7
2017	11	14	2	20:5:13.81479	-26.219	28.095	-19	0.5	0.1	0.1	76.2	17	12
2017	11	14	1	20:5:13.85892	-26.205	28.079	-43	0.9	0.3	0.1	174.8	19	12
2017	11	14	1	20:5:13.917	-26.206	28.080	-12	0.1	0.1	0.1	109.5	14	9
2017	11	14	1	20:5:13.95434	-26.207	28.080	-12	0.3	0.3	0.1	0.1	11	8
2017	11	14	1	20:5:14.84261	-26.207	28.082	-7	0.6	0.5	0.2	102.7	2	3

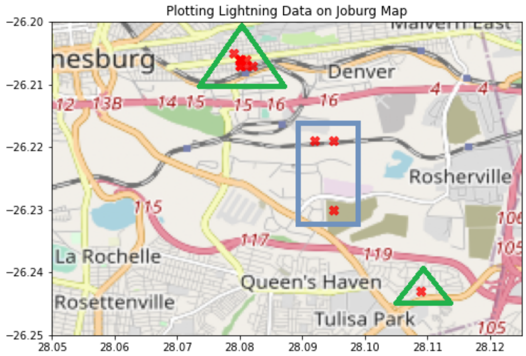


FIGURE 4.2: A picture of flash 74 plot on Johannesburg map. Strokes that are at strike point 1 are enclosed in green triangles and those that are at strike point 2 are enclosed in the blue rectangle.

they a part of the same lightning flash. 4.3c is a stacked image of the two images to show the GSPs of the lightning strokes.

(A) A low resolution image of lighting stroke 1 with its time stamp at the first GSP captured with high speed camera by Dr Carina Schumann

(B) A low resolution image of lightning stroke 2 with time stamp at the second GSP captured with high speed camera by Dr Carina Schumann

(C) A superimposed low resolution images of strokes 1 and 2 each at its GSP

FIGURE 4.3: Images of strike point 1 and 2 from flash 74 captured with a high speed camera by Dr Carina Schumann. Image 4.3a is of the strike point 1 and image 4.3b is of strike point 2. Strike point 1 appears to be on on the left while strike point 2 is on the right and slightly farther. Image 4.3c is a stacked image of images 4.3a and 4.3b to show the two GSPs [53]

4.2.2 Preprocessing

The semi-major and semi-minor distances were converted from km to degrees to make the variable units consistent. Time was also converted from hours: minutes: seconds to seconds so that the it is able to be read by the program as a float. The data set had 463 flashes whereby 213 of them are single-stroke and 250 of them are multi-stroke. Single-stroke flashes were excluded from the research as they do not have a PSP variable.

4.3 Methods

The aim of this analysis was to determine the GSP of stroke given the GSP of the previous stroke in flashes that have more than one stroke. To make such predictions, another feature was added, Previous Strike Point (PSP). First, the data was categorised according the flashes, then single stroke flashes were removed as they are not significant for the purpose of the project. The strokes labels in each flash were observed and used to determine the value of the PSP.

If a stroke label is the same as the previous one, PEC, then the PSP label for that entry is 1, NGC, if it is not the same then the PSP label is 0. The data has more strike points classified as NGC than PEC, this is alignment with the observation made in the research of global GSP characteristics in negative downward flashes [8], which then creates a class imbalance. It should also be noted that the initial stroke of each flash was removed since it does not have a PSP value.

BN-learn is an R package that is capable of Bayesian network modelling analysis by means of structure learning, parameter learning and inference [54]. The package contains algorithms for data pre-processing, inference, parameter learning and structure learning combining data and prior knowledge [54], that are necessary for Bayesian network modelling. It is capable of handling discrete Bayesian networks, Gaussian Bayesian networks and conditional linear Gaussian Bayesian networks using real data [54].

Five Bayesian network models were created using the data. For each model a k-fold cross validation was run 50 times on the data to obtain the mean classification errors and their standard deviations. This is done to evaluate how well the models are expected to perform. K-fold cross validation is a training method whereby the training data is divided into $k = 10$ samples, then $k - 1$ samples trained while the other sample is used as the testing sample. The algorithm is repeated until all the samples have been used as a testing sample then the average loss is calculated as the performance measure.

The models were learned using a score-based structure learning procedure with tabu search. The first model included all the variables and was learned from the raw data. The second model was learned with all the variables with the date information set as conditional dependencies of PSP. The third model excluded dates and times while the fourth model excluded dates, times and ellipse information. Both these models had number of sensors and degrees of freedom set as parent nodes of PSP. The last model was created using Time, Strike point, PSP, Latitude and Longitude and ellipse information.

4.3.1 Algorithm

The algorithm used to create the models using a structure learning procedure with tabu search algorithm and Bayesian Information Criterion. The full code can be accessed on Github repository: `https://github.com/Lwano31/BN_models`

1. Set up conditional dependencies from the data
2. Learn the best fitted DAG from data with dependencies
3. Do a cross validation on the DAG where; data = data,DAG,runs = 50
4. Calculate the loss with target = PSP
5. Predict the output from the fitted DAG
6. Plot DAG
7. Repeat algorithm for other models

4.4 Analysis

The performance of the models were evaluated using a confusion matrix which is suitable for the analysis of a binary classification model [55]. Figure 4.4 is a schematic of a confusion matrix with positive and negative values of the true and predicted classes. It computes the ratio of true predictions and false predictions by comparing the model results to the actual data. From a confusion matrix, other performance measures can be computed such accuracy, precision and recall [55]. The ground truth data can be used to validate the results from the Bayesian network using this analysis.

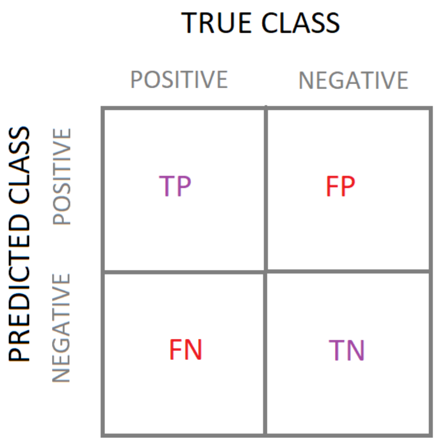


FIGURE 4.4: A schematic of a confusion matrix, adapted from Mohajon [56].

TP is the number of true positives, where the predicted value and the true value are both positive. FP is the number of false positives, where the predicted value is positive while the true value is negative. TN is the number of true negatives, where both the predicted value and the true value are negative. FN is the number of false negatives, where the predicted value is positive but the true value is negative.

Precision measures the model's ability to predict the positive labels correctly [55],

$$Precision = \frac{TP}{TP + FP}. \tag{4.10}$$

Recall measures the model's ability to obtain all positive labels [55],

$$Recall = \frac{TP}{TP + FN}. \tag{4.11}$$

F1-score measures the harmonic mean of precision and recall [55],

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{4.12}$$

Accuracy measures the percentage of accurately classified labels [55],

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.13}$$

Other performance measures that are used are balanced accuracy and kappa statistic. Kappa statistic is measure of agreement between the actual target and the predicted target [57] and a balanced accuracy is the arithmetic mean of the recall of each label [55], both these performance measures are able to take into account a class imbalance.

4.4.1 Baselines

- **Naive Bayes** classifier constructs a Bayesian probabilistic model that allocates a posterior class probability to a target value. It is based on the assumption that all the other variables are independent from each other but have a conditional dependence on the target value [58].
- **Multilayer (ML) perceptron** intakes multiple variables, assigns weights to them, sum them up to produce a binary output [59]
- **Logistic regression** is similar to *ML perceptron* but employs an activation function such as a sigmoid function to obtain an output [59].
- **Random tree** is a decision tree created by a learning algorithm that generates individual learners from random data sets [60]
- **Random forest** is a combination of tree classifiers whereby each tree depends on a random vector of which it was sampled independently and with the same distribution of all the trees in the forest [61].

- **Sequential Minimal Optimization** is an algorithm that employs Lagrangian multipliers to optimize Support Vector Machine (SMV), a hyperlane that separates binary classes [62].

4.5 Technical Contributions

This method introduces an algorithm that does not require a a user defined distance parameter, it relies on LLS data features. The method introduces the use of probability to classify GSPs. The method could possibly lead to the discovery of the most significant features of the data for lightning GSP classification from the conditional dependencies between the the features and target value.

4.6 Limitations

The methods has various limitations due to the tools used and the type of data. The data is a hybrid of discrete and continuous data, of which BN-learn has only one loss function that is suitable. BN-learn only has two algorithms for maxima and minima search. Another limitation of BN-learn is that discrete nodes can only be parent nodes of continuous nodes, even with prior knowledge of a continuous node being a parent of a discrete node, the library does not allow that.

Chapter 5

Results and Discussion

The aim of this research report was to classify lighting strokes in each flash by means of a Bayesian Network. The method is a different approach to the current methods found in the literature that use various algorithms of K-means. The objectives of the research are; to obtain lightning data, to select the relevant features, to categorise the data set into relevant lightning flashes according to the IEC 62858, to develop a Bayesian network model to fit the data and to analyse the performance of the model.

This chapter describes the results obtained from applying the methods and performance measures discussed in Chapter 4 in the first section. The second section of this chapter of the research discusses the results. The third section discusses the contribution of the research, the fourth section discusses the limitations and the fifth section summarises the results and discussion section.

5.1 Results

5.1.1 Directed Acyclic Graphs

Figure 5.1-Figure 5.5 are the Bayesian networks produced using a score-based structure learning procedure on bn-learn with tabu search with BIC scoring function that is suitable for data that hybrid of continuous and discrete. The green highlighted nodes are the PSP, which the target value and its parent node, and the edges from the parent nodes to the target node are highlighted in navy blue.

In the first model, Figure 5.1, the probability distribution of PSP has a conditional dependence on year. The second model, Figure 5.2, shows that PSP depends on year, month, day and number of sensors and third Bayesian network model, Figure 5.3, shows that in these the probability distribution of PSP has a conditional dependence on flash number and number of sensors. The fourth model, Figure 5.4, shows that on this model probability distribution of PSP has a conditional dependence on flash number and number of sensors. On the last model, Figure 5.5, PSP is a parent node and has no dependence on other features.

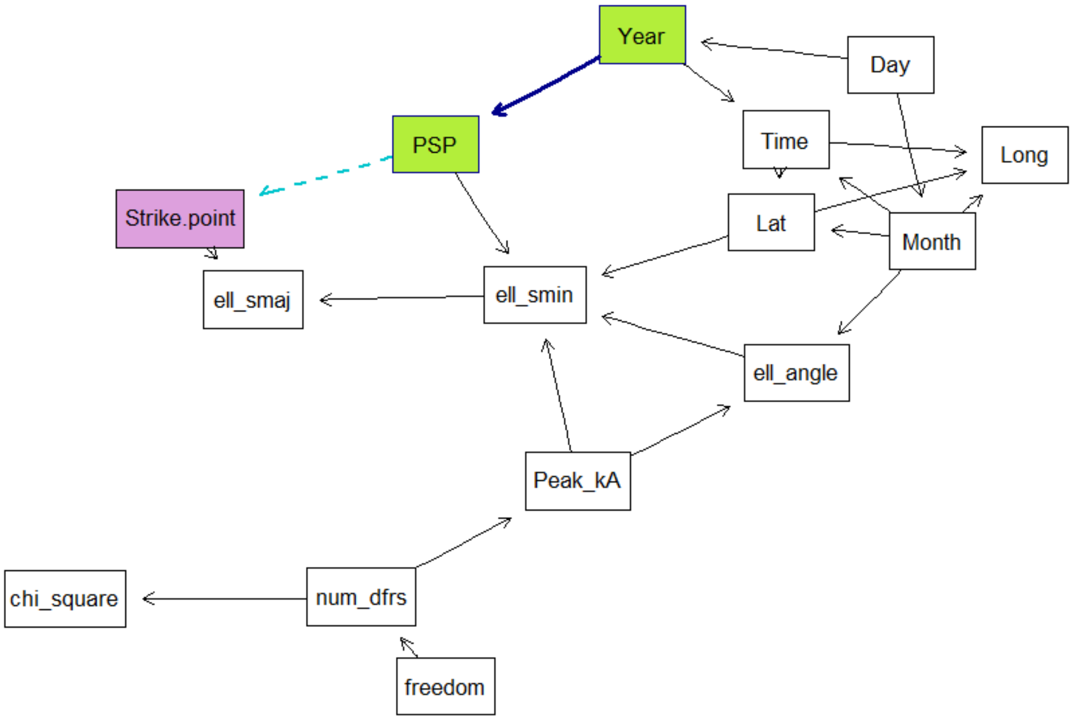


FIGURE 5.1: Model 1 Bayesian network diagram. The parent node of PSP is Year and strike point is a descendant of PSP

The first model, Figure 5.1, was learned with all the variables from the LLSs data without any conditional dependencies set. This resulted in the probability distribution of PSP has a conditional dependence on Year. The descendent node of PSP is the semi-minor axis of the uncertainty ellipse.

The second model, Figure 5.2, was learned with all the variables from LLSs. The model had set conditional dependencies of the target value. The model shows that PSP depends on year, month, day and number of sensors. The descendent node of

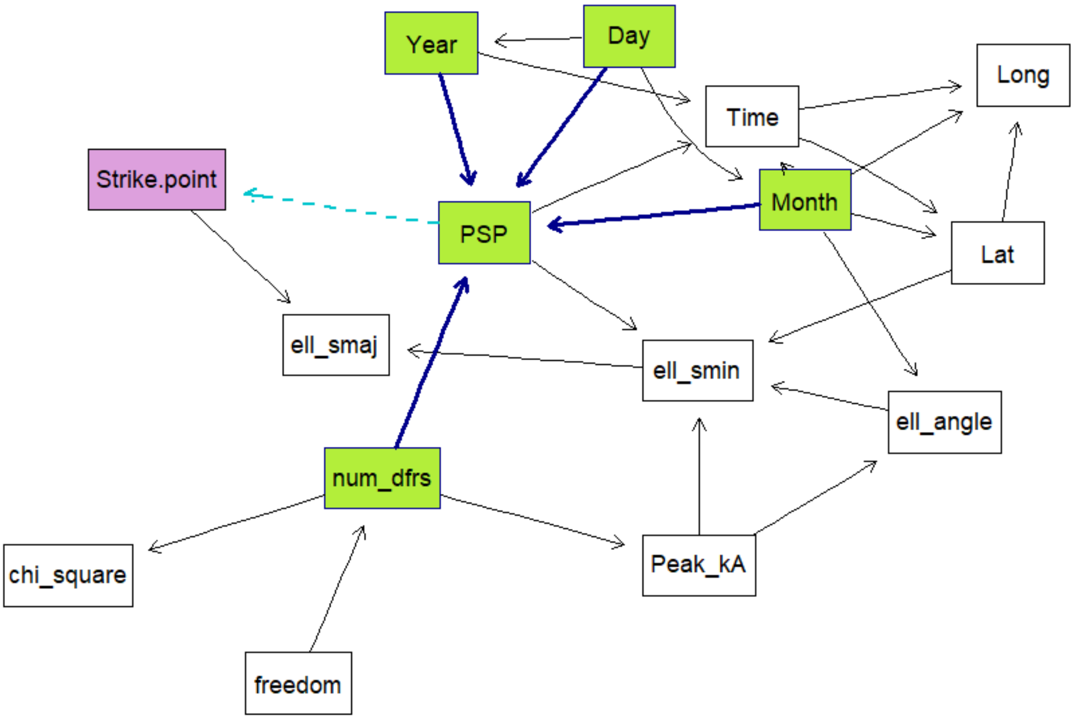


FIGURE 5.2: Model 2 Bayesian network diagram. The parent nodes of PSP are number of sensors,Year, Month and Day while strike point is a descendant of PSP.

PSP is the semi-minor axis of the uncertainty ellipse.

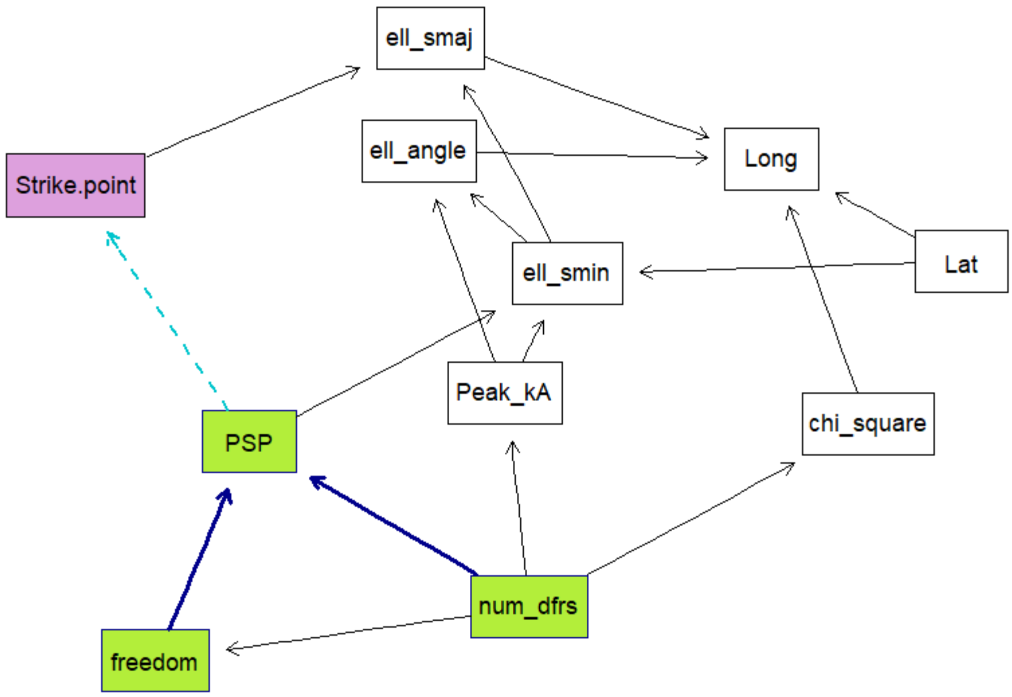


FIGURE 5.3: Model 3 Bayesian network diagram. The parent nodes of PSP are number of sensors and flash number, PSP is a parent node of strike point.

The third model, Figure 5.3, was learned with the exclusion of Year, Month, Day and Time features from the LLSs data. The BN model was learned with conditional dependencies set on the target value. The parent nodes of PSP are number of sensors and degrees of freedom. The descendent node of the target value is the semi-minor axis of the uncertainty ellipse.

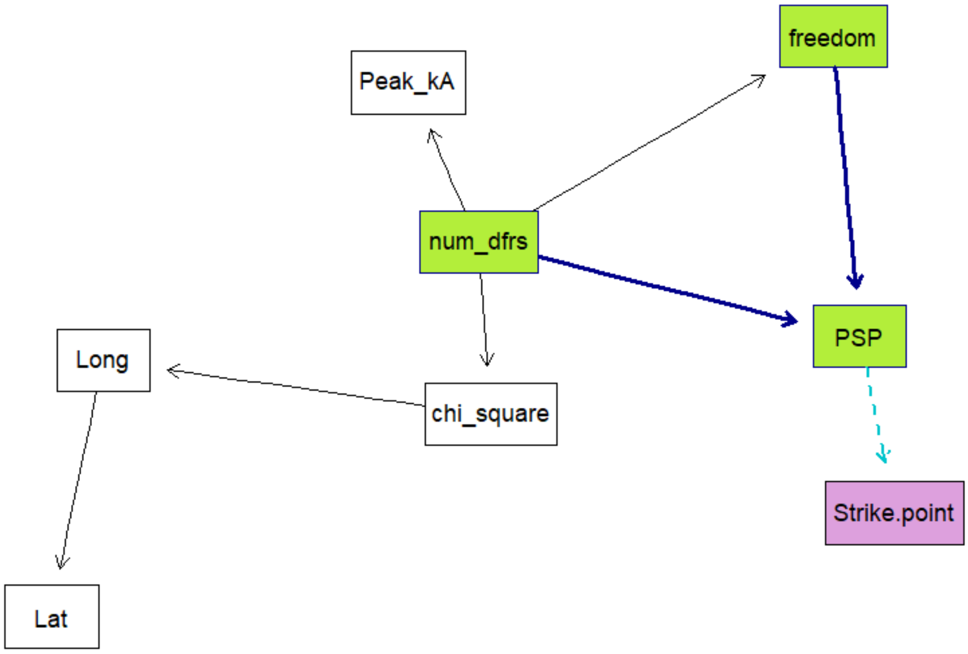


FIGURE 5.4: Model 4 Bayesian network diagram. The parent nodes of PSP are number of sensors and flash number, PSP is a parent node of strike point.

The fourth model, Figure 5.4, was learned with only a few selected LLSs data variables; latitude and longitude positions of the GSP, peak current of the lightning stroke, chi-square, degrees of freedom and number of sensors. The model was set with the target value having conditional dependencies on degrees of freedom and number of sensors.

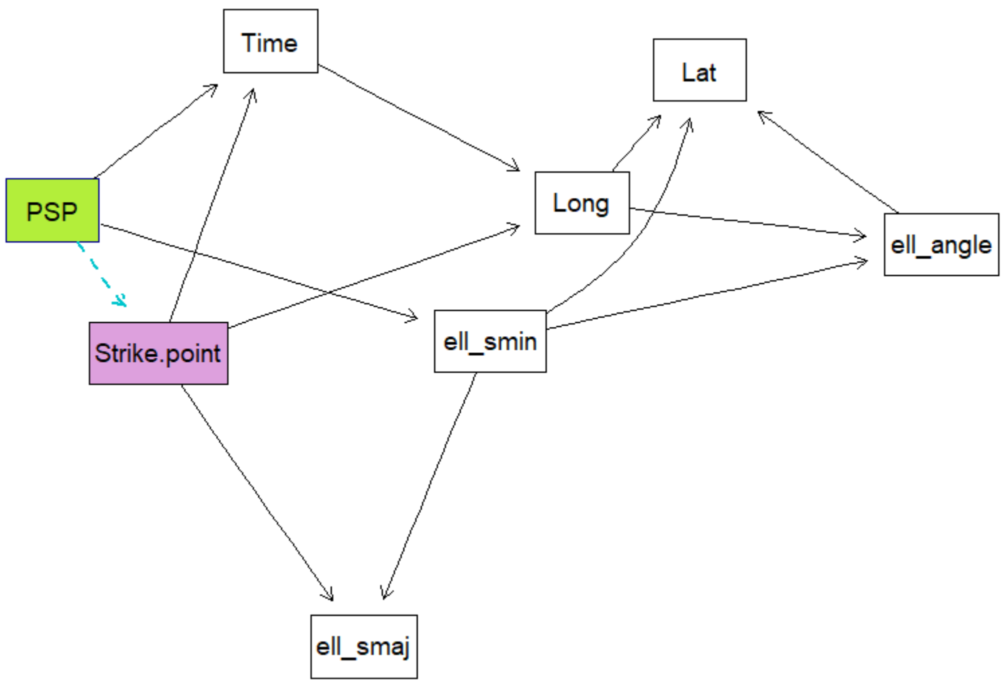


FIGURE 5.5: Model 5 network diagram. The parent node of strike point is PSP.

The fifth model, figure Figure 5.5, was learned with time, longitude, latitude, semi-major and semi-minor axes of the uncertainty ellipse and the angle of the uncertainty ellipse. The target value, PSP is a parent node and has no conditional dependence on other features. The descendants of PSP are time, longitude and the semi-major axis of the uncertainty ellipse.

5.1.2 Performance Measures

Table 5.1 summarises the overall performances of the models. Model 1 has an accuracy of 86.7%, its balanced accuracy is 73.6% and its kappa statistic is 0.53. Model 2 has an accuracy of 93.9%, a balanced accuracy of 88.0% and its kappa is 0.76. Model 3 has an accuracy of 90.5%, a balanced accuracy of 81.8% and its kappa is 0.64. Model 4 produced an accuracy of 90.8%, a balanced accuracy of 81.1% and kappa of 0.65. The last model output an accuracy of 88.8%, balanced accuracy of 77.3% and a kappa statistic of 0.57.

TABLE 5.1: Overall model performances with accuracy, balanced accuracy and kappa statistic

Model	Accuracy	Balanced Accuracy	Kappa
Model 1	86.7%	73.6%	0.53
Model 2	**93.9**%	**88.0**%	**0.76**
Model 3	90.5%	81.8%	0.64
Model 4	90.8%	81.1%	0.65
Model 5	88.8%	77.3%	0.57

Table 5.2 summarises how well the models able to classify the labels. Model 1 produced precision scores of 70.0% and 91.2%, recall scores of 51.5% and 95.8%, and F1-score of 59.3% and 93.4% for NGC and PEC respectively. The second model resulted with precision scores of 81.8% and 96.1%, recall scores of 79.4% and 96.6%, and F1-score of 80.6% and 96.4% for NGC and PEC respectively.

The third model has precision scores of 70.4% and 94.1% for NGC and PEC respectively, their recall scores are 69.1% and 94.5% and F1-scores are 69.9% and 94.3%. Model 4 output a precision score of 73.4% and 93.8%, a recall of 66.9% and 95.4%, and an F1-score of 70.0% and 94.6% NGC and PEC. The last model output a precision score of 66.7% and 92.6%, a recall of 60.3% and 94.2%, and an F1-score of 63.3% and 93.4% for NGC and PEC.

Figure 5.6 depicts the box plots of the classification errors of the k-fold cross validation of the five BN models over 50 runs. The classification errors of the first model reach a maximum of 0.126, a minimum of 0.108 and a median of 0.118. There an outlier below the minimum at 0.106. The lower and upper quartile of the classification errors are 0.114 and 0.12 respectively.

Table 5.3 summarises the K-fold cross validations of the data on each model and the mean classification errors and standard deviations were obtained, these were done with k value of 10 and 50 runs. The first model has a mean error of 0.12 and a standard deviation of 0.0042. Model 2 has a mean error of 0.15 and a standard deviation of 0.0051. Model 3 has mean error of 0.13 with a standard deviation of 0.0045. The

TABLE 5.2: Predictive performance of target value for each model with precision, recall and f1-scores

Model	Precision	Recall	F1-score
Model 1 NGC	70.0%	51.5%	59.3%
Model 1 PEC	91.2%	95.8%	93.4%
Model 2 NGC	81.8%	79.4%	80.6%
Model 2 PEC	96.1%	96.6%	96.4%
Model 3 NGC	70.7%	69.1%	69.9%
Model 3 PEC	94.1%	94.5%	94.3%
Model 4 NGC	73.4%	66.9%	70.0%
Model 4 PEC	93.8%	95.4%	94.6%
Model 5 NGC	66.7%	60.3%	63.3%
Model 5 PEC	92.6%	94.2%	93.4%

fourth model has a mean error of 0.13 with a standard deviation of 0.0046. The last model has a mean of 0.12 and a standard deviation of 0.0032.

TABLE 5.3: K-fold cross validation results, with a mean classification errors of the target value and their standard deviations.

Model	Mean Classification error	Standard deviation
Model 1	0.1174	0.0042
Model 2	0.1494	0.0051
Model 3	0.1294	0.0045
Model 4	0.1293	0.0046
Model 5	0.1248	0.0032

The classification errors of the second model reach a maximum of 0.159 with two outliers above the maximum at 0.162 and 0.163. They have a minimum of 0.138 and their median is 0.149. The lower and upper quartile are 0.146 and 0.152 respectively. The classification errors of the third model reach a maximum of 0.151 and a minimum of 0.122. Their median is at 0.129, the lower quartile is 0.127 and upper quartile is 0.134.

The classification errors of the fourth model have a minimum of 0.119, a median of 0.130 and the maximum of 0.140. The lower quartile is 0.127 and the upper quartile is 0.134. The lower quartile The last model output classification errors that have a minimum of 0.118, a median of 0.124 and a maximum of 0.132. The lower quartile is 0.123 and the upper quartile is 0.128.

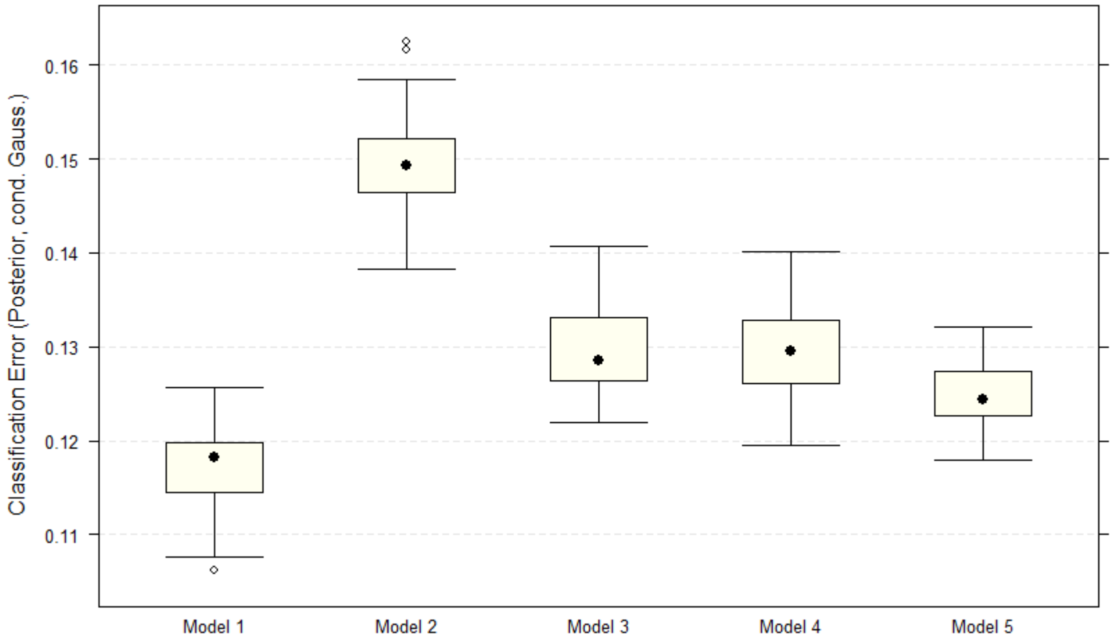


FIGURE 5.6: Classification error bar with confidence intervals

5.2 Discussion

This section discusses the results obtained from the five BN models and their implications. These results will also be compared to the results obtained in global GSP characteristics in negative downward lightning flashes research from the K-means models that are used in current literature [10]. The performance of the BN models will be compared with the performance of a variety of other classifier models that are used as baselines in this research.

5.2.1 BN models Discussions

Model 1 in Figure 5.1 which was learned using all the variables and it shows that the target value depends only on the year. The model produced the lowest accuracy and the lowest kappa index. This could be an indication that Year as the only parent node of the target value may not be sufficient. The model was able to predict PEC strokes with more precision compared to NGC strokes.

This model produced lowest mean classification error during k-fold cross validation. The model classification error is slightly skewed downwards, this implies that the classification error is most likely to be less than 0.118, which is confirmed by the mean classification error at 0.117. The classification error has a low inter-quartile range and standard deviation, this means the mean classification error lies within a narrow region.

Model 2 in Figure 5.2 was learned using all the variable. The conditional dependence of the target value were set to be date information and number of sensors. This model produced the highest accuracy of 93.9% and a kappa statistic of 0.76 even though it had the highest mean classification error. This model predicted strokes in the same GSP as the previous strokes with a precision of 96.1% and strokes at the same GSP with a precision of 81.8% which is the highest.

The model classification error is the highest of all models but it is not skewed, the error is evenly spread around the mean which is also the median. The classification error has a highest standard deviation, which means it has a wide spread around the mean. This implies that the mean classification error has a higher uncertainty but this is due to the outliers above the maximum. The inter-quartile range is 0.021, which infers that the mean classification error is most like to be between a narrower region and its uncertainty is lower than what the standard deviation suggests.

Model 3 in Figure 5.3 was learned with the exclusion of date and time. The Bayesian network shows that the target value has degrees of freedom and number of sensors as its parents. The model produced the third highest accuracy and kappa statistic. This implies that the exclusion of time and date reduces the model's ability to identify GSPs particularly because from the previous model, date and time were parent nodes of the target value.

The median of the classification error is the third highest but the data skewed upwards, which implies that the mean classification error is most likely to be higher than the median which can be observed on Table 5.3. The model also has the highest inter-quartile range which suggest the spread around the mean is wide thus the uncertainty of the mean classification error is high.

Model 4 in Figure 5.4 excluded date, time and the ellipse information. The model has the target value with degrees of freedom and number of sensors as its parents nodes just as the third model. The model produced results that are slightly higher than in the previous model, making it the second best performing model by a small margin. This implies that ellipse information does not affect the target value which is evident given that the ellipse information are not parent nodes of target value.

The classification error of the fourth model slightly different from the third model's classification error. The median is slightly higher but the data is less skewed and also skewed downwards which implies that the mean is lower than the median, this observed in Table 5.3. The model also has a the highest inter-quartile range just as the third model, implying that the uncertainty around the mean classification error is high.

Model 5, Figure 5.5 was learned using only the ellipse information and time. The target value is parent node has multiple dependencies, this is due to the architecture of bn-learn library as it does not allow continuous data to be parent nodes of discrete data. This model produced the lowest accuracy and kappa statistic even though it produced a mean classification error that almost he same as the two previous model.

The classification error of the fifth model has the smallest range and inter-quartile range, which implies that the data is spread across a narrow region which is evident from the standard deviation of the mean classification error in Table 5.3. The data is also slightly skewed upwards which indicates that the mean classification error is most likely to be above the median of which it can observed in Table 5.3.

Although model 1 was expected to perform better than all the other models due to the mean classification error being the lowest at 0.1174, the model 2 outperformed the other models even with having the highest mean classification error. This is evidence that BN works better with set conditional dependencies, even though BNs are capable of finding conditional dependencies of the features, some prior knowledge of how the features relate to each other creates a better working model.

All the models were affected slightly negatively by the class imbalance as the results show that the precision, recall and f1-scores were lower for strokes with different GSPs than their previous strokes. It also reflected by the balanced accuracies, which are lower than the accuracies. The kappa statistics for all the models is high, which shows that there is high agreement between the actual target values and the predicted targets.

5.2.2 Baselines Discussions

The data was trained using other Machine Learning algorithms with k-fold cross validation. The algorithms are all obtained on Weka, a java software for machine learning algorithms [63]. The algorithms; Naive Bayes, Logistic regression, Multilayer perceptron, Random trees, Random forest and Specific Minimal Optimization (SMO) were compared with BN results as shown on Table 5.4.

BN produced better predictive performances than all the algorithms. All the algorithms show a similarity with lower performance when identifying NGCs. This may be due to the class imbalance of which BN handled it better given that it produced significantly higher results. Naive Bayes and SMO produced the second best results in identifying the GSPs while Random tree had the poorest results of all the algorithms.

TABLE 5.4: Predictive performance of target value using other machine learning algorithms as baselines with precision, recall and f1-scores

Classifier	Precision	Recall	F1-score
Bayesian Network NGC	81.8%	79.4%	80.6%
Bayesian Network PEC	96.1%	96.6%	96.4%
Naive Bayes NGC	61.6%	56.6%	59.0%
Naive Bayes PEC	91.8%	93.3%	92.5%
Logistic NGC	54.1%	44.1%	48.6%
Logistic PEC	89.7%	92.8%	91.2%
ML Perceptron NGC	58.6%	47.8%	52.6%
ML Perceptron PEC	90.4%	93.5%	91.9%
Random forest NGC	50.8%	22.8%	31.5%
Random forest PEC	86.7%	95.8%	91.0%
Random tree NGC	43.0%	40.4%	41.7%
Random tree PEC	88.8%	89.7%	89.2%
SMO NGC	61.6%	56.6%	59.0%
SMO PEC	91.8%	93.3%	92.5%

TABLE 5.5: Overall model performances using other machine learning algorithms as baselines with accuracy and kappa statistic

Classifier	Accuracy	Kappa
Bayesian Network	93.9%	0.76
Naive Bayes	87.4%	0.52
Logistic regression	85.0%	0.40
ML Perceptron	86.2%	0.45
Random forest	84.1%	0.24
Random tree	81.8%	0.31
SMO	87.4%	0.52

The model accuracies for the algorithms and their kappa statistic are shown on Table 5.5. Even though the all the algorithms produced high accuracies, BN output the highest model accuracy. BN also had the highest kappa statistic, with Naive Bayes and SMO having the second best kappa statistic. This means that there is a moderate agreement between the actual target values and predicted target values from the Naive Bayes and SMO algorithms while BN has a substantial agreement between the actual and predicted target values.

The ML perceptron also had a moderate agreement between the actual target values and the predicted target values. Logistic regression, random forest and random tree had the lowest model accuracies and the agreement between the actual target values and predicted target values is fair, which is the lowest strength of agreement from the algorithm.

5.2.3 K-means and BN Discussion

Table 5.6 summarises the results obtained in global GSP characteristics in negative downward lightning flashes research [10] using the three K-means algorithms used in the current literature; Matsui K-means, GroupGCP and Meteorage K-means algorithms that are described in Chapter 4 and the results obtained from the BN models. The table summarises the performances of the models in identifying NGC strokes, PEC strokes and overall model performances.

TABLE 5.6: Summarised results from global GSP characteristics in negative downward lightning flashes research [10] using three algorithms used in the current literature combined with the results from BN models

Algorithm	NGC (%)	PEC (%)	All strokes (%)
Meteorage K-means	63.8 – 92.0	84.4 – 99.4	79.9 – 90.6
GroupGCP	64.6 – 95.3	87.6 – 99.4	80.3 – 94.4
Matsui K-means	98.3 – 99.1	63.8 – 79.3	83.1 – 89.7
Bayesian Network	66.7 – 73.4	91.2 – 96.1	86.7 – 93.9

The K-means algorithms rely on a predetermined distance threshold and other various conditions to achieve GSP analyses. For the purpose of their investigation, global GSP characteristics in negative downward lightning flashes research created models whereby the distance threshold was used a control variable [10]. Six distances were selected as the threshold; 0.2 km, 0.5 km, 1 km, 2 km, 5 km and 10 km, all of which are within the distance that the IEC 62585 has set as the maximum distance between strokes that belong to the same flash.

The K-means algorithms designed at Meteorage is able to identify NGC strokes at the range between 63.8% and 92.0%, and PEC strokes at a range between 84.4% and 99.4%. This implies that the model is capable of identifying PEC strokes accurately better than it identifies NGC strokes. The model is able to identify 90.6% of all the strokes accurately at most.

The GroupGCP algorithm is able to identify NGC strokes at the range between 64.6% and 95.3%, and PEC strokes at a range between 87.6% and 99.4%. This implies that the model is identifies the PEC strokes better than it is capable of identifying NGC strokes accurately like the Meteorage algorithm. The model is able to identify all 94.4% of all the strokes accurately, which is better than Meteorage algorithm.

The K-means algorithms designed for the research about characteristics of negative flashes with multiple GSPs [31] is able to identify NGC strokes at the range between 98.3% and 99.1%, and PEC strokes at a range between 63.8% and 79.3%. This implies that this model is better at identifying NGC strokes than than PEC strokes which is very different from the GroupGCP and Meteorage models. The model is capable of identifying 89.7% of all the strokes accurately, which is the lowest of all the K-means algorithms.

BN model is able to identify NGC strokes at a range between 66.7% and 73.4%, and PEC strokes at range between 91.2% and 96.1%. This implies that BN models perform better when identifying PEC strokes than NGC strokes, similar to Meteorage and GroupGCP algorithms. The model is able to identify 93.9% of all the strokes accurately at most.

Comparing the results in global GSP characteristics in negative downward lightning flashes research [10] with the BN model results, it can be observed that the range of accuracies of the BN models are within the range of all three algorithms, GroupGCP has the best overall performance. The BN models identified NGC strokes at a range of accuracies that were within the range of the Meteorage K-means and GroupGCP algorithms but below the Matsui K-means. The K-means algorithms are outperformed the BN models when identifying NGC strokes.

The accuracy range of the identification of strokes at the same GSP from the BN models are within range of Meteorage K-means and GroupGCP algorithm and above than the range of Matsui K-means algorithm. GroupGCP and Meteorage algorithms are the best at identifying strokes that are at the same GSP as the previous strokes.

The use of BN for GSP analyses proves to not improve the lightning stroke location validation, the K-means methods remain the best method thus far. Although the K-means method rely on pre-determined distance parameter and other parameters that are specific to each algorithm, they are easy to implement and produce high accuracies.

5.3 Contributions

Bayesian Networks provide an algorithm that does not rely on a predetermined parameter, distance, such as in the K-means algorithms. It relies on the data features obtained from the LLS network. DAGs from BN models allow for one to visualise how the data features are dependent or independent from each other and with the model learning processes and inference one can deduce which features are most significant given the conditional probabilities.

5.4 Limitations

BN-learn is a useful, particularly for obtaining DAG diagrams but it has some limitations. BN-learn has only two score-based learning algorithms and four score functions for data with continuous and discrete features, which limits options for this research. Continuous features could not be set to be parents of discrete features due to the BN-learn. This impacted on negatively the model learning as it limits how the models can be configured.

5.5 Summary

The lightning data was obtained supplied by co-supervisor, Dr. Hunt. All the features that were from the LLS data were used for this research. Flash Number parameter, which is determined from Time, was used to categorize the data into flash groups according to the IEC 62858 but was excluded in the model learning processes as it is not data feature of the LLS. The data was categorised. BN models were learned through a structure learning procedure and they were analysed using a confusion matrix and a kappa statistic.

There were five BN models created, one of which was learned from raw data and the rest were learned by setting conditional dependencies on the target value, PSP. All the models output good results with accuracy range of 86.7% − 93.9%. The highest accuracy is slightly below the accuracy obtained with GroupGCP algorithm and higher the Matsui and Meteorage algorithms.

Six baselines were created as comparisons of the BN models and they produced good results but not better than BN. BN proved to be more capable of handling imbalanced data as it output a high kappa statistic while the baseline algorithms yield a poor kappa index with the exception of naive Bayes and SMO algorithms.

Chapter 6

Conclusions and Future Work

The purpose of this research report was to introduce new application of BN to the analysis of GSPs that has not been done yet to validate the locations of lightning GSPs obtained by LLSs. A Bayesian network is a probabilistic directed acyclic graph and is able to learn the joint probability distribution of random variables from a data set which can be used to predict the GSPs of each lightning stroke.

The objectives of the research were; to obtain lightning data, to select the relevant features, to categorise the data set into relevant lightning flashes according to the IEC 62858, to develop a Bayesian network model to fit the data and to analyse the performance of the model. The objectives of the research were achieved and the results were obtained and evaluated.

The lightning data was supplied by co-supervisor, Dr. Hunt. All the features that were from the LLS data were used for this research. Flash Number parameter, which is determined from Time, was used to categorize the data into flash groups according to the IEC 62858 but was excluded in the model learning processes as it is not data feature of the LLS. The data was categorised. BN models were learned through a structure learning procedure and they were analysed using a confusion matrix and a kappa statistic.

Five Bayesian Network models were created using various combinations of the variables. The first model was learned from the raw data and the other models were learned from the data with assumptions about the joint probabilities. It was expected that model 1 would outperform the other models given that it output the lowest mean classification error during cross validation but model 2 produced the best results with an accuracy of 93.9% and a kappa index of 0.73.

There were six algorithms used as baselines for the BN models; Naive Bayes, logistic regression, multi-layer perceptron, random tree, random forest and sequential minimal optimization. They all produced good results but BN outperformed them. The baselines and BN models were affected by the class imbalance of the data set as the the minor class had significantly lower prediction results than the major class.

The K-means algorithms used global GSP characteristics in negative downward lightning flashes research [10] performed better than the Bayesian network models. The highest accuracy of the Bayesian network models was slightly lower than GroupGCP but higher than Meteorage and Matsui K-means. The other ML algorithms produced good results but still lower than BN. This implies that BN is good alternative to K-means algorithms.

The results from this research and the global GSP characteristics in negative downward lightning flashes research [10] show that the BN models do not improve the performance of GSP analyses. The K-means method remain the best methods for GSP analyses. Although they rely on pre determined parameters, they are still easy to implement.

6.1 Future Work

This analysis can be improved by using a distance threshold to determine the target value as used in the K-means algorithms. Even though k-fold cross validation works well for imbalanced data, the problem could be alleviated by applying a SMOTE technique during the pre-processing of data to make the minor class the same amount as the major class.

Making use of other programming tools and libraries with Bayesian Network capabilities could be an interesting investigation, particularly to compare the results with the ones obtained in this research. Another investigation that could be done in future, would be to use the obtained DAGs from the BN models to learn the parameters of other lightning data from various LLSs such as the data used in global

GSP characteristics in negative downward lightning flashes research from Brazil, Austria, France, Spain and United States of America [10].

Appendix A

Appendix Title

A.1 Code Sample

```
net = tabu(flash_data, score = 'bic-g')
graphviz.plot(net,shape = 'rectangle',
layout = 'fdp',highlight =
list(nodes = c("PSP", parents(net, "PSP")),
arcs = incoming.arcs(net, "PSP"),
col = "darkblue", fill = "tomato", lwd = 3))

bn.cv(data=flash_data,net,runs=50
loss.args = list(target = 'PSP'))

predicted = predict(bn.fit(net,flash_data), "PSP", flash_data)
```

A.2 Assumptions and Definitions

Leader An initial propagation of charges from one region to a region of opposite polarity [13].

Streamer Secondary propagation of charges induced by a nearby leader [13].

Stroke The connection between a leader and a streamer that creates a channel for electron to flow [13].

Ground-Strike Point (GSP) A point on the ground where a stroke makes a connection [13].

Flash One or multiple strokes that occur within a short period of time in the same area [16, 17].

Cloud-to-ground (CG) A leader that connect from the clouds to the ground [1, 13].

Negative CG A CG with negatively charged leaders [1, 13].

Positive CG A CG with positively charged leaders [1].

Multiplicity The number of strokes per flash [21].

Bibliography

[1] Sanketa Shivalli. “Lightning Phenomenon, Effects and Protection of Structures from Lightning”. In: *IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE) Volume* 11 (2016), pp. 2278–1676.

[2] RD Hill. “A survey of lightning energy estimates”. In: *Reviews of Geophysics* 17.1 (1979), pp. 155–164.

[3] Robert Lichtenberg et al. “Cardiovascular effects of lightning strikes”. In: *Journal of the American College of Cardiology* 21.2 (1993), pp. 531–536.

[4] JP Parmantier, F Issac, and V Gobin. “Indirect effects of lightning on aircraft and rotorcraft”. In: *AerospaceLab* 5 (2012), p–1.

[5] Irshad Ullah et al. “Smart lightning detection system for smart-city infrastructure using artificial neural network”. In: *Wireless Personal Communications* 106.4 (2019), pp. 1743–1766.

[6] Stéphane Pédeboy. “Introduction to the IEC 62858: Lightning density based on Lightning Locating Systems”. In: *International Lightning Protection Symposium*. 2018.

[7] Brandon J Vogt and Stephen J Hodanish. “A high-resolution lightning map of the state of Colorado”. In: *Monthly Weather Review* 142.7 (2014), pp. 2353–2360.

[8] Dieter R Poelman et al. “Global ground strike point characteristics in negative downward lightning flashes–Part 1: Observations”. In: *Natural Hazards and Earth System Sciences* 21.6 (2021), pp. 1909–1919.

[9] Kenneth L Cummins and Martin J Murphy. “An overview of lightning locating systems: History, techniques, and data uses, with an in-depth look at the US NLDN”. In: *IEEE transactions on electromagnetic compatibility* 51.3 (2009), pp. 499–518.

[10] Dieter R Poelman et al. "Global ground strike point characteristics in negative downward lightning flashes–Part 2: Algorithm validation". In: *Natural Hazards and Earth System Sciences* 21.6 (2021), pp. 1921–1933.

[11] Hugh GP Hunt. "Lightning Location System Detections as Evidence: A Unique Bayesian Framework". In: *IEEE Transactions on Geoscience and Remote Sensing* 59.3 (2020), pp. 1848–1858.

[12] Phillip M Bitzer, Jeffrey C Burchfield, and Hugh J Christian. "A Bayesian approach to assess the performance of lightning detection systems". In: *Journal of Atmospheric and Oceanic Technology* 33.3 (2016), pp. 563–578.

[13] Łukasz Staszewski. "Lightning Phenomenon–Introduction and Basic Information to Understand the Power of Nature". In: *International Conference Environment and Electrical Engineering*. 2010.

[14] Clive Saunders. "Charge separation mechanisms in clouds". In: *Planetary Atmospheric Electricity* (2008), pp. 335–353.

[15] Vladislav Mazur. "Physical processes during development of lightning flashes". In: *Comptes Rendus Physique* 3.10 (2002), pp. 1393–1409.

[16] MMF Saba, MG Ballarotti, and O Pinto Jr. "Negative cloud-to-ground lightning properties from high-speed video observations". In: *Journal of Geophysical Research: Atmospheres* 111.D3 (2006).

[17] FJ De Miranda, O Pinto Jr, and MMF Saba. "A study of the time interval between return strokes and K-changes of negative cloud-to-ground lightning flashes in Brazil". In: *Journal of Atmospheric and Solar-Terrestrial Physics* 65.3 (2003), pp. 293–297.

[18] Caixia Wang et al. "Characteristics of downward leaders in a cloud-to-ground lightning strike on a lightning rod". In: *Atmospheric Research* 203 (2018), pp. 246–253.

[19] Vernon Cooray and KPSC Jayaratne. "Characteristics of lightning flashes observed in Sri Lanka in the tropics". In: *Journal of Geophysical Research: Atmospheres* 99.D10 (1994), pp. 21051–21056.

[20] MG Ballarotti, MMF Saba, and O Pinto Jr. "High-speed camera observations of negative ground flashes on a millisecond-scale". In: *Geophysical Research Letters* 32.23 (2005).

[21] Dieter R Poelman et al. “Global ground strike point characteristics in negative downward lightning flashes–part 1: Observations”. In: *Natural Hazards and Earth System Sciences Discussions* (2021), pp. 1–17.

[22] William C Valine and E Philip Krider. “Statistics and characteristics of cloud-to-ground lightning with multiple ground contacts”. In: *Journal of Geophysical Research: Atmospheres* 107.D20 (2002), AAC–8.

[23] Matthew A Pasek and Marc Hurst. “A fossilized energy distribution of lightning”. In: *Scientific reports* 6.1 (2016), pp. 1–9.

[24] Amitabh Nag et al. “Lightning locating systems: Insights on characteristics and validation techniques”. In: *Earth and Space Science* 2.4 (2015), pp. 65–93.

[25] VA Rakov. “Lightning electromagnetic environment: From continuing-current fields to X-rays”. In: *Intl. Conf. on Grounding and Earthing*. 2008.

[26] Joseph R Dwyer and David M Smith. “A comparison between Monte Carlo simulations of runaway breakdown and terrestrial gamma-ray flash observations”. In: *Geophysical Research Letters* 32.22 (2005).

[27] EA Lewis, RB Harvey, and JE Rasmussen. *Hyperbolic direction finding with sferics of transatlantic origin*. Tech. rep. AIR FORCE CAMBRIDGE RESEARCH LABS HANSCOM AFB MA, 1962.

[28] EP Krider et al. “Lightning direction-finding systems for forest fire detection”. In: *Bulletin of the American Meteorological Society* 61.9 (1980), pp. 980–986.

[29] DJ Boccippio et al. “The Optical Transient Detector (OTD): Instrument characteristics and cross-sensor validation”. In: *Journal of Atmospheric and Oceanic Technology* 17.4 (2000), pp. 441–458.

[30] LZS Campos. “On the mechanisms that lead to multiple ground contacts in lightning”. PhD thesis. Ph. D. Thesis, Instituto Nacional de PesquisasEspaciais INPE, Brazil, 2016.

[31] Michihiro Matsui, Koji Michishita, and Shigeru Yokoyama. “Characteristics of Negative Flashes With Multiple Ground Strike Points Located by the Japanese Lightning Detection Network”. In: *IEEE Transactions on Electromagnetic Compatibility* 61.3 (2019), pp. 751–758.

[32] Stéphane Pédeboy and Wolfgang Schulz. "Validation of a ground strike point identification algorithm based on ground truth data". In: *Proceedings, International Lightning Detection Conference ILDC*. 2014.

[33] Shi Na, Liu Xumin, and Guan Yong. "Research on k-means clustering algorithm: An improved k-means clustering algorithm". In: *2010 Third International Symposium on intelligent information technology and security informatics*. Ieee. 2010, pp. 63–67.

[34] LZ Campos, KL Cummins, and O Pinto. "An algorithm for identifying ground strike points from return stroke data provided by Lightning Location Systems". In: *Asia-Pacific Conference on Lightning (APL)*. 2015, pp. 475–478.

[35] Judea Pearl. "Bayesian networks". In: (2011).

[36] Todd Andrew Stephenson. *An introduction to Bayesian network theory and usage*. Tech. rep. IDIAP, 2000.

[37] Tyler J VanderWeele and James M Robins. "Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect". In: *American journal of epidemiology* 166.9 (2007), pp. 1096–1104.

[38] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[39] Nevin L Zhang and David Poole. "A simple approach to Bayesian network computations". In: *Proc. of the Tenth Canadian Conference on Artificial Intelligence*. 1994.

[40] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[41] Arthur P Dempster. "A generalization of Bayesian inference". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 30.2 (1968), pp. 205–232.

[42] Fred Glover. "Tabu search—part I". In: *ORSA Journal on computing* 1.3 (1989), pp. 190–206.

[43] James L Beck and Ka-Veng Yuen. "Model selection using response measurements: Bayesian probabilistic approach". In: *Journal of Engineering Mechanics* 130.2 (2004), pp. 192–203.

[44] Nikos Vlassis and Aristidis Likas. "A greedy EM algorithm for Gaussian mixture learning". In: *Neural processing letters* 15.1 (2002), pp. 77–87.

[45] Joseph E Cavanaugh. "Unifying the derivations for the Akaike and corrected Akaike information criteria". In: *Statistics & Probability Letters* 33.2 (1997), pp. 201–208.

[46] Andrew A Neath and Joseph E Cavanaugh. "The Bayesian information criterion: background, derivation, and applications". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.2 (2012), pp. 199–203.

[47] Sumio Watanabe. "A widely applicable Bayesian information criterion". In: *Journal of Machine Learning Research* 14.Mar (2013), pp. 867–897.

[48] Harish S Bhat and Nitesh Kumar. "On the derivation of the bayesian information criterion". In: *School of Natural Sciences, University of California* 99 (2010).

[49] Gideon Schwarz. "Estimating the dimension of a model". In: *The annals of statistics* (1978), pp. 461–464.

[50] Marcelo MF Saba et al. "High-speed video observations of positive lightning flashes to ground". In: *Journal of Geophysical Research: Atmospheres* 115.D24 (2010).

[51] Marcelo MF Saba et al. "Upward lightning flashes characteristics from high-speed videos". In: *Journal of Geophysical Research: Atmospheres* 121.14 (2016), pp. 8493–8505.

[52] Carina Schumann et al. "High-speed observation of upward lightning flashes in Brazil". In: *International Lightning Detection Conference (ILDC)*. 2014.

[53] Carina Schumann et al. "High-speed video observation of lightning flashes over Johannesburg, South Africa 2017-2018". In: *2018 34th international conference on lightning protection (ICLP)*. IEEE. 2018, pp. 1–7.

[54] Marco Scutari, Maintainer Marco Scutari, and Hiton-PC MMPC. "Package 'bnlearn'". In: *Bayesian network structure learning, parameter learning and inference, R package version* 4.1 (2019).

[55] Xinyang Deng et al. "An improved method to construct basic probability assignment based on the confusion matrix for classification problem". In: *Information Sciences* 340 (2016), pp. 250–261.

[56] Joydwip Mohajon. *Confusion Matrix for Your Multi-Class Machine Learning Model*. 2021. URL: https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826.

[57] Stephen V Stehman. “Selecting and interpreting measures of thematic classification accuracy”. In: *Remote sensing of Environment* 62.1 (1997), pp. 77–89.

[58] David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[59] Michael A Nielsen. *Neural networks and deep learning*. Vol. 25. Determination press San Francisco, CA, 2015.

[60] Sushilkumar Kalmegh. “Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news”. In: *International Journal of Innovative Science, Engineering & Technology* 2.2 (2015), pp. 438–446.

[61] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[62] John Platt. “Sequential minimal optimization: A fast algorithm for training support vector machines”. In: (1998).

[63] Aaditya Desai and R Sunil. “Analysis of machine learning algorithms using WEKA”. In: *International Journal of Computer Applications* 975 (2012), p. 8887.