A -based space weather conditions machine learning classification model for GNSS PNT performance analysis

# Introduction

The Global Navigation Satellite System (GNSS) and its Positioning, Navigation, and Timing (PNT) service have matured to become an essential part of national infrastructure, public goods, and enablers of a vast number of emerging technology and socio-economic applications citestuff(filjar2022application). Maintaining the GNSS PNT quality has appeared fundamental for the sustainable development of modern economy and society citestuff(filjar2022application).

Overcoming the shortcomings and vulnerabilities of GNSS PNT is a scientific challenge, and the need of a wide variety of scientists, developers, operators, regulators, and users of GNSS-based systems and services citestuff(filjar2024ambient, filjar2022application). The Earth’s ionosphere, a part of the Earth’s atmosphere stretching from to above the Earth’s surface and comprised of rare but mostly charged atoms and molecules, is the major natural cause of GNSS PNT degradation citestuff(davies1990ionospheric, filic2018modelling).

This phenomenon especially affects most currently used GNSS receivers, which work as single-frequency receivers exposed to GNSS ionospheric effects citestuff(spilker1996global, filjar2022application). Driven by powerful and random flows of energy and particles from the Sun (space weather conditions), the ionospheric conditions define the properties of GNSS signal propagation through the Earth’s atmosphere and the resulting ionospheric delay citestuff(davies1990ionospheric, oxley2017uncertainties, filic2018modelling). The GNSS ionospheric delay causes errors in GNSS PNT algorithm deployment, designed to produce position, velocity, and time estimates for a GNSS receiver citestuff(spilker1996global, filic2018modelling).

The ionosphere affects GNSS satellite signals for position estimation by introducing signal propagation delay citestuff(spilker1996global, filic2018modelling). The GNSS ionospheric delay is a stochastic variable, whose value is determined by complex physical processes of space weather citestuff(filic2018modelling, filjar2022application).

How space weather affects GNSS PNT performance was explained using the Space weather – GNSS PNT performance coupling model citestuff(filic2018modelling), as depicted in Table [1](#tab:SpaceWeather).

The Space weather – GNSS PNT performance coupling model, after citestuff(filic2018modelling).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SPACE | GEOMAGNETIC | IONOSPHERE | GNSS | GNSS |
| WEATHER | FIELD |  | PSEUDO- | POSITIONING |
|  |  |  | RANGES | PERFORMANCE |

The GNSS signal encounters a certain number of charged particles from the satellite aerial to a mobile unit’s (GNSS receiver’s) aerial citestuff(filic2018modelling, filic2018modelling). Such an encounter is quantified using the Total Electron Content () defined by Equation 1 in citestuff(filic2018modelling, davies1990ionospheric), where denotes the height above the Earth’s mean sea level in , represents the vertical ionospheric profile, the volume density of charged particles at height , in . The dataset used in this study was derived from Receiver Independent Exchange Format (RINEX) observations using Global Positioning System (GPS) software by Seemala citestuff(SEEMALA202363).

It should be noted that in the sense of Equation 1, is defined as a result, a consequence, of the ionospheric conditions, and not their descriptor citestuff(filjar2022application). The GNSS ionospheric delay may be determined by derivation from the Appleton-Hartree equation, as given in Equation 2, where denotes the GNSS ionospheric delay in , denotes the velocity of an electromagnetic wave in vacuum in , and denotes the carrier wave frequency of the satellite signal in citestuff(spilker1996global, filic2018modelling).

Combining Equation 1 and Equation 2, one can conclude the linear relation between and , as given in Equation 3, a sub-model incorporated in the Space weather – GNSS PNT performance coupling model citestuff(filic2018modelling).

The GNSS ionospheric delay has been identified as a source of GNSS PNT degradation since the dawn of GNSS. GNSS systems offer various standard GNSS ionospheric delay estimation (correction) models to mitigate the deteriorating effects on GNSS PNT, such as the Klobuchar model citestuff(spilker1996global, filic2018modelling). The standard ionospheric correction models are global, and insufficiently flexible to update to mitigate GNSS ionospheric delay to satisfy rising demands on GNSS PNT performance citestuff(filjar2024ambient, filic2018modelling). The development of regional and local models attempts to solve the problem of GNSS PNT sustainable performance in various ionospheric conditions citestuff(sikirica2021risk). In recent developments, our team has proposed the Ambient-Aware Application-Aligned ((AA)2) PNT to take into account the actual ionospheric and geomagnetic conditions near a mobile unit (a GNSS receiver) citestuff(filjar2022application, filic2018modelling, filjar2024ambient). Direct measurements of the immediate geomagnetic and ionospheric condition variables may be supplied to a machine learning-based adapted GNSS ionospheric correction model, thus solving the single-frequency GNSS problem citestuff(filjar2022application). Previous research has identified predictors and target variables (descriptors of geomagnetic, ionospheric, and GNSS PNT conditions) citestuff(filic2018modelling, natras2022ensemble, natras2023regional). Current space weather severity scales, such as the one provided by National Oceanic and Atmospheric Administration (NOAA) citestuff(NOAA2024eSpace), are based on global space weather and geomagnetic indices averaged over a certain period (for instance, hours for the global index). The current space weather severity scales do not directly address classifying scenarios of GNSS performance deterioration and have limited potential in deployment for a GNSS ionospheric correction model. The classification of different scenarios of GNSS ionospheric conditions with adverse effects on the GNSS PNT remained an unsolved precondition needed for the development of a machine learning-based GNSS ionospheric delay correction model to render the GNSS PNT algorithm ionospheric conditions-agnostic citestuff(filjar2022application, filjar2024ambient).

A methodology for a machine learning-based classification of ionospheric conditions based entirely on observations of geomagnetic indices is described in this study. The proposed method is inspired by previous lightweight machine-learning techniques citestuff(filjar2020comparison) for classifying sensor readings. The model is sufficiently simple to be applied on computationally capable platforms with suitable geomagnetic field sensors, such as smartphones and connected/autonomous vehicles. The research presented acquires ambient data and analyses its statistical properties. The dataset is split into training and test sets. Several candidates for the GNSS ionospheric delay model are developed based on Disturbance Storm-Time () data taken from the INTERMAGNET citestuff(Intermagnet2022-cj) dataset, and reformatted to match the format of data. The machine learning (ML) models include a Support Vector Machine (SVM) with a Polynomial Kernel, C5.0 Decision Tree (DT), Naive Bayes (NB), shallow Neural Network (NN), Partial Least Squares (PLS), Flexible Discriminant Analysis (FDA) and shallow Neural Network (NN) using Principal Component Analysis (PCA) of the input data. A tailored set of validation methods is used to assess their performance. The optimal GNSS ionospheric delay correction model is identified based on GNSS PNT-related objective criteria, and its performance is demonstrated in an independent case study.

# Method and Data

A Support Vector Machine (SVM) with a Polynomial Kernel, C5.0 Decision Tree (DT), Naive Bayes (NB), Neural Network (NN), Partial Least Squares (PLS), Flexible Discriminant Analysis (FDA) and shallow Neural Network (NN) using Principal Component Analysis (PCA) of the input data were tested based on their ability to classify a set of observations of the geomagnetic field in , and other predictors, into one of the scenarios of space weather conditions based on .

Multiple -dependent classes were predefined using theoretical knowledge. Statistical analysis of the data confirmed that distributions of other variables change for different ranges, supporting the validity of the classification. The study assumes that the dependent output variable, the class, can be predicted based on the independent variables used as input.

data was obtained using GPS software by Seemala to process RINEX observations citestuff(SEEMALA202363). The INTERMAGNET citestuff(Intermagnet2022-cj) dataset contains and data from 2014 for a measuring station maintained by Geoscience Australia in Kakadu, referred to as KDU in the database, at degrees latitude and degrees longitude near Darwin, Nothern Territory, Australia. The two datasets are merged based on location, year, month, day, and time of day in hours.

## Method

The models were selected because they represent larger families of classification methods. SVM models are supervised maximum margin models. DT models also apply supervised learning. NB classifiers are probabilistic classifiers that can be parametric or non-parametric, but this study uses a non-parametric approach. PLS is a non-parametric linear regression model. FDA uses multiple non-parametric linear regression models to create a non-linear classification. PCA is a linear dimensionality reduction technique that extracts a predefined number of components for training an NN model. NN models imitate the brain using artificial neurons to produce outputs based on the input and the activation function. NN models require that the structure be predefined, and hyperparameters are tuned.

All NN models were applied based on research by Kuhn for the *R* *caret* package citestuff(kuhn2013applied, KuhnCaret2024, rprojectProjectStatistical).

### Support Vector Machine

In machine learning, a Support Vector Machine (SVM) or Support Vector Network (SVN) model is a supervised maximum margin model with associated learning algorithms that analyzes data for classification, regression citestuff(Cortes1995), or outlier detection citestuff(scikit-learn2023). SVM models are also effective for non-linear classification using the hyperplane kernel trick citestuff(Boser1992, Aizerman1964). Intuitively, a good separation is achieved by the hyperplane with the greatest distance to the nearest point in the training data belonging to any class citestuff(HastieRosset2009). Support Vector Clustering (SVC) citestuff(BenHur2001) is a related method for clustering unlabeled data.

Meyer, Leisch, and Hornik compared SVM models with other classifiers citestuff(Meyer2003). However, it is unclear whether SVM predictions perform better than other linear models, such as logistic, and linear regression.

A multidimensional feature space increases the generalization error of SVM models, so additional samples are needed to enhance performance citestuff(Jin2012). To keep the computational burden reasonable, a kernel probability density function is chosen to fit the problem citestuff(Press2007).

The performance of SVM depends on the kernel probability density function, its parameters, and the "soft" margin parameter chosen in cross-validation to train the final classification model citestuff(Hsu2003).

### Decision Tree

Decision Tree (DT) models are used for supervised learning in statistics, data mining citestuff(Rokach2014), and machine learning. Classification trees use a discrete target variable, while regression trees can handle continuous values citestuff(Studer2011). DT models are popular due to their comprehensibility and simplicity citestuff(Wu2008).

A tree is recursively partitioned by dividing the original set, or root node, into subsets that form descendants, or successors, using classification rules based on features citestuff(ShalevShwartz2014). This process of Top-Down Induction of Decision Trees (TDIDT) citestuff(Quinlan1986) is an example of a greedy algorithm that is the most common strategy for learning DT models from data citestuff(Rokach2005). C5.0, used in the *caret* package in *R*, has a similar approach and improves the ID3 and C4.5 algorithms.

### Naive Bayes

In statistics, Naive Bayes (NB) models, simple Bayes, or independent Bayes citestuff(Hand2001) classifiers are a family of linear "probabilistic classifiers" that assume that, given a target class, the features are conditionally independent. These classifiers are among the simplest Bayesian network models citestuff(McCallum2011), but a NB classifier model is not necessarily a Bayesian method. Maximum likelihood training evaluates a closed-form expression citestuff(Russell1999) in linear time instead of using iterative approximation.

Despite their simplicity, NB classifier models have performed well in real-world situations citestuff(Metsis2006). However, a comprehensive comparison in 2006 showed that Bayesian classification performed worse than boosted trees or Random Forest (RF) citestuff(Caruana2006). An advantage of NB over other models is a smaller amount of training data citestuff(John2013, Mccallum2001).

In the abstract, NB models assign conditional probabilities to each of the possible outcomes or classes for an instance to be classified, represented by the vector encoding features citestuff(Murty2011). The model can be reformulated using Bayes’ theorem in Equation 4.

### Neural Networks

The neurons of human or animal brains provide the basis for a Neural Network (NN) or Artificial Neural Network (ANN) with connected units or nodes called artificial neurons in machine learning citestuff(mitExplainedNeural, brahme2014comprehensive).

Shallow NN models typically contain only a few hidden layers for processing between the input layer that receives the data and the final layer that produces the output citestuff(olden2002illuminating, ozesmi1999artificial). A network with at least two hidden layers citestuff(bishop2006pattern) is a deep NN model.

Gradient-based methods such as backpropagation estimate ANN parameters citestuff(vapnik2013nature) to minimize the difference or empirical risk between the output and target labels, expressed in a loss function citestuff(goodfellow2016deep). The hyperparameters may also be modified to suit the problem citestuff(probst2019tunability) during an extensive tuning process, like the one used in this study.

Principal Component Analysis (PCA) citestuff(pearson1901liii, stewart2019introduction) is a linear dimensionality reduction technique in exploratory data analysis, visualization citestuff(jolliffe2016principal), and preprocessing. The method in the *caret* package in *R* citestuff(ripley2007pattern) uses PCA in preprocessing.

### Partial Least Squares

Partial Least Squares (PLS) regression, or projection to latent structures, citestuff(wold2001pls, abdi2010partial), is a linear regression statistical model that transforms the predicted and the observable variables to a new space. PLS methods are bilinear factor models because the and are projected to new spaces. In Partial Least Squares Discriminant Analysis (PLS-DA), is categorical citestuff(saebo2008lpls).

Using paired observations . PLS finds the normalized direction that maximizes the covariance in the first step , shown in Equation 5.

The scores (projections) form an orthogonal basis in PLS regression, and the loadings are chosen to achieve this. In PCA orthogonality is not imposed on scores citestuff(lindgren1993kernel) but onto loadings instead citestuff(de1994comments, dayal1997improved). Many versions of PLS exist citestuff(de1993simpls, rannar1994pls) for estimating the factor and loading matrices, such as the PLS1 algorithm citestuff(takane2016pls, hoskuldsson1988pls).

### Flexible Discriminant Analysis

Flexible Discriminant Analysis (FDA) is a general methodology that creates the discriminant surface for a multigroup non-linear classification model citestuff(fisher1936use, mclachlan2005discriminant) based on a mixture of non-parametric linear regression models citestuff(hastie1995penalized), such as Multivariate Adaptive Regression Splines (MARS) and Linear Discriminant Analysis (LDA).

Many predictors can be used at once in FDA citestuff(HastieTibshirani2009), and variable interactions are automatically noted citestuff(hastie1994flexible). FDA is complex but execution time and computational load are adequate citestuff(reynes2006choice). The algorithm is not largely affected by outliers citestuff(phillips2017applying).

Changed settings significantly alter FDA citestuff(hallgren2019species), and estimation might fail if predictors are correlated. FDA is prone to overfitting citestuff(thuiller2016ensemble), and challenging to comprehend citestuff(quillfeldt2017influence). Feature normality and equal group covariances are assumed citestuff(zhang2019using, wetcher2011analyzing).

LDA, Normal Discriminant Analysis (NDA), or discriminant function analysis citestuff(cohen2013applied) is a generalization of Fisher’s linear discriminant defined in 1936 citestuff(fisher1936use, archive765Discriminant, rao1948utilization). The results of LDA may be utilized directly, as demonstrated in this experiment, but they are more often used to reduce dimensionality citestuff(hansen2005using).

## Data Description and Analysis

Dynamic space weather conditions, such as solar activity and geomagnetic storms, can affect GNSS PNT performance and high-frequency GPS signals passing through the ionosphere, motivating work on error modeling citestuff(klobuchar1986design, zolesi2014ionospheric). Changes in the ionosphere include the density distribution, , and the current balance citestuff(komjathy2023global).

Geomagnetic storms cause signal deterioration by affecting Global Electric Current (GEC) variability citestuff(NOAA2024aRadioCommunications). The ionosphere may show changes related to location, geomagnetic and solar activity, sunspots, local time citestuff(kotz2005encyclopedia), seasonality, thunderstorms citestuff(vellinov1992ionospheric), nuclear experiments, earthquakes citestuff(liu20142013), and other phenomena.

Many parameters can describe the Earth’s geomagnetic field citestuff(ulukavak2018analysis). This study focuses on parameters describing disturbances, most importantly indices calculated from and indices, Total Electron Content (), standard deviation of Total Electron Content (), and Disturbance Storm-Time ().

Incorporating parameters such as the -indices and -indices, which provide global measures of geomagnetic activity, alongside local and -index values, allows for a more detailed assessment of the space environment and its potential effects on GNSS signals. Values of , , , were used with the , , and components of the Earth’s magnetic field to train machine-learning models.

### Magnetic Field Indices

The Earth’s magnetic field has similarities to that of a bar magnet. However, plasma gushes from the solar corona and the domain of the Sun influence the interplanetary magnetic field citestuff(schwenn2001solar, melnikov1990relationships). The , , and vectors represent interplanetary magnetic field indices. and are parallel to the plane of orbits, and the third component is perpendicular. Widely available hand-held devices, such as Android smartphones citestuff(Android2024SensorTypes), measure magnetic field indices in micro-Tesla (). The Android magnetometer reports accuracy through a status variable. Readings are calibrated using temperature compensation, factory (or online) soft-iron, and online hard-iron calibration.

### Geomagnetic Storm Indices

The geomagnetic storm -index is an integer from to measuring disturbances in global geomagnetic activity. The maximum positive and negative fluctuations of the horizontal components of the Earth’s magnetic field citestuff(de2007ionosphere), and , during hours, relative to a quiet day, are added to determine the total maximum fluctuation.

Each observatory uses different threshold values to convert the maximum (nano-Tesla) fluctuation to a -index value. The thresholds for each observatory are adjusted so that the historical rate of occurrence for each -index value is similar across all observatories. Observatories with a lower geomagnetic latitude use a lower fluctuation in and to achieve each -index value.

The threshold valid in Boulder, Colorado citestuff(NOAA2024bKindex) for transforming the maximum and fluctuation measured in using a magnetometer into a -index value is given in Table [2](#tab:K).

The relationship between the -index and magnetometer observations of and in in Boulder, Colorado.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| and () | - | - | - | - | - |
|  |  |  |  |  |  |
| and () | - | - | - | - |  |

### Planetary Geomagnetic Storm Indices

The planetary geomagnetic storm -index citestuff(Matzka2024) is derived from -hour-based indices from magnetometer stations between and degrees of north and south latitude. Announcements and warnings citestuff(NOAA2024cAlertsWatches) of geomagnetic changes and disturbances in the Earth’s magnetic field are based on the -index.

Hourly geomagnetic storm index data are available on the NASA Goddard SPDF citestuff(Papitashvili2024About, Papitashvili2024Explorer) web pages. The scale values of the -index are determined by the change of the geomagnetic field and the geomagnetic storm effect in .

The official planetary -index is a weighted average of -indices from multiple observatories. When -index data is not available in real-time, operators such as The National Oceanic and Atmospheric Administration (NOAA) Space Weather Prediction Center (SWPC) calculate near real-time estimates of the -index citestuff(Myint2022).

The -index as related to geomagnetic storm descriptions and warnings using the NOAA G scale citestuff(NOAA2024eSpace, NOAA2024fScales) is given in Table [3](#tab:Kp).

In March 2021, was assigned a DOI with a dataset citestuff(Matzka2021a) and a scientific publication citestuff(Matzka2021b) for reference.

The -index as related to geomagnetic storm descriptions and warnings using the NOAA G scale citestuff(NOAA2024eSpace, NOAA2024fScales).

| -index |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| NOAA Space Weather Scale Geomagnetic Storm Level | G0 | G1 | G2 | G3 | G4 | G5 |

### Equivalent Three Hourly Range Geomagnetic Storm Indices

The -index citestuff(NOAA2024gGeomagneticIndices) represents a daily average level of magnetic activity. Because the relationship between the -index and magnetometer fluctuations is not linear, the -index values are not directly used for calculating average values. Each -index or -index is converted into the "equivalent three hourly range" -index or -index that uses a linear scale. Table [4](#tab:a) illustrates the conversion between -index and -index values citestuff(NOAA2024bKindex). Table [5](#tab:ap) illustrates the conversion between -index and -index values citestuff(NOAA2024gGeomagneticIndices). An average of -indices (lowercase) is used as the daily -index (uppercase). As an example, if -indices were , , , , , , , and , the -indices equal , , , , , , , and , and the -index equals

.

The -index values for each -index, from citestuff(NOAA2024bKindex).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |

The -index values for each -index, from citestuff(NOAA2024gGeomagneticIndices).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

### Disturbance Storm-Time

The next relevant parameter is Disturbance Storm-Time (), also known as the geomagnetic activity -index, another measure of geomagnetic storm intensity. The -index is a geomagnetic indicator of magnetic flux changes derived from measurements taken by a network of ground-based magnetometer stations near the magnetic equator, which continuously monitor and , the horizontal components of Earth’s magnetic field citestuff(zolesi2014ionospheric). The -index describes ring currents forming above the sub-equatorial region and affecting the ionospheric regions in mid-latitudes.

To calculate the -index, variations in and , the horizontal magnetic field, are obtained from multiple stations and averaged. The average is subtracted from a baseline value representing the quiet-time magnetic field. The resulting value in measures the intensity of geomagnetic disturbances, with increasingly negative values indicating stronger geomagnetic storms.

The -indice measurements as an hourly average were evaluated and published on a web interface by the NASA Goddard SPDF citestuff(Papitashvili2024About, Papitashvili2024Explorer), and the Geomagnetism and Space Magnetism Data Analysis Center of the Institute of Science, Kyoto University in Japan.

Loewe and Prölss citestuff(loewe1997classification) classified magnetic activity -indices into five storm classes in 1997, as shown in Table [6](#tab:loewe1997). Gonzalez et al. citestuff(gonzalez1994geomagnetic) used three groups for the same data in 1994, as shown in Table [7](#tab:gonzalez1994), similar to Kamide et al. in 1998 citestuff(kamide1998two), Rozhnoi et al. in 2004 citestuff(rozhnoi2004middle), and Contadakis et al. in 2012 citestuff(contadakis2012total).

Storm classification from Loewe and Prölss, in 1997 citestuff(loewe1997classification).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| Storm class | Fraction | | -Range |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| weak |  |  | to |  |  |  |  |
| moderate |  |  | to |  |  |  |  |
| strong |  |  | to |  |  |  |  |
| severe |  |  | to |  |  |  |  |
| great |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| The bars above the magnetic indices indicate median values. | | | | | | | |

Storm classification from Gonzales et al., in 1994 citestuff(gonzalez1994geomagnetic).

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | , | , | , hours |
|  |  |  |  |
|  |  |  |  |
| Intense |  |  |  |
| Moderate |  |  |  |
| Small (typical substorm) |  |  |  |
|  |  |  |  |

### Training and Testing Dataset

Among samples, were considered outliers because the is larger than or equal to . The remaining samples are used for training and testing. The samples are divided into training and testing datasets as close as possible to a ratio of for training and for testing. The division was stratified so that an approximately equal ratio of classes was present in both the training and testing data, which is a feature of the *createDataPartition* function from the *caret* *R* library that was used citestuff(KuhnDataSplitting2024, hyndman2018forecasting, createDataPartition2024).

Samples were split into class ranges, based on values derived from theoretical knowledge of different storm phases and similar to Loewe and Prölss citestuff(loewe1997classification). The upper range limits are excluded, while the lower ones are included.

Table [8](#tab:Dstranges) lists the class ranges used in this study, the total number of samples in each class, and the number of samples used for testing, and training. It is evident from this division that the normal (N), and recovery (R) classes are far more common than other classes with very high or low values, impacting model performance.

-based classification rules used in this study, the total number of samples in each class, and the number of samples used for testing, and training.

|  | Storm phase classification | Total samples | Test samples | Train samples |
| --- | --- | --- | --- | --- |
|  | positive phase (P) |  |  |  |
|  | normal (N) |  |  |  |
|  | recovery phase (R) |  |  |  |
|  | through (T) |  |  |  |
|  | extreme (E) |  |  |  |
| Any | Any |  |  |  |

### Data Preprocessing

Data preprocessing can increase classification accuracy citestuff(Fan2008). There are many ways to standardize data, such as minimum-maximum, normalization by decimal scaling, and Z-score citestuff(Mohamad2013). Subtracting the mean and dividing by the variance for each feature are commonly used for SVM models citestuff(Fennell2019) and other models tested in this study, so this approach was chosen.

The values *scale* and *center* were used in the code for this study in the *preProcess* parameter for the *train* function from the *caret* package developed for *R*. The option *center* subtracts the mean of each feature while *scale* divides by the standard deviation.

### Correlation

Figure [1](#fig:correlation) contains a heat map of the correlation between all variables used in this study. The highest correlation of is between and , as expected based on theoretical knowledge, and supporting the hypothesis that can be predicted using . The second highest correlation is between and and equals , meaning can be used to predict with an opposite trend.

![](data:application/pdf;base64,)

A heat map of the correlation between all variables used in this study, when the is less than . Red represents a high positive correlation, blue represents a high negative correlation, and white represents a low correlation. Variables are fully correlated with themselves, so values on the secondary diagonal equal . The matrix is symmetrical concerning the secondary diagonal because the same combination of correlated variables is achieved when swapping the row and column.

The box plots of all variables for different ranges of values in Figure [2](#fig:iono3boxplot) demonstrate that the minimum, maximum, and arithmetic mean of decrease for larger . The opposite is true for , as indicated by a high correlation in Figure [1](#fig:correlation). exhibits a reverse trend when compared to , but it is less prominent. is the most stable with the smallest changes related to when observing , , and .

![](data:application/pdf;base64,)

Box plots of all variables, when the is less than , for different ranges of values defining the class label used in this study.

Table [9](#tab:minmax) provides the minimum, quartile, median, arithmetic mean, quartile, and maximum values for all variables when the is less than . These values suggest that variables are not normally distributed. The Kolmogorov-Smirnov and Shapiro-Wilk normality tests, using *R* functions *ks.test* citestuff(ks.test2024) and *shapiro.test* citestuff(shapiro.test2024), did not yield a -value larger than the selected -value of for any variable, further strengthening the claim that variables do not follow a normal (Gaussian) distribution.

The minimum, quartile, median, arithmetic mean, quartile, and maximum values for all variables when the is less than .

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Min. |  |  |  |  |  |  |  |
| Qu. |  |  |  |  |  |  |  |
| Median |  |  |  |  |  |  |  |
| Mean |  |  |  |  |  |  |  |
| Qu. |  |  |  |  |  |  |  |
| Max. |  |  |  |  |  |  |  |

## Performance Metrics

The metrics and terminology used to evaluate the classifier performance were taken from the confusion matrix defined in the *R* function *confusionMatrix* in the *caret* library citestuff(kuhn2008building, confusionMatrix2024).

The examples for each performance metric use only two groups (Yes and No, positive and negative). For multiple classes, results are calculated by a "one versus all" approach, viewing each class as positive and all others as negative.

### Confusion Matrix

A True Positive (TP), or True Negative (TN) result in the confusion matrix correctly indicates that the sample belongs to a class. A False Positive (FP), or False Negative (FN) classification result incorrectly claims that the sample belongs to a class and represents a type I or type II error, respectively.

Sensitivity, recall, hit rate, or True Positive Rate (TPR), calculated as , and specificity, selectivity, or True Negative Rate (TNR), calculated as , represent the proportion of correctly classified samples among those that truly belong to a class citestuff(altman1994diagnostic1, altman1994diagnostic2).

Precision, or Positive Predictive Value (PPV), calculated as , and Negative Predictive Value (NPV), calculated as , represent the proportion of correctly classified samples among those classified to a class.

Prevalence, calculated as , and Detection Prevalence (DP), calculated as represent the proportion of samples that truly belong to or are classified to each class, respectively. Prevalence values other than an even split by the number of classes indicate a class imbalance.

The Detection Rate (DR), calculated as , and accuracy (Acc), calculated as , represent the proportion of correctly classified samples for separate classes or any class, respectively.

Since accuracy is unusable for unbalanced classes, Balanced Accuracy (BA) citestuff(velez2007balanced) is calculated as and represents the arithmetic mean of TPR and TNR, which are centered on the positive and the negative classes, respectively.

# Research Results

The accuracy for all models when using all variables, all variables except (no ), all variables except , , and (no ), only , , and (coord), only , , and ( ), only , , and ( ), or only , , and ( ) as input is displayed in Table [10](#tab:acc:all). Accuracy results can lead to an incorrect conclusion if the dataset is unbalanced. Out of samples for training and testing, belonged to the largest N class, and to the smallest E class, as listed in Table [8](#tab:Dstranges). More comprehensive results by class are presented only for the Naive Bayes (NB) classifier, the second most successful model. The Neural Network (NN) classifier made no errors, so further discussion is redundant.

The accuracy for each model using different combinations of variables as input.

| Model | all | no | no | coord |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SVM Poly |  |  |  |  |  |  |  |
| C5.0 DT |  |  |  |  |  |  |  |
| NB |  |  |  |  |  |  |  |
| NN |  |  |  |  |  |  |  |
| PLS |  |  |  |  |  |  |  |
| FDA |  |  |  |  |  |  |  |
| PCA NN |  |  |  |  |  |  |  |

The execution time in seconds () utilizing the *R* *system.time* function citestuff(system.time2024) for each model and different combinations of variables as input is displayed in Table [11](#tab:time:total). The experiment was run on *Windows* 11 using *R Studio* version 2024.04.2+764 and *R* version 4.4.1, the AMD Radeon RX 6600 Graphics Processing Unit (GPU), GB of Random Access Memory (RAM), and the AMD Ryzen 5 PRO 4650G Central Processing Unit (CPU) with cores. Execution time is significant because built-in systems for mobile devices using GNSS PNT have low computational capabilities.

The execution time in seconds () for each model when using different combinations of variables as input.

| Model | all | no | no | coord |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SVM Poly |  |  |  |  |  |  |  |
| C5.0 DT |  |  |  |  |  |  |  |
| NB |  |  |  |  |  |  |  |
| NN |  |  |  |  |  |  |  |
| PLS |  |  |  |  |  |  |  |
| FDA |  |  |  |  |  |  |  |
| PCA NN |  |  |  |  |  |  |  |

## Comprehensive Results for the Naive Bayes Model

The confusion matrix and the performance indicators derived from it for the Naive Bayes (NB) model, when using different combinations of variables as input, are depicted in Table [12](#tab:cm:merged), Table [13](#tab:cs:reverse:all:nb), Table [14](#Xaeecd2ecd40307013bc0d03f5868e5e72a85912), and Table [15](#tab:cs:reverse:yzap:nb). The NB model was chosen because of a low computational load and high accuracy. The results for combinations of input variables with a accuracy using the NB model are ommited.

The confusion matrix for the Naive Bayes model when using all variables as input (left), all variables except as input (center), and only , , and as input (right).

|  | Reference - all | | | | | Reference - no | | | | | Reference - | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pred. | E | N | P | R | T | E | N | P | R | T | E | N | P | R | T |
| E |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| N |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| P |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| R |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| T |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

The performance indicators derived from the confusion matrix for the Naive Bayes model when using all variables as input.

|  | Class | | | | |
| --- | --- | --- | --- | --- | --- |
| Statistics | E | N | P | R | T |
| Sensitivity |  |  |  |  |  |
| Specificity |  |  |  |  |  |
| PPV |  |  |  |  |  |
| NPV |  |  |  |  |  |
| DR |  |  |  |  |  |
| DP |  |  |  |  |  |
| BA |  |  |  |  |  |

The performance indicators derived from the confusion matrix for the Naive Bayes model when using all variables except as input.

|  | Class | | | | |
| --- | --- | --- | --- | --- | --- |
| Statistics | E | N | P | R | T |
| Sensitivity |  |  |  |  |  |
| Specificity |  |  |  |  |  |
| PPV |  |  |  |  |  |
| NPV |  |  |  |  |  |
| DR |  |  |  |  |  |
| DP |  |  |  |  |  |
| BA |  |  |  |  |  |

The performance indicators derived from the confusion matrix for the Naive Bayes model when using only , , and as input.

|  | Class | | | | |
| --- | --- | --- | --- | --- | --- |
| Statistics | E | N | P | R | T |
| Sensitivity |  |  |  |  |  |
| Specificity |  |  |  |  |  |
| PPV |  |  |  |  |  |
| NPV |  |  |  |  |  |
| DR |  |  |  |  |  |
| DP |  |  |  |  |  |
| BA |  |  |  |  |  |

# Discussion

The Neural Network (NN) model has the highest execution time for any subset of the input variables due to its complexity and extensive training, evident from the data in Table [11](#tab:time:total). However, it achieved a accuracy, shown in Table [10](#tab:acc:all). The Naive Bayes (NB) model has the second highest accuracy, over . The NB model achieved a accuracy, except when using all input variables, all input variables except Disturbance Storm-Time (), or only , , and . The difference in accuracy between the NN and the NB model is negligible, and the training time for the NN model is more than twice as long. An analysis in 2004 showed reasonable theoretical reasons for the seemingly incredible performance of NB classifiers citestuff(Zhang2004).

All other models fail to achieve an accuracy over , except when using all input variables, including . The class was derived by thresholding the continuous value and discretizing it by converting it to a single character, so the original value should not be used as model input. This labeling method explains the increase in accuracy when adding the input variable in all except the two best-performing models. For example, the C5.0 DT model did not consider any variable except , indicated by a variable importance. The Support Vector Machine (SVM) model with a Polynomial Kernel is consistently the worst-performing for any subset of input variables, never achieving an accuracy over . None of the models achieved an accuracy under , so they were all moderately successful.

When studying the performance of the NB model by individual class in Table [12](#tab:cm:merged), Table [13](#tab:cs:reverse:all:nb), Table [14](#Xaeecd2ecd40307013bc0d03f5868e5e72a85912), and Table [15](#tab:cs:reverse:yzap:nb), samples of the T class are all correctly classified except for one case, but the testing is less extensive since it is the second smallest class. Samples of the P class are all correctly classified. Samples of the R class are sometimes erroneously assigned to the N, T, or less commonly the E class. Two of the most common misclassifications were assigning samples of the N class with the highest detection rate and prevalence to the P or R classes with the third and second highest detection rate and prevalence. This is due to the largest number of samples in the N class between the P and R classes with the narrowest range. Samples of the N class are least often included in the E class, whose range is the furthest apart from the N class. A sample of the E class was attributed to the N class only once. The E class is the smallest with only two testing samples, which must be accounted for when interpreting these results, such as the lowest balanced accuracy of and sensitivity of when using only , , and as input, or the lowest positive predictive value of when using all variables as input.

# Conclusion

The presented study aims to classify ambient conditions of space weather events for sub-equatorial regions. Global Navigation Satellite System (GNSS) Positioning, Navigation, and Timing (PNT) performance is significantly affected by such events. It would be beneficial to warn users of a geomagnetic/ionospheric storm.

Classification models using machine learning were applied to descriptions of the geomagnetic field expressed in Total Electron Content () and other input data. It was assumed observations contained independent variables to generate the dependent variable representing the Disturbance Storm-Time () class.

Statistical analysis confirmed that other variables change distribution based on , not . Continuous values in different ranges were converted into discrete classes based on statistics, previous theories, and research.

An Support Vector Machine (SVM) with a Polynomial Kernel, C5.0 Decision Tree (DT), Naive Bayes (NB), Neural Network (NN), Partial Least Squares (PLS), Flexible Discriminant Analysis (FDA), and Principal Component Analysis (PCA) NN model created a -based classification from multiple combinations of input variables.

The NB model achieved perfect accuracy for most tested combinations of input variables, and over in all cases. It is slightly less successful than the NN model, and the total execution time is at least two times shorter, making it more suitable for use in compact, low-performance, and low-cost portable devices such as smartphones.

# Acronyms

*(AA)2* Ambient-Aware Application-Aligned

*ANN* Artificial Neural Network

*BA* Balanced Accuracy

*CCA* Canonical Correlation Analysis

*CI* Confidence Interval

*CNN* Convolutional Neural Network

*CPU* Central Processing Unit

*DBSCAN* Density-Based Spatial Clustering

*DP* Detection Prevalence

*DR* Detection Rate

*DT* Decision Tree

*Dst* Disturbance Storm-Time

*EOF* Empirical Orthogonal Functions

*EPB* Equatorial Plasma Bubbles

*EUV* Extreme ultraviolet

*EVD* Eigenvalue Decomposition

*FDA* Flexible Discriminant Analysis

*FN* False Negative

*FP* False Positive

*GEC* Global Electric Current

*GFZ* German Research Centre for Geosciences

*GNSS* Global Navigation Satellite System

*GPS* Global Positioning System

*GPU* Graphics Processing Unit

*ION* The Institute of Navigation

*KLT* Karhunen–Loève Theorem

*LDA* Linear Discriminant Analysis

*MANOVA* Multivariate Analysis Of Variance

*MARS* Multivariate Adaptive Regression Splines

*ML* machine learning

*NAS* Neural Architecture Search

*NASA* National Aeronautics and Space Administration

*NB* Naive Bayes

*NDA* Normal Discriminant Analysis

*NIR* No Information Rate

*NN* Neural Network

*NOAA* National Oceanic and Atmospheric Administration

*NPV* Negative Predictive Value

*NRCAN* Natural Resources Canada

*PCA* Principal Component Analysis

*PLS* Partial Least Squares

*PLS-DA* Partial Least Squares Discriminant Analysis

*PNT* Positioning, Navigation, and Timing

*POD* Proper Orthogonal Decomposition

*PPV* Positive Predictive Value

*RAM* Random Access Memory

*RF* Random Forest

*RIN* The Royal Institute of Navigation

*RINEX* Receiver Independent Exchange Format

*SAR* Synthetic Aperture Radar

*SPDF* Space Physics Data Facility

*SVC* Support Vector Clustering

*SVD* Singular Value Decomposition

*SVM* Support Vector Machine

*SVN* Support Vector Network

*SWPC* Space Weather Prediction Center

*TDIDT* Top-Down Induction of Decision Trees

*TEC* Total Electron Content

*TID* Traveling Ionospheric Disturbance

*TN* True Negative

*TNR* True Negative Rate

*TP* True Positive

*TPR* True Positive Rate

*URSI* Union Radio-Scientifique Internationale

*dTEC* standard deviation of Total Electron Content

# Declarations

## Availability of data and materials

The datasets used during the current study are available from the corresponding author upon reasonable request.

## Competing interests

The authors declare no conflict of interest.

## Funding

The authors have no funding sources to declare.

## Authors’ contributions

LŽ and DK contributed to conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing of the original draft, reviewing and editing the text, and visualization. TI contributed to validation, formal analysis, investigation, review and editing, supervision, and project administration. RF contributed to conceptualization, methodology, investigation, resources, writing of the original draft, reviewing and editing the text, and supervision. All authors read and approved the final manuscript.

## Acknowledgements

Not applicable

## Authors’ information

LŽ is a doctoral student at the Faculty of Engineering, University of Rijeka, currently employed as an assistant at the Department of Computer Science. Her research interests include applied machine learning in biology, chemistry, medicine, and transportation. Maritime transportation is also the subject of her doctoral research.

DK is a doctoral student at the Faculty of Engineering, University of Rijeka, currently employed as an assistant at the Department of Computer Science. His research interests include applied machine learning in object recognition and image segmentation. Image classification and labeling are also the subject of his doctoral research.

TI earned his MSc and PhD in telecommunication engineering from the University of Ruse “Angel Kanchev,” Ruse, Bulgaria, in 1999 and 2007, respectively. He is a full professor at the Department of Telecommunication, University of Ruse “Angel Kanchev.” His research interests include mobile communications networks, signal processing, wireless technologies, and satellite navigation. Professor Iliev is a Member of IEEE.

RF is an external Professor of Electronics Engineering with the Department for Computer Science, Faculty of Engineering, and Center for Artificial Intelligence and Cybersecurity, both at the University of Rijeka, Croatia, and Head of the Laboratory for Spatial Intelligence at Hrvatsko Zagorje Krapina University of Applied Science, Krapina, Croatia. He holds a BSc, MSc, and PhD in electrical engineering, obtained in 1987, 1994, and 2007, respectively, from the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. His professional interests include Ambient-Aware Application-Aligned ((AA)2) Global Navigation Satellite System (GNSS) Positioning, Navigation, and Timing (PNT), spatial uncertainty quantification, spatial statistical learning, predictive modeling, statistical signal processing, trajectory analysis and prediction, and occupancy modeling. Professor Filjar is a Fellow of The Royal Institute of Navigation (RIN) (London, United Kingdom), a Member of The Institute of Navigation (ION) (Manassas, Virginia), a Senior Member of Union Radio-Scientifique Internationale (URSI) (Ghent, Belgium) and a Member of The Society for Industrial and Applied Mathematics (Philadelphia, Philadelphia).