Classification of Solar Flares using Data Analysis and Clustering of Active Regions

H. Baeke[1], J. Amaya[1], G. Lapenta[1]

[1]Center for mathematical Plasma Astrophysics, Department of Mathematics, KU Leuven, Celestijnenlaan 200B, 3001 Leuven, Belgium

Key Points:

- Sparse autoencoders are a successful method to create a higher-dimensional hyperplane where the spread between the data is larger.
- Unsupervised method, K-means, is able to distinguish inactive regions from C-/M-flaring active regions, and very active X-flare regions.
- The magnetic field properties of C- and M-flaring active regions are very similar.

Corresponding author: Hanne Baeke, `hanne.baeke@kuleuven.be`

Abstract

We devised a new data analysis technique to identify the threat level of solar active regions by processing a combined data set of magnetic field parameters and flaring activity. The data set is composed of two elements: a reduced factorization of SHARP parameters of the active regions, and information about the flaring activity at the time of measurement of the SHARP parameters. Machine learning is used to reduce the data and to subsequently classify the active regions. For this classification we used both supervised and unsupervised methods. The following processing steps are applied to reduce and enhance the SHARP data: outlier detection, redundancy elimination with common factor analysis, addition of sparsity with autoencoders, and construction of a balanced data set with under- and over-sampling. The supervised method, K-nearest neighbors, produces very good results on the strong X- and M-flares, with TSS scores of respectively 0.94 and 0.75. As unsupervised methods we used clustering models, K-means and Gaussian Mixture Models. We find that both techniques are able to distinguish non-flaring and flaring active regions. Moreover, K-means is able to distinguish C-/M-flaring active regions from X-flaring active regions. For processing purposes an unsupervised method is more useful, since the type of flare will not be available. Therefore, we conclude that K-means provides the most promising results.

Plain Language Summary

One of the main sources of space weather activity are solar active regions. In these zones the magnetic activity of the Sun is increased and can produce the two most energetic events in the solar system: flares and coronal mass ejections. We investigate the magnetic field parameters of active regions, and the amount of energy they release. Our end goal is to produce an automatic model that can forecast the energy level released by a flare from solar active regions, using only their current magnetic field parameters.

For this study, we used machine learning techniques that recognize patterns in data, without being explicitly told what to look for. These techniques can sometimes find patterns that escape the human intuition. The technique classifies different active regions, based on their magnetic parameters, identifying those that can release large amounts of energy in the near future.

Our technique is able to discover differences between flaring and non-flaring active regions. We are also able to distinguish weakly flaring active regions from the strongest flaring active regions. However, there is still room for improvement. Future research should incorporate other data types and time series to get better results.

1 Introduction

Solar flares are sudden electromagnetic explosions on the solar surface. These flares are triggered by the rapid reconfiguration of the magnetic field in solar active regions. The most intense solar flares are frequently accompanied by coronal mass ejections (CMEs) and solar energetic particles. Kawabata et al. (2018) show that CMEs are associated with approximately all events whose X-ray flux is larger than $10^{-3.9} Wm^{-2}$, which correspond to the X-flares. Solar activity, like solar flares and CMEs, pose a serious threat to the near-Earth environment. They can produce streams of highly energetic particles, which can affect the Earth's magnetosphere within a few hours or minutes (Cinto et al., 2020). These particles pose radiation hazards to astronauts and spacecrafts (Mikaelian, 2009). Flares are also associated with radio communication disruptions (Knipp et al., 2016; Redmon et al., 2018), and the associated high energy particles can ionize our atmosphere at low altitudes (Liu et al., 2021). CMEs can trigger geomagnetic storms, which can disable satellites (Dang et al., 2022) and even knock out electrical power grids (Pulkkinen et al., 2005). Should such a large geomagnetic storm happen nowadays, it would have catastrophic results, causing considerable economic damage (Eastwood et al., 2017). For example, the 1977 New York City blackout cost is estimated at $624 million dollars (Sorkin, 1982). A similar event today would have an even higher cost. Therefore, it is critical to study the magnetic field of active regions in order to understand, and ultimately predict, solar activity, like flares.

Since solar flares mostly occur in active regions, it is common in solar flare prediction research to first produce predictions for active regions. The differentiation of solar active regions often involves the use of sunspot classifications - Mount Wilson (Hale et al., 1919) and McIntosh (McIntosh, 1990) - which are still performed manually. These classes are based on human observations in the visible light spectrum. This leads to inference of the subjectivity of the experts. Moreover, the visible light spectrum provides very limited information regarding the critical properties of solar active regions. Today it is possible to automatize the classification of solar active regions, reducing the influence of human bias. This will allow to produce fast solar flare forecasting systems.

There is already a long history of research on solar flare prediction. One of the earliest flare-prediction systems is THEO (McIntosh, 1990). This system is based on the McIntosh classification, representing the morphological characteristics, and uses more than 500 decision rules to predict X-ray solar flares. The predictions from THEO were subjective, since the decision rules were incorporated by a human expert and had not been evaluated statistically (Qahwaji & Colak, 2007). Afterwards, a lot of other researchers

also based themselves on the McIntosh classification to construct a solar flare prediction method, but used more objective statistical methods (Bornmann & Shaw, 1994; Gallagher et al., 2002) and machine learning methods (Li et al., 2008; Qahwaji & Colak, 2007). Other researchers based themselves on the properties of the solar photospheric magnetic field instead (Barnes et al., 2007; Leka & Barnes, 2007; Wang et al., 2008).

Leka and Barnes (2003) pioneered the use of vector magnetic field data for flare prediction. Since then, numerous other researchers have followed and used a multitude of different vector magnetic field features and studied the relevance of them for flare prediction. For example, magnetic field topology parameters have been used by Schrijver (2007). Fisher et al. (2012) found that the integrated Lorentz force exerted by active regions can be related to the flaring activity as well. Also, the presence of polarity inversion lines is believed to be strongly correlated with flaring activity (Zirin & Wang, 1993), leading to the development of features measuring this (Jing et al., 2006; Mason & Hoeksema, 2010).

Since the launch of the Solar dynamics Observatory (SDO) in 2010, the magnetic features on the Sun have been continuously monitored by the Helioseismic and Magnetic Imager (HMI). This instrument maps the full-disk photospheric magnetic field with high cadence and continuity. (Bobra et al., 2014) developed the SHARP parameters from this data. These parameters are physical features calculated from the full-disk photospheric magnetic field for each active region identified on the disk. Sixteen of the SHARP parameters are features derived by Leka and Barnes (2003). Seven other parameters are adapted from Fisher et al. (2012).

With the launch of the HMI instrument at SDO and the development of the SHARP parameters, the amount of data available and the quality of the data has increased enormously, leading to the need for a more automated and faster approach. Barnes et al. (2007) were the first to use vector magnetograms to investigate solar flare forecasting using a statistical classifier. Bobra and Couvidat (2015) followed this with the first solar flare forecast using machine learning on features derived from a vector magnetogram. It was also the first time that a large data set of vector magnetograms had been used to forecast solar flares. They used 25 parameters of the SHARP vector magnetic field data and applied a support vector machine on them to predict X- and M-class flares.

Many new approaches have been developed to process the large amounts of solar observational data. Different statistical methods have been applied in many solar flare

prediction studies, e.g. (Bloomfield et al., 2012; Barnes et al., 2016; Mason & Hoeksema, 2010; Song et al., 2009). In the last 15 years machine learning methods have become increasingly popular for analysing the occurrence of solar flares. Multiple machine learning methods have been considered, including regression models (Lee et al., 2007; Song et al., 2009), k-Nearest Neighbors (Huang et al., 2013; Li et al., 2008; Nishizuka et al., 2017; Winter & Balasubramaniam, 2015), Bayesian networks (D. Yu et al., 2009), C4.5 decision trees (Huang et al., 2010; D. Yu et al., 2009), hybrid LASSO (Campi et al., 2019), support vector machines (SVM) (Bobra & Couvidat, 2015; Ilonidis et al., 2015; Nishizuka et al., 2017; Qahwaji & Colak, 2007; Raboonik et al., 2016; Sadykov & Kosovichev, 2017; Yuan et al., 2010), artificial neural networks (Ahmed et al., 2013; Colak & Qahwaji, 2009; Higgins et al., 2011; Li & Zhu, 2013; Nishizuka et al., 2018; Qahwaji & Colak, 2007), random forest (Florios et al., 2018; Liu et al., 2017), convolutional neural networks (CNN) (Huang et al., 2018; Park et al., 2018; Sun et al., 2022; Zheng et al., 2019), long-short term memory (LSTM) (Jiao et al., 2020; Sun et al., 2022) and ensemble learning (Colak & Qahwaji, 2009; Guerra et al., 2015, 2020; Huang et al., 2010; Sun et al., 2022).

The above studies are some examples of research that used machine learning to predict solar flares. More related to the study of (Bobra & Couvidat, 2015) and our study is the literature where machine learning methods are applied on the SHARP parameters, e.g. (Abduallah et al., 2020; Jiao et al., 2020; Liu et al., 2017; Ma et al., 2017; Sun et al., 2022; Tang et al., 2021). Abduallah et al. (2020) developed a machine-learning-as-a-service framework, DeepSun, for predicting solar flares on the Web, based on the SHARP parameters. Jiao et al. (2020) applied a mixed LSTM regression model on the SHARP parameters to identify the flare intensity, a continuous variable. Liu et al. (2017) applied a random forest algorithm to the SHARP data and predicted the occurrence of a certain class of flares in a given active region within 24 hours. Ma et al. (2017) clustered the SHARP features one-by-one, identified the most interesting ones, e.g. the ones that were able to find the most different clusters, and applied a decision tree on the clustering results to find the final binary flare prediction. Sun et al. (2022) applied two deep-learning methods, LSTM and CNN, on SHARP data from solar cycle 24 and SMARP data from solar cycle 23. Tang et al. (2021) used a combination of a deep neural network, a CNN and a bidirectional LSTM to predict whether a sunspot group will have a flare event $\leq$ C or $\leq$ M, based on the SHARP features.

As mentioned in the previous paragraph, Sun et al. (2022) used photospheric magnetic field data from both solar cycle 23 and 24, while the SHARP data only exist for solar cycle 24, since the HMI instrument did not exist before. However, the HMI instru-

ment had a precursor, MDI (Hoeksema et al., 1992) of SOHO. A new data set has been created by Bobra et al. (2021), called SMARPs. These are similar to SHARPs, but constructed from the solar images taken by MDI. As such, they attempt to extend backwards the SHARP database to the more active Solar Cycle 23. However, MDI's measurements are restricted to only the line-of-sight component, with a a lower spatial resolution (2" per pixel, in contrast to 0.5" per pixel for HMI), lower signal-to-noise ratio, and shorter cadence. As such, the SMARPs do not include as much information as the SHARPs and the data quality is lower (Sun et al., 2022).

A number of studies have investigated the importance of each of the SHARP parameters for solar flare prediction (Ran et al., 2022; Sinha et al., 2022; Zhang et al., 2022). They found that the most influential SHARP parameters are `TOTUSJH`, `TOTUSJZ`, `MEANPOT`, `TOTPOT`, `USFLUX` and `R_VALUE`. See Table 1 for the physical meaning of these parameters.

The SHARP parameters are selected based on physical understanding of what quantities are related to flare prediction. As such, they may be thought of as handcrafted features and are not objective. Recent research focuses on automatic characterization of useful features. There have been some studies combining the SHARP magnetic field parameters with features that are automatically generated from the solar images with machine learning methods, e.g (Chen et al., 2019; Jonas et al., 2018). Chen et al. (2019) compared the results of LSTM models trained on the SHARP data and on autoencoder-derived features and found that they were very similar in performance. Therefore, the autoencoder-derived features could be a viable alternative for the SHARP parameters.

Finally, most solar flare prediction methods focus on a binary classification. There are two ways to do this: flaring versus non-flaring active regions (Bloomfield et al., 2012; Bobra & Couvidat, 2015; Hamdi et al., 2017; Gallagher et al., 2002; Park et al., 2018; Wheatland, 2004) or strong versus weakly flaring active regions (Benvenuto et al., 2018; Huang et al., 2018; Nishizuka et al., 2018; Sun et al., 2022; Tang et al., 2021). In both cases, sometimes the intermediate C-flares are ignored, e.g. (Bobra & Couvidat, 2015; Sun et al., 2022). There are also some studies that make multiclass predictions of solar flares (Bloomfield et al., 2012; Colak & Qahwaji, 2009; Liu et al., 2017; Zheng et al., 2019), distinguishing between the different types of flares.

The goal of the present work is to classify the flaring activity of solar active regions, based only on the SHARP parameters extracted from the SDO HMI instrument. We apply rigorous and comprehensive pre-processing techniques to divide the useful from the

194 superfluous information, contained in the SHARP database. The novelty of this work
195 is the use of both supervised and unsupervised machine learning methods, where the lat-
196 ter does not take the flaring activity of the training data into account. This allows the
197 computer to extract patterns unknown to human experts. Many of the classification meth-
198 ods used in the literature are based on supervised learning only. We show how the un-
199 supervised classes that we obtain correlate with the flaring activity of active regions. An-
200 other aspect that differentiates our work is that we try to distinguish the different lev-
201 els of flaring activity, implementing a multi-class approach, whereas most studies focus
202 on binary classification.

203 The paper is structured as follows. Active regions and solar flares are briefly in-
204 troduced in section 2. Section 3 discusses the data used, followed by section 4, which ex-
205 plains the data processing methods and results. Sections 5 and 6 introduce the cluster-
206 ing methods and types of evaluation. The clustering results are shown in section 7, fol-
207 lowed by the discussion in section 8. Finally, section 9 summarizes the main conclusions
208 of the research results.

209 2 Active Regions and Solar Flares

210 Solar active regions are large areas on the Sun where the magnetic activity tem-
211 porarily and locally increases. The magnetic field there is complex and intense. Mag-
212 netic fields in active regions can be a thousand times stronger than the average solar mag-
213 netic field of a few Gauss (Sheeley, N.R., 2020). The number of active regions observed
214 in the solar disk varies over the course of the solar cycle and are most common during
215 its peak.

216 A solar flare is a sudden, intense brightening of a small area on the Sun, lasting min-
217 utes to a few hours. Flares occur in the solar corona when magnetic field lines of oppo-
218 site polarity are forced together, by the convective motion of their foot-points in the con-
219 vection zone, or by travelling coronal pressure waves. This causes magnetic reconnec-
220 tion, a sudden transformation of magnetic energy into kinetic and thermal energy. Streams
221 of highly energetic particles travel along magnetic field lines, generating high intensity
222 electromagnetic radiation on their path and during their interaction with matter. So-
223 lar flares typically erupt from solar active regions, because their complex and intense mag-
224 netic field is the perfect locus of magnetic reconnection (Priest & Forbes, 2002).

225 Flares are classified according to the strength of their soft X-ray emission, as recorded
226 by the GOES satellites located in geostationary orbit. The types of flares, ordered from
227 weakest to strongest are A, B, C, M and X. Strong solar flares occur very infrequently,

compared to weak solar flares. Therefore, solar flare data is by definition largely imbalanced. This always has to be taken into account during the processing of the data and the interpretation of the results.

3 Data Set

The open source data set of Angryk et al. (2020b) is used for this research. The authors developed a data set (henceforth called the Angryk data set), extracted from the Space Weather HMI Active Region Patch series (SHARP) (Bobra et al., 2014), integrated with information from solar flare catalogs. These SHARP patches and their magnetic field parameters are derived from solar photospheric vector magnetograms obtained by the Helioseismic and Magnetic Imager (HMI) from the Solar Dynamics Observatory (SDO). The HMI instrument provides information on the magnetic field in the solar photosphere. These observations are bundled in patches for each active region. Magnetic field parameters are extracted from these patches and integrated over the whole area. They give an indication of the magnetic activity of the complete patch.

The Angryk data set contains sixteen SHARP parameters and eight additional parameters proposed by Angryk et al. (2020a). These 24 parameters are listed in Table 1. The data set also contains parameters `BFLARE`, `CFLARE`, `MFLARE` and `XLFARE`. These express the number of flares of each flare class occurring at the time of measurement of the SHARP and therefore indicate the concurrent solar flare activity of that active region. For simplicity, in this work, each data point has been assigned to only one of four classes: No-flare, C-flare, M-flare or X-flare. These correspond to the strongest occurring flare originating from the active region at that time. The No-flare class signifies the flare-quiet instances, but also the weakest, A- and B-class, flares. This because the A- and B-flares are hard to distinguish against the background brightness of the Sun (Chen et al., 2019). The assignment of flare types to the data points leads to the following ratio: 2 602 509 No-flares, 6717 C-flares, 680 M-flares and 47 X-flares. The data was collected between May 2010 and December 2018. This corresponds with solar cycle 24 (December 2008 - December 2019) and includes the solar maximum in April 2014. This solar cycle was an unusual quiet one, and the data set contains only few strong flares. The Angryk data set is meant to serve as a benchmark data set for testing flare prediction algorithms (Angryk et al., 2020a).

Table 1: Magnetic field parameters from Angryk et al. (2020b). Parameters with * are derived by Angryk et al. (2020a), the others are contained in SHARP. Units from Liu et al. (2017) and SDO.

Parameters	**Description**	**Formula**
`ABSNJZH` [$10\mathrm{G}^2/\mathrm{m}$]	Absolute net current helicity	$H_{c_{abs}} \propto \lvert\sum B_z \cdot J_z\rvert$
`EPSX`* [-10^{-1}]	Sum normalized Lorentz force (X)	$\delta F_x \propto \dfrac{\sum B_x B_z}{\sum B^2}$
`EPSY`* [-10^{-1}]	Sum normalized Lorentz force (Y)	$\delta F_y \propto \dfrac{-\sum B_y B_z}{\sum B^2}$
`EPSZ`* [-10^{-1}]	Sum normalized Lorentz force (Z)	$\delta F_z \propto \dfrac{\sum (B_x^2 + B_y^2 - B_z^2)}{\sum B^2}$
`MEANALP` [1/Mm]	Mean twist parameter	$\alpha_{total} \propto \dfrac{\sum J_z \cdot B_z}{\sum B_z^2}$
`MEANGAM` [°]	Mean inclination angle	$\overline{\gamma} = \dfrac{1}{N}\sum \arctan\left(\dfrac{B_h}{B_z}\right)$
`MEANGBH` [G/Mm]	Mean horizontal field gradient	$\overline{\nabla B_h} = \dfrac{1}{N}\sum \sqrt{\left(\dfrac{\partial B_h}{\partial x} + \dfrac{\partial B_h}{\partial y}\right)}$
`MEANGBT` [G/Mm]	Mean total field gradient	$\overline{\nabla B_{tot}} = \dfrac{1}{N}\sum \sqrt{\left(\dfrac{\partial B}{\partial x} + \dfrac{\partial B}{\partial y}\right)}$
`MEANGBZ` [G/Mm]	Mean vertical field gradient	$\overline{\nabla B_z} = \dfrac{1}{N}\sum \sqrt{\left(\dfrac{\partial B_z}{\partial x} + \dfrac{\partial B_z}{\partial y}\right)}$
`MEANJZD` [mA/m^2]	Mean vertical current density	$\overline{J_z} \propto \dfrac{1}{N}\sum \left(\dfrac{\partial B_y}{\partial x} - \dfrac{\partial B_x}{\partial y}\right)$
`MEANJZH` [G^2/m]	Mean current helicity	$\overline{H_c} \propto \dfrac{1}{N}\sum B_z \cdot J_z$
`MEANPOT` [$10^3\mathrm{ergs}/\mathrm{cm}^3$]	Mean photospheric excess magnetic energy density	$\overline{\rho} \propto \dfrac{1}{N}\sum (\mathbf{B}^{Obs} - \mathbf{B}^{Pot})^2$
`MEANSHR` [°]	Mean shear angle	$\overline{\Gamma} = \dfrac{1}{N}\sum \arccos\left(\dfrac{\mathbf{B}^{Obs} \cdot \mathbf{B}^{Pot}}{\lvert B^{Obs}\rvert\lvert B^{Pot}\rvert}\right)$
`R_VALUE`* [Mx]	Total unsigned flux around high gradient polarity inversion lines	$\phi = \sum \lvert B_{los}\rvert \cdot dA$ (within R mask)
`SAVNCPP` [$10^{12}\mathrm{A}$]	Summed absolute value of net current per polarity	$J_{\Sigma z} \propto \left\lvert\sum^{B_z^+} J_z dA\right\rvert + \left\lvert\sum^{B_z^-} J_z dA\right\rvert$
`SHRGT45` [%]	Area with shear angle $> 45°$	$\dfrac{\text{Area with Shear} > 45°}{\text{Total Area}}$
`TOTBSQ`* [$10^{10}\mathrm{G}^2$]	Total magnitude of Lorentz force	$F \propto \sum B^2$
`TOTFX`* [-10^{23}dyne]	Sum X-component of Lorentz force	$F_x \propto \sum B_x B_z dA$
`TOTFY`* [-10^{23}dyne]	Sum Y-component of Lorentz force	$F_y \propto \sum B_y B_z dA$
`TOTFZ`* [-10^{23}dyne]	Sum Z-component of Lorentz force	$F_z \propto \sum \left(B_x^2 + B_y^2 - B_z^2\right) dA$
`TOTPOT` [$10^{23}\mathrm{ergs}/\mathrm{cm}^3$]	Total photospheric magnetic energy density	$\rho_{tot} \propto \sum \left(\overrightarrow{\mathbf{B}^{Obs}} - \overrightarrow{\mathbf{B}^{Pot}}\right)^2 dA$
`TOTUSJH` [$10^2\mathrm{G}^2/\mathrm{m}$]	Total unsigned current helicity	$H_{c_{total}} \propto \sum B_z \cdot J_z$
`TOTUSJZ` [$10^{12}\mathrm{A}$]	Total unsigned vertical current	$J_{z_{total}} = \sum \lvert J_z\rvert dA$
`USFLUX` [$10^{21}\mathrm{Mx}$]	Total unsigned flux	$\phi = \sum \lvert B_z\rvert dA$

4 Data Processing

Some pre-processing of the data set was carried out by Angryk et al. (2020a). Further processing includes outlier removal, data transformation and dimensionality reduction. These steps are explained in more detail in the following sections.

There is a large class imbalance present in the data set, with 2 602 509 No-flares, 6717 C-flares, 680 M-flares and only 47 X-flares. This class imbalance needs to be taken into account when processing the data. To reduce the impact of class imbalance, the No-flare class is randomly under-sampled to 50 000 No-flares. This is done by randomly selecting 50 000 data points from the 2 602 509 No-flares, without selecting the same data point twice.

The selected number of No-flares is determined after multiple tests of the autoencoding procedure, described in section 4.3.2, the most data-intensive processing step in this work. In short, in an autoencoder a compression and decompression of the data set is performed, and the active region parameters before and after the procedure should be exactly the same. We applied the procedure with different sample sizes. For each case the error is computed. When the sample size is too small, the error is large. Increasing the size of the sample reduces the error. A plot of the sample size versus the error presents an optimal inflection point, which in this work corresponds to the selected sample size: 50 000 data points are sufficient to obtain an accuracy comparable to the full 2 602 509 data points.

In section 4.4 we show how we handle additional class imbalances using over- and under-sampling techniques.

4.1 Outlier Removal

Multiple entries in the data set contain one or more empty parameters (NaN values). We eliminate from the original data set every entry where at least one of the parameters was empty. We also perform a detection and elimination of outliers. These were identified using the hierarchical clustering algorithm HDBSCAN. This method is able to automatically determine the optimal clustering of a cloud of points in an N-dimensional space. The points that are detached from the core cloud of points are identified as outliers. A more detailed explanation of HDBSCAN can be found in Campello et al. (2013).

We found 586 outliers. About 20% of the outliers come from HMI magnetogram images taken during rotation or re-positioning of the SDO spacecraft, causing distortions in the data. In addition, 36 outliers were identified and removed by hand, of which 33 were due to the same parameter, `MEANPOT`. The other three were due to the parameter

TOTFZ. The fact that they were missed by HDBSCAN is due to a combination of the standardization and some extreme outliers. The standardization transforms the data to zero mean and to unit variance. If there are a few extreme outliers, this will shift the majority of the data to very small values. This causes a difference of $\sim 2-3$ orders of magnitude with respect to the other parameters, hindering HDBSCAN to detect all outliers.

4.2 Data Transformation

To be able to differentiate groups of points in the parameter space, it is necessary to identify high concentrations of points that can be separated by a hyper-plane. Some of the parameters have a very skewed or narrow distribution, which will complicate the classification process. The log transformation is one of the most popular transformations to transform skewed data (Feng et al., 2014; West, 2022). This technique has been used abundantly in statistical analysis, for a review see (Curran-Everett, 2018; Keene, 1995), and has also been used with SHARP data in the past (Zhang et al., 2022). Therefore, we applied log transformations to some of the parameters to perform a rebinning of the data distributions. The transformations used are listed in Table 2.

Table 2: Data transformations used to expand narrow and/or skewed distributions.

Parameter (Table 1)	**Transformation**
TOTUSJH	$\ln(x + \lvert\min(x)\rvert + 0.01)$
TOTBSQ	$\ln(x + \lvert\min(x)\rvert + 0.01)$
TOTPOT	$\ln(x + \lvert\min(x)\rvert + 0.01)$
TOTUSJZ	$\ln(x + \lvert\min(x)\rvert + 0.01)$
ABSNJZH	$\ln(x + \lvert\min(x)\rvert + 0.01)$
SAVNCPP	$\ln(x + \lvert\min(x)\rvert + 0.01)$
USFLUX	$\ln(x + \lvert\min(x)\rvert + 0.01)$
MEANPOT	$\ln(x + \lvert\min(x)\rvert + 0.0001)$
TOTFZ	$\ln(-x + \lvert\max(x)\rvert + 0.01)$
TOTFY	$\ln(\lvert x\rvert)$
TOTFX	$\ln(\lvert x\rvert)$

4.3 Dimensionality Reduction

High-dimensional data is computationally expensive to process. If possible, it is important to reduce the number of dimensions. In addition, clustering methods and other techniques based on the calculation of distances in an Eulerian space are subject to the 'curse of dimensionality': in high dimensions every point tends to be equidistant to each

other point. Moreover, we want to reduce high correlations by removing redundant features. Figure 1 (left) illustrates the presence of correlations between the magnetic field parameters.The redundancy in magnetic field parameters, like SHARP, has been noted before by several studies, e.g. (Ahmed et al., 2013; Barnes et al., 2016; Bobra & Couvidat, 2015; Chen et al., 2019; Leka & Barnes, 2007). These studies all found that when a subset of only four to six parameters is used, the performance of their algorithm is similar compared to when the whole data set is used. Redundant features do not add relevant information and may hinder the learning algorithm, possibly causing overfitting (Yu & Liu, 2004). Therefore, we applied Common Factor Analysis (CFA) on the data set to remove the redundancy.

4.3.1 Common Factor Analysis

Common Factor Analysis (CFA) (Spearman, 1904) is a technique that searches for latent, unobserved variables, called factors, from a set of observed variables. For example, three observed variables (Y_1, Y_2, Y_3) can be modeled as a linear combination of two factors (F_1, F_2) and three error terms (e_1, e_2, e_3), as shown in Eq. (1) (Navlani, 2019). Applied to the data set, there are 24 equations for Y_1, Y_2, ..., Y_{24}. The amount of factors depends the amount of redundant features in the data, but can be at most 24.

$$\begin{aligned} Y_1 &= \theta_{10} + \theta_{11}F_1 + \theta_{12}F_2 + e_1 \\ Y_2 &= \theta_{20} + \theta_{21}F_1 + \theta_{22}F_2 + e_2 \\ Y_3 &= \theta_{30} + \theta_{31}F_1 + \theta_{32}F_2 + e_3 \end{aligned} \tag{1}$$

The package `FactorAnalyzer` of Biggs (2019) is used. The number of factors is determined with the help of *Horn's Parallel Analysis*. For more information about this method we would like to refer the reader to Horn (1965).

Figure 1 (right) shows the resulting factor loadings, corresponding to the θ parameters in Eq. 1. The loadings are a measure of how much a factor explains the associated magnetic field parameters. It is a correlation coefficient for the original parameters and the resulting factors and is defined like this:

$$X - M = LF - \varepsilon \tag{2}$$

where $X \in \mathbb{R}^{p\times n}$ is the observation matrix, $L \in \mathbb{R}^{p\times k}$ is the loading matrix, $F \in \mathbb{R}^{k\times n}$ is the factor matrix, $\varepsilon \in \mathbb{R}^{p\times n}$ is the error term matrix and $M \in \mathbb{R}^{p\times n}$ is the mean matrix, with p the number of parameters, n the number of observations and k the number of factors.

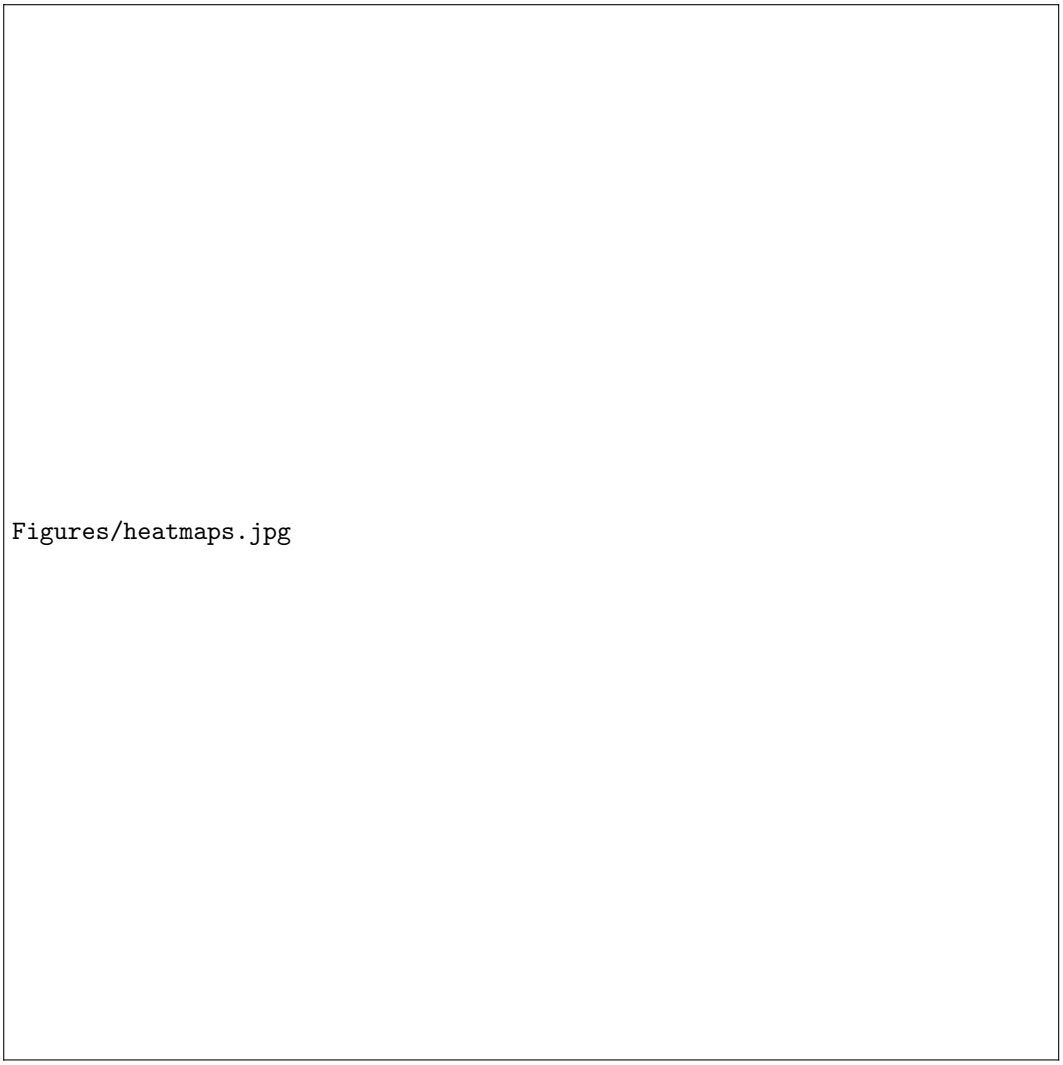


Figure 1: Left: Covariance matrix of the data set before applying CFA. Many parameters are strongly correlated with each other. Right: Heatmap of factor loadings of CFA.

4.3.2 Sparse Autoencoders

Makhzani and Frey (2014) show improvement in classification tasks when sparse data representations are used. To improve sparsity in our data set, we applied sparse autoencoders. These are able to transform the data into a higher dimensional space, where it is possible to create hyperplanes that allow to separate different clouds of points.

Sparse autoencoders are a type of unsupervised neural networks. For an explanation on neural networks, we refer the reader to the notes of Ng et al. (2011). The underlying mathematics of autoencoders are the same as for neural networks. The special property of autoencoders is that the target values ($\hat{X}$) are set equal to the input values (X) (Hinton & Salakhutdinov, 2006): $f : X \rightarrow \hat{X}$, where $X \approx \hat{X}$. The model learns

an approximation of the identity function. This may seem like a trivial task, but by reinforcing an information bottleneck in the hidden layers it is possible to discover interesting structures in the data.

In a basic (vanilla) autoencoder, also called encoder-decoder, $AE = \{f, f'\}$, the applied constraint consists to limit the number of nodes in an intermediary hidden layer to less than the number of input features of the model: the autoencoder functions are defined as $f : X \in \mathbb{R}^n \rightarrow Z \in \mathbb{R}^m$, followed by $f' : Z \in \mathbb{R}^m \rightarrow \hat{X} \in \mathbb{R}^n$, where $n > m$. A second autoencoder category corresponds to sparse autoencoders (Jiang et al., 2015), where the constraint is applied by forcing sparsity in the intermediary hidden layer. The dimension of the hidden layer is not smaller than the input layer. Instead, a sparsity constraint ensures that only a few hidden nodes are allowed to be active at the same time, i.e. most of the hidden nodes will have a value of zero. Sparse autoencoders provide an information bottleneck without having to reduce the number of nodes. This means that low dimensional data sets can be projected into higher dimensions where sparsity is encouraged, allowing for a better differentiation between different clouds of data points.

4.3.2.1 Implementation Details The sparse autoencoder is implemented using Python, together with libraries `Tensorflow` (Abadi et al., 2015) and `Keras` (Chollet et al., 2015). Any kind of neural network learns by minimizing a loss function, obtained by comparing the output of the model with the expected output. The loss function, Eq. 3, consists of two terms: (1) a reconstruction error and (2) a sparsity penalty. As reconstruction error the mean squared error is used. The sparsity penalty is a regularization acting on the activations of the hidden nodes, $a_i^{(h)} \in Z$, using the L1-norm. In the sparsity term of Eq. 3, λ is the pre-factor that determines the influence of the regularization.

$$L = \frac{1}{n}\sum_i (X_i - \hat{X}_i)^2 + \lambda \sum_i \left| a_i^{(h)} \right| \tag{3}$$

The autoencoder is optimized following the traditional error minimization techniques used in classical neural networks. As optimization algorithm, `Adam` (Kingma & Ba, 2015), is used. This is an extension to stochastic gradient descent that maintains separate dynamic learning rates for each parameter. For the Adam optimization we have to define the starting learning rate. The optimal value is determined with the method introduced by Smith (2017). This method trains a network starting with a low learning rate, which is exponentially increased throughout the epochs (training cycles). The optimal learning rate corresponds to the fastest decrease in loss throughout the training. An additional method to determine the optimal learning rate is to run the algorithm for multiple values of the

learning rate for a limited number of epochs, and to select the one with the lowest validation loss. The combination of these methods yields an optimal learning rate of 0.0005.

To determine the accuracy of the output, the R-squared metric is used, see Eq. 4.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(X_i - \hat{X}_i)^2}{\sum_{i=1}^{N}(X_i - \overline{X_i})^2} \text{ with } \overline{X_i} = \frac{1}{N}\sum_{j=1}^{N} X_j \tag{4}$$

To reduce the influence of the class imbalance, different weights have been assigned to the data samples corresponding to different flare classes. A weight of respectively 1, 4, 16 and 64 has been assigned to classes No-flare, C-flare, M-flare and X-flare.

Our data set is split into three sub-groups: 60% training, 20% validation and 20% testing data. The split is performed using stratification, which means that in each data portion the percentage of each flare type is preserved.

4.3.2.2 Architecture Optimization To find the optimal autoencoder architecture, three parameters need to be optimized: (1) the magnitude λ of the sparsity constraint, (2) the number of hidden nodes and (3) the activation function.

If the sparsity pre-factor is too high, all hidden nodes will tend to produce values of zero; if this parameter is too small, no sparsity will be introduced. The optimal value of λ is obtained by finding a balance between the level of sparsity and the activity on the hidden nodes. The pre-factor needs to be set to ensure that only part of the nodes (less than the number of input nodes) are active at the same time, without leaving inactive nodes. This balance is found for $\lambda = 0.1$.

The most adequate architecture is selected by comparing the loss function between the training and the validation set. The optimal architecture contains one hidden layer with seven hidden nodes and uses SELU (Klambauer et al., 2017) activation function.

4.3.2.3 Resulting Distributions The resulting optimal sparse autoencoder is used to increase the dimensionaliy, generating sparsity in the data set. The R-squared metric returns a value of 0.9942, indicating that the model is able to nearly perfectly mimic the original distributions. A two-dimensional projection of the distribution of each pair of parameters in the final data set is shown in Figure 2.

4.4 Data Sampling

Solar flare data is by definition largely imbalanced, since strong solar flares are scarce. Machine learning methods tend to favor the dominant class, here the non-flaring active

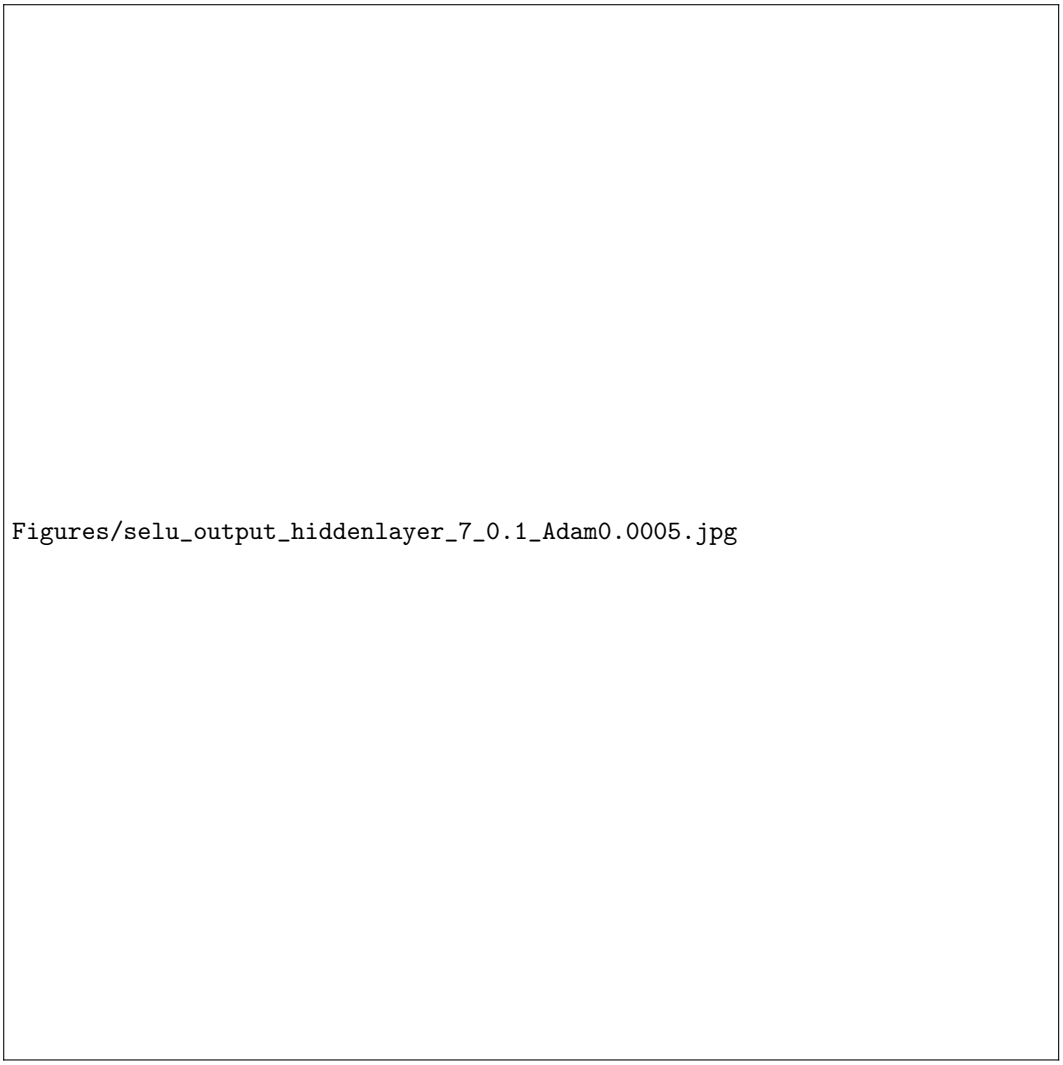


Figure 2: Distributions of the encoded data produced by the hidden layer of the sparse autoencoder. The autoencoder includes one hidden layer, with seven neurons, and SELU activation functions. The pre-factor λ for the activity regularization is set to 0.1. The parameters H1 until H7 denote the output learned by each of the seven hidden neurons of the sparse autoencoder.

regions. The four different flare activity classes are either over-sampled or under-sampled to construct a balanced data set. A random under-sampling of the No-flares was already presented in section 4, but the imbalance among flare classes is still large.

4.4.1 Random Sampling

Random sampling can be applied to either under-sample or over-sample data. The methods `RandomUnderSampler` and `RandomOverSampler` of the package `imbalanced-learn` (Lemaître et al., 2017) are used. Random under-sampling picks samples from the ma-

jority classes without replacement, while over-sampling picks samples from the minority classes with replacement. However, random over-sampling of the minority class leads to duplication, which might lead to overfitting. Therefore an alternative over-sampling method is used.

4.4.2 SMOTE Sampling

The alternative Synthetic Minority Over-sampling TEchnique (SMOTE) (Chawla et al., 2002) technique is also included in the `imbalanced-learn` package. SMOTE does not duplicate any samples, but generates new data points by randomly selecting a minority class instance (a), and finding its k nearest neighbors. Subsequently, one of those k neighbors (b) is chosen at random and a synthetic example is created at a random point on the line segment between the instance (a) and its selected neighbor (b).

4.4.3 Resulting Data Set

It has been shown by Chawla et al. (2002) that the combination of SMOTE and under-sampling performs better than plain under-sampling. In our work the majority classes, No-flare and C-flare, are randomly under-sampled, while the minority classes, M-flare and X-flare, are over-sampled with SMOTE. Every class is sampled to 6000 samples, making the data set balanced.

4.4.4 Justification Sampling

The sampled data set has to be representative of the real data set in order to be reliable. To check this, we compare the distribution functions for each parameter before and after sampling. Since the under-sampling of non-flaring and C-flaring active regions was random, and the amount of data sampled was sufficiently high (6000 samples), the distribution functions before and after sampling are almost identical. The minority classes, M- and X-flares, were largely over-sampled, which makes it less evident that the sampled data set is representative of the real data set. Figures 3 and 4 show the density functions for each of the seven parameters before and after over-sampling. The distributions of the M-flares before and after over-sampling are very similar. The width of the density distribution is the same and the peaks occur at the same values. The peaks are slightly overestimated by the over-sampling technique. The over-sampling of the X-flares does not preserve the data distribution equally well. The overall shape is preserved, but the width of the distribution is more narrow after over-sampling. This is due to the SMOTE algorithm. This algorithm creates a synthetic sample between two real data samples. Therefore, SMOTE never creates samples at the border of the distribution and over-estimates

the center. The peaks in the distribution are also overestimated more significantly than for the M-flares. However, given that the X-flares are over-sampled by a factor of 160, these effects are limited in magnitude and the overall distributions are similar. We recognize that the effects of the over-sampling method will influence our results. Since the distributions are overestimated in the peaks and underestimated at the borders, our method will be limited in its reliability at the borders.

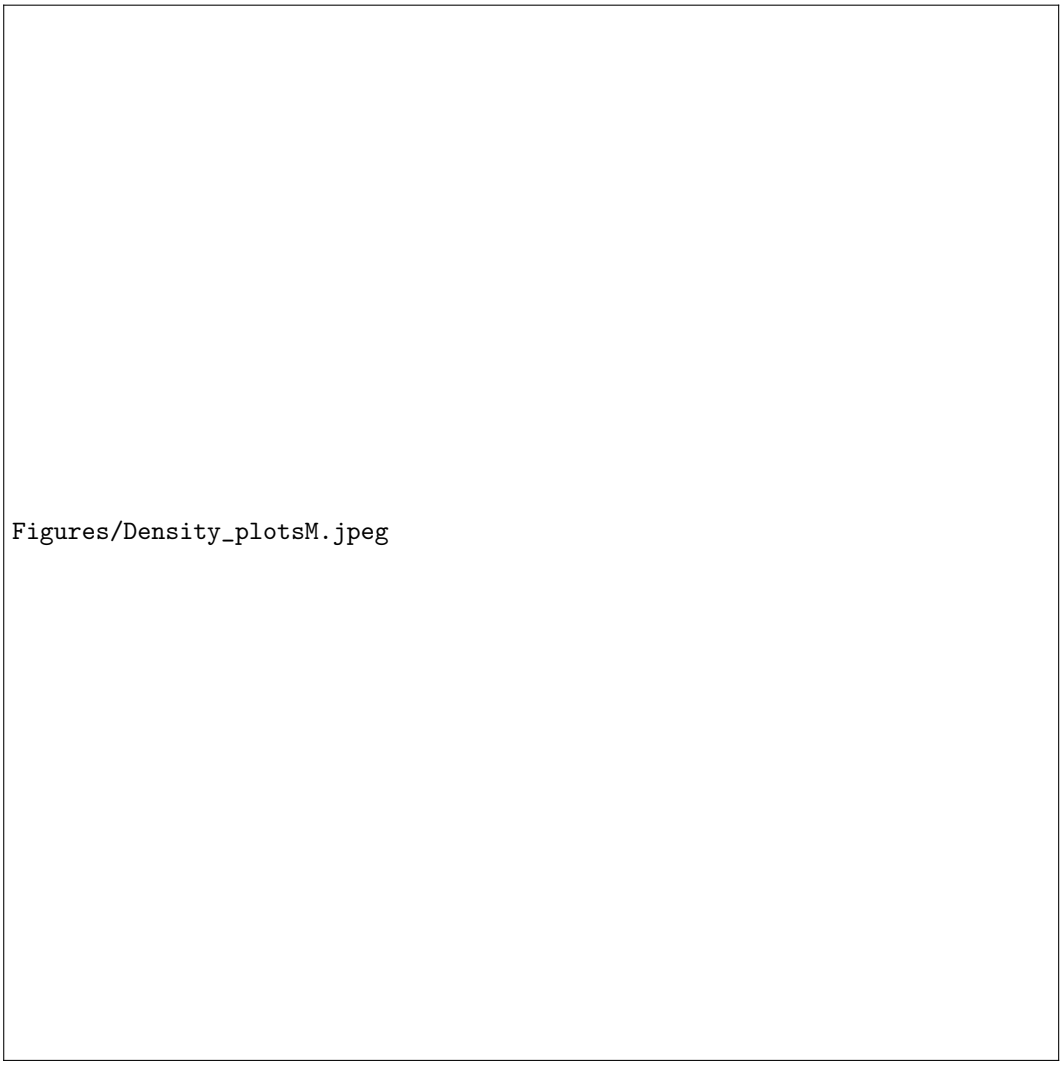


Figure 3: Density plots showing the distribution of M-flares for all seven parameters. The distribution is shown for the original data, and after over-sampling with SMOTE.

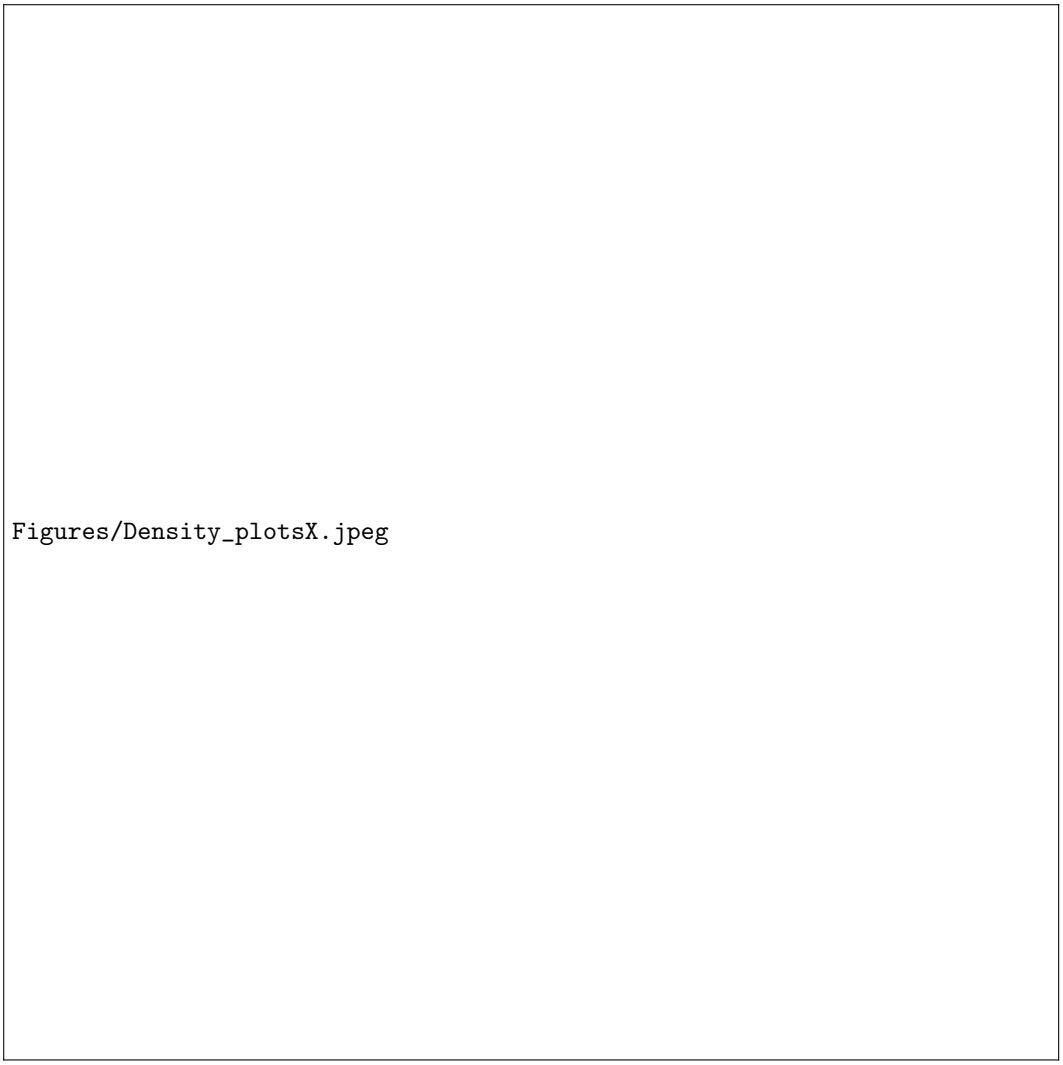


Figure 4: Density plots showing the distribution of X-flares for all seven parameters. The distribution is shown for the original data, and after over-sampling with SMOTE.

5 Classification of Active Regions

We tested multiple algorithms on the data set to classify the solar active regions based on their processed magnetic field parameters and found common aspects among the corresponding active regions.

5.1 k-Nearest Neighbors (supervised)

k-Nearest Neighbors (KNN), explained in e.g. Cunningham and Delany (2007), is a supervised and instance-based classifier. It assumes similar objects exist in close proximity to the evaluated data point. The class of a data point is determined based on the most frequent class among its k nearest neighbors.

The optimal number of neighbors k is the one that minimizes the error, the percentage of wrong predictions, while maintaining the ability to make accurate predictions on new data. The method minimizes the loss on the validation data, without overfitting on the training data. In general, lower k makes the predictions less stable. Increasing the number of neighbors makes the predictions more stable due to averaging and therefore more likely to produce reliable results. We selected the optimal k by performing the KNN algorithm for a range of k-values, fitting a fourth order polynomial to the corresponding error values and selecting the k corresponding to the minimum error.

5.2 K-means (unsupervised)

K-means (Lloyd, 1982; MacQueen, 1967) is a centroid-based clustering method, which assumes that the clusters are spherical and equally sized. Clustering is a machine learning method which groups data in subgroups that share similar properties (here, similar reduced magnetic field parameters). A good clustering method minimizes intra-cluster distances, while maximizing inter-cluster distances (Zhang & Tsai, 2005). The implementation and how clusters are defined differ from method to method. K-means works best when the clusters are equally dense and not too contaminated by noise or outliers. The clustering is achieved by iteratively assigning each data point to its nearest centroid and creating new centroids by computing the mean of each cluster.

The optimal number of clusters is determined by a *scree* plot (Cattell, 1966), where the 'knee' point is associated to the optimum value, and corresponds to the inflection point of the curve. The position of this 'knee' is determined through the *Kneedle algorithm* (Satopaa et al., 2011). The scree plot is configured by computing the error for different runs for a range of different number of clusters. A line is plotted between the first and last point of the curve and the distances between each point and the line are computed. The point with maximal distance between the two lines marks the maximum of curvature, i.e. the elbow.

5.3 Gaussian Mixture Models (unsupervised)

Gaussian Mixture Models (GMM) are a probabilistic clustering method, which assume that all data points are generated from a mixture of Gaussian distributions. It identifies for each data point the probabilities of belonging to each Gaussian distribution. This allows the detection of more elongated clusters. The Gaussian distributions are approximated by the Expectation-Maximization method (Dempster et al., 1977).

To determine the number of clusters for GMM, we used the gradient of the Bayesian Information Criterion (BIC). BIC (Schwarz, 1978) gives an estimation on how accurately

the model represents the existing data, with lower BIC value indicating a better estimation. BIC is defined in Eq. 5, with k the number of unknown model parameters (mean and variance for each cluster), n the number of samples and $\hat{L}$ the maximum likelihood.

$$BIC = k \ln n - 2 \ln \hat{L} \tag{5}$$

A high number of clusters corresponds to low BIC scores, but the error curve shows an inflection point. This point can be found by checking the gradient of BIC. The optimal number of clusters is the point where the gradient no longer changes, i.e. when the second derivative is zero (Lavorini, 2018).

6 Evaluation Methods

To determine the quality of an algorithm a good evaluation method is essential. An Area Under the Curve Receiver Operating Characteristics (AUC-ROC) plot (Fawcett, 2006) is a good evaluation technique for supervised classification methods, when the data is severely imbalanced (Brownlee, 2020).

ROC curves are in general used for binary classifications, but can be extended to multi-class data by using one-vs-rest for each class, providing one ROC curve per class. The macro-average is computed by averaging all ROC curves, treating all classes equally. The ROC curve is a visual measure of the predictive quality of a model, that visualizes the trade-off between sensitivity and specificity. The plot of a ROC curve displays the True Positive Rate (TPR), see Eq. 6, on the y-axis and the False Positive Rate (FPR), see Eq. 7, on the x-axis. These rates are computed for different thresholds. The threshold is the lowest probability necessary to be assigned to the positive cluster.

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{TN + FP} \tag{7}$$

An AUC score can be computed from the ROC, by computing the area under the curve. AUC is a measure of the ability of a classifier to distinguish between classes, where e.g. 0.7 means that in 70% of the cases the model is able to distinguish between the positive and the negative class (Narkhede, 2018).

In addition, the True Skill Statistic (TSS), also called the Hanssen score (Hanssen & Kuipers, 1965), will be computed for the supervised method, k-Nearest Neighbors, see

Eq. 8. The value of TSS lies between -1 and 1, with a higher value indicating a better forecast. It is one of the most used evaluation metrics to assess solar flare forecasts.

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN} = \frac{TP}{P} - \frac{FP}{N} \tag{8}$$

It is a lot harder to assess whether unsupervised methods perform well, because no labels are present. A viable alternative are validation methods that check whether there is a high separation between clusters and a high cohesion within the clusters. Examples of such metrics are the Calinsky-Harabasz (CH) coefficient (Caliński & Harabasz, 1974) and the Silhouette coefficient (SC) (Rousseeuw, 1987). The Calinski-Harabasz coefficient is defined as the ratio between the within-cluster dispersion and the between-cluster dispersion. This coefficient should be maximized. The Silhouette coefficient is computed, for each sample, using: (a) the mean inter-cluster distance, and (b) the mean nearest-cluster distance, see Eq. 9. The total Silhouette score is found by averaging over all samples. The best value is 1, the worst is -1 and values near 0 indicate that the clusters overlap. If the value is negative it is generally an indication that samples are assigned to the wrong cluster, as it is found that a different cluster is more similar.

$$SC = \frac{b - a}{max(a, b)} \tag{9}$$

7 Results

Figure 5 shows the mean value and standard deviation of each of the seven reduced parameters, for each flare class. In general, the parameters are very similar for all flaring active regions (C, M and X-flares). X-flare classes present only slight differences with respect to the other flaring classes. Parameters H2, H5 and H6 have a larger absolute mean value for these stronger flare classes. The mean value of the data without flares (No) is clearly different. It can be expected that flaring active regions will be distinguishable from non-flaring active regions, while distinguishing between the different flare classes will be more challenging with the available data. The same conclusions can be made from Figure 6, where the distribution function of magnetic field parameter `ABSNJZH` is shown for the different types of flares. The distribution of non-flaring active regions is clearly different from the distributions of flaring active regions. Meanwhile, the distributions of C-, M- and X-flares show a similar shape, only shifted a bit, but largely overlapping. We show here the magnetic field parameter with the largest differences in distributions between the types of flares.

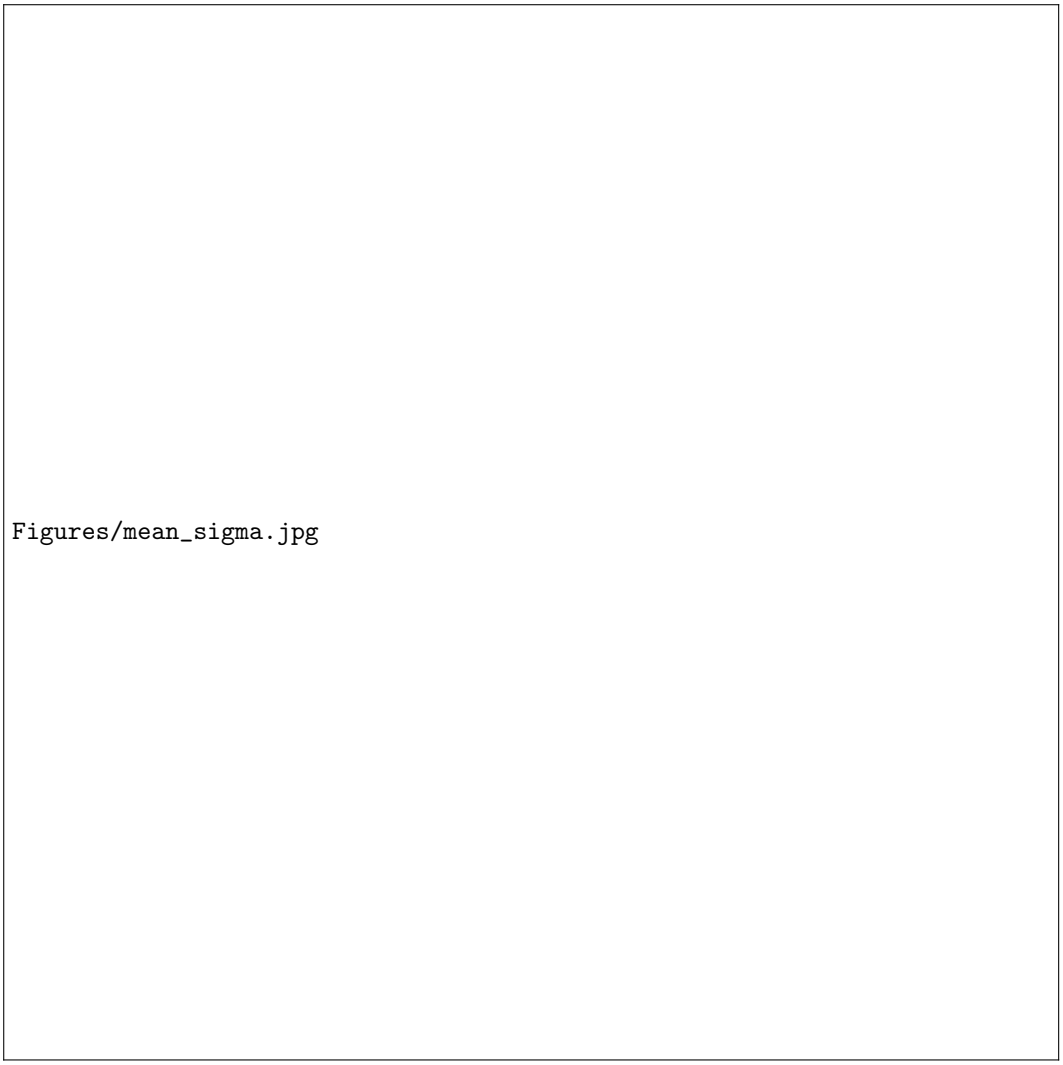


Figure 5: Mean and standard deviation of the features resulting from the sparse autoencoder, per flare label. The flaring data looks very similar, while the non-flaring data has distinct parameter values.

7.1 Supervised (KNN)

7.1.1 Hyperparameter selection

We based the hyperparameter selection for KNN on the unsampled data set, to avoid using under-/over-sampled data points. Performing the hyperparameter selection on the sampled data would yield an optimal number of neighbors of one, leading to unstable results. By applying the hyperparameter selection on the unsampled data set, we find an optimal number of neighbors of ten. To validate this selection method, KNN is conducted multiple times, testing the use of one, three, six and ten nearest neighbors. The resulting ROC curves are shown in Figure 7. These figures demonstrate that when more

Figure 6: Density distribution function per flare type, for magnetic field parameter ABSNJZH, with the true flare labels.

neighbors are taken into account, the results improve, producing a higher value for the area-under-the-curve. This is the case for the macro-average, as well as for the individual flare types. This shows that taking only one neighbor into account is not optimal.

7.1.2 Classification results

Figure 8 shows the normalized confusion matrix for the KNN results. On the x-axis are the predictions and on the y-axis the true classes. The following conclusions can be derived from the confusion matrix. Almost all X-flares are correctly identified. However, this is influenced by the over-sampling of the X-flares by a factor of ∼160. We minimized this influence by applying cross-validation. 87% of the true M-flares are correctly identified. This high percentage is also somewhat influenced by the over-sampling. When M-flares are misclassified, it is ∼ 37% of the time as an X-flare and ∼ 61% of the time as a C-flare. 76% of the non-flaring active regions are correctly classified as well. The non-flaring active regions are most often mistaken for C-flares. Finally, C-flares are hardest to identify, with only 50% of the active regions correctly identified as C-flares. They are mostly mistaken for their neighboring classes, in terms of X-ray flux strength. For

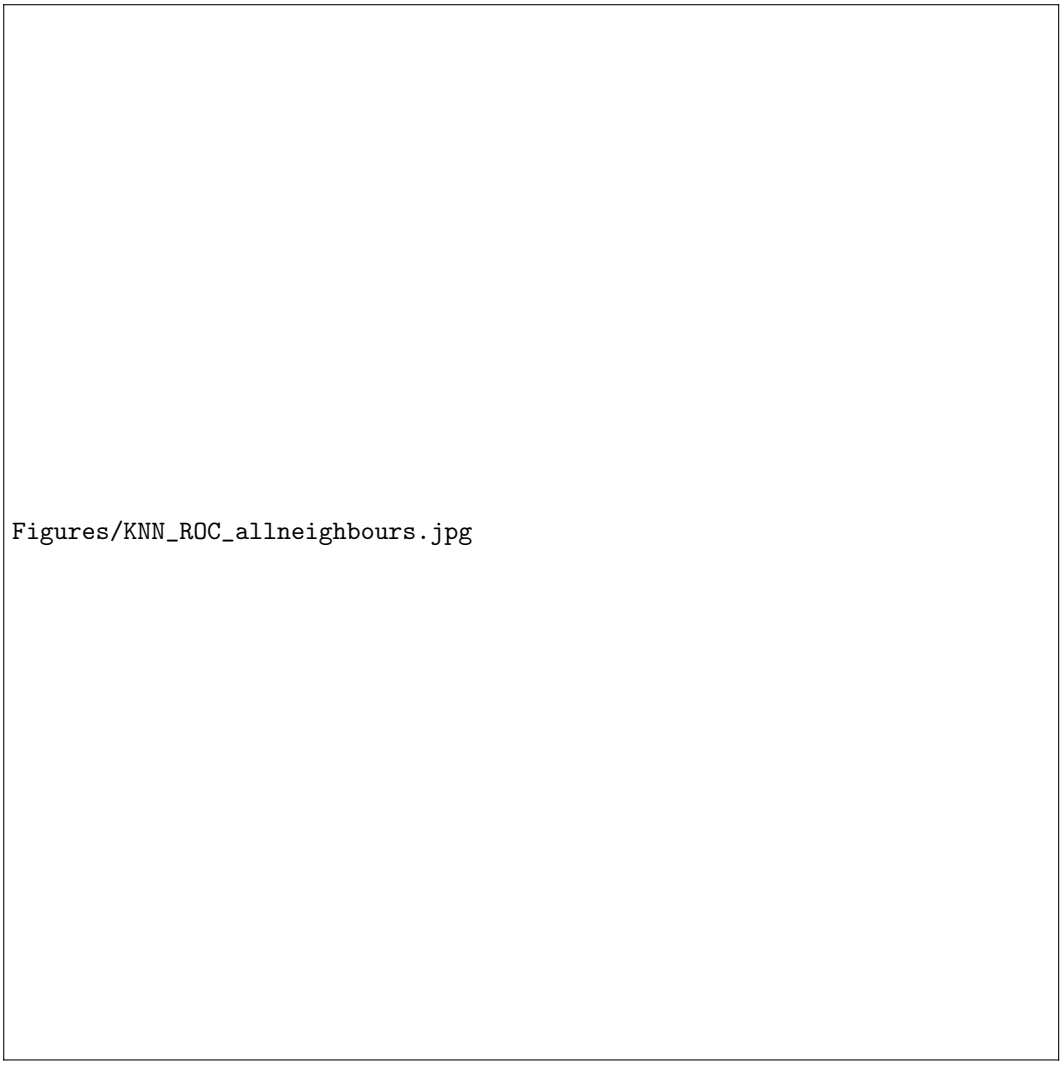


Figure 7: AUC ROC plot of KNN, performed on the sampled data set, for varying number of neighbors.

example, almost 30% of the C-flares are misclassified as M-flares. This indicates that the distributions of the magnetic field parameters of C- and M-flares are overlapping.

The TSS has been calculated for each of the flare types. A TSS of 0.94 is found for X-flares, 0.75 for M-flares, 0.42 for C-flares and 0.72 for the non-flaring active regions

7.2 Unsupervised (K-means + GMM)

Unsupervised methods are more useful in practice, since there is not always information present about the flaring nature of an active region. These methods only process the reduced magnetic field parameters and not the flare types. For both unsupervised methods used in this work (K-means and GMM), the number of clusters needs to

Figure 8: Normalized confusion matrix of the results of KNN with ten nearest neighbors taken into account.

be determined using a hyperparameter optimization technique, as described in sections 5.2 and 5.3. For K-means an optimal number of four (4) clusters is found, while GMM has an optimal number of three (3) clusters.

Table 3 shows the Calinski-Harabasz (Caliński & Harabasz, 1974) and Silhouette (Rousseeuw, 1987) coefficients, which evaluate the clusters found through K-means and GNN. The first one should be maximized, while the latter should be as close to 1 as possible. Both coefficients indicate that K-means does a better job at clustering the data. However, a relatively low Silhouette score of 0.25 indicates that the clusters are either not well separated or the points within a cluster are distributed far apart.

Table 3: Evaluation coefficients for K-means and GNN.

	K-means	**GMM**
Calinski-Harabasz	7506	1886
Silhouette	0.25	0.12

With unsupervised machine learning methods we can not construct a confusion matrix, since no labels are used. Instead, we constructed a matrix that shows the percentage of each flare included in each cluster in order to analyse the relation between the generated clusters and the flaring activity of the active regions. This matrix is shown for both K-means and GMM in Figure 9. Normalization is performed per flare type.

Analyzing the clusters of K-means learns us that 66% of the non-flaring active regions are included in Cluster 3. Cluster 3 also includes 17% of the C-flares, 12% of the M-flares and 5% of the X-flares. This cluster can be considered as one with mostly non- and weakly flaring active regions. An active region that is classified in Cluster 4 has a high probability to be an X-flare, since there are double as many X-flares as C- and M-flares, and only 12% non-flaring active regions present. If an active region is classified in Cluster 2, it is very probable to be flaring, because there are only 7% non-flaring active regions and $\sim$ 40% of each flaring type. However, nothing can be concluded about the type of flare from this cluster.

The resulting clusters found with GMM are visualized in Figure 9 (right). Cluster 3 contains 52% of the non-flaring active regions and 14 to 18% of flaring active regions. Meanwhile, Cluster 2 contains 34% of the non-flaring active regions and 8 to 18% of flaring active regions. Active regions that are classified into Cluster 2 and Cluster 3 are therefore mostly non-flaring. Cluster 1 contains only 14% of non-flaring active regions and 68 to 78% of each type of flare. This cluster can therefore with a high success rate identify flaring active regions. In all clusters found with GMM, the percentages of the different types of flaring active regions are very similar. Therefore, in contrast to K-means, the clustering with GMM is not able to distinguish the strength of the flares.

To get a more quantitative analysis, Figure 10 is a useful addition to Figure 9. They show the same data, but in Figure 10 the normalization is performed per cluster. Therefore, it can be used to determine the probability that an active region is of a certain flare type if it belongs to a certain cluster. We clarify this by giving a few examples. When an active regions is assigned to Cluster 3 by the K-means algorithm, it is with 66% probability non-flaring, with 17% a C-flare and with 12% probability an M-flare. An active regions that is assigned to Cluster 2 by K-means will with 94% probability (31% + 33% + 30%) be flaring, with approximately equal probability to be a C-flare, M-flare or X-flare. If an active region belongs to Cluster 1, found with GMM, there is only 6% chance that it is not flaring. However, when the active region is assigned to Cluster 2 or 3 by GMM, there is respectively a chance of 48% and 53% that it is not flaring.

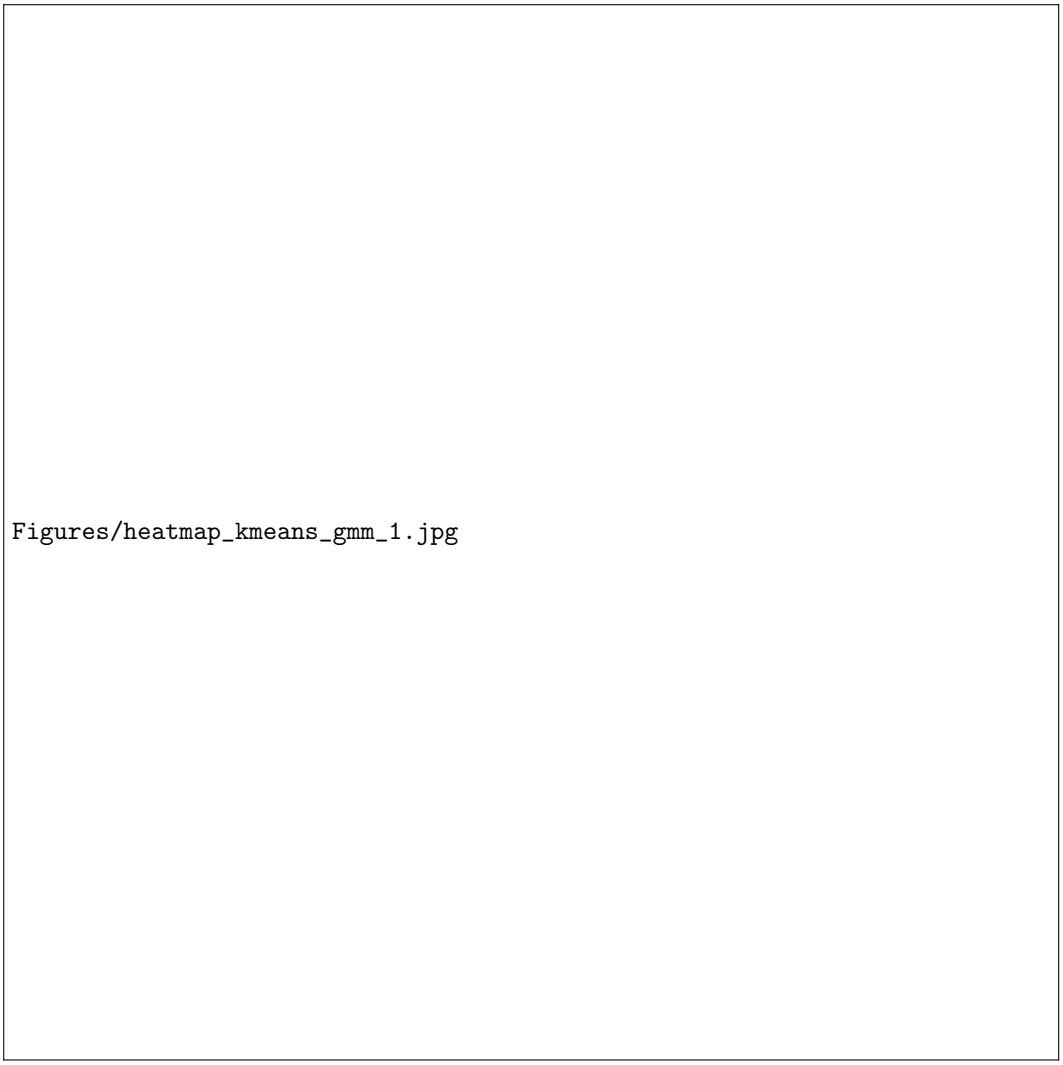


Figure 9: Clustering results of K-means (left) and GMM (right) on the sampled data set. The percentage of each flare included in each cluster is shown, where normalization is performed per flare type.

8 Discussion

The supervised method (KNN) has good performance for the M- and X-flares as well as for the non-flaring active regions with TSS scores of respectively 0.75, 0.93 and 0.72. The performance on the C-flares is less accurate, due to overlapping distributions of magnetic field parameters with M-flares and non-flaring active regions.

The unsupervised methods (K-means and GMM) can distinguish non-flaring from flaring active regions. Moreover, K-means is able to differentiate active regions producing strong flares from active regions producing weak flares, within some probability.

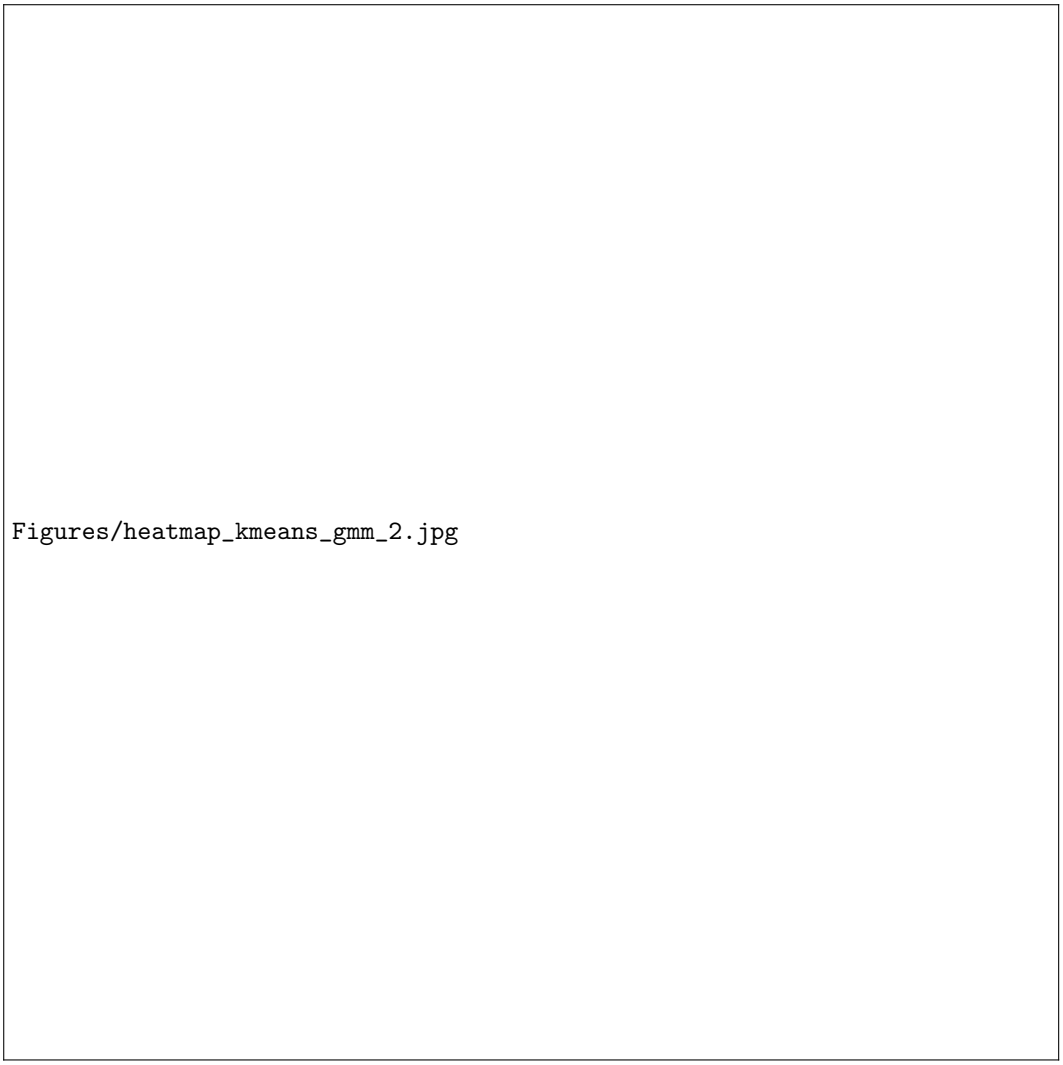


Figure 10: Clustering results of K-means (left) and GMM (right) on the sampled data set. The percentage of each flare included in each cluster is shown, where normalization is performed per cluster.

For processing purposes, unsupervised methods are more useful, since the type of flare will not be available. Out of the two unsupervised methods we studied, K-means stands out because it can distinguish between weakly and strongly flaring active regions.

Although K-means shows very promising results, it is not possible to differentiate C- from M-flares. The difficulty of differentiating between these flare types is partly inherent to the data, as predicted by analysis of Figure 5. The magnetic field parameters are very similar for C-flaring and M-flaring active regions. Moreover, the difficulty of differentiating C-, M- and X-flares is also caused by the arbitrary class boundaries, determined by their peak X-ray flux. A C9-flare is very similar to an M1-flare, but they were

for this work considered as strictly different classes of flares. The difference between background radiation (non-flaring active regions) and weak C-flares can be very small as well. The strength of flares is a continuous parameter, but was here treated as strictly discrete.

Discarding C-flares and distinguishing non-flaring active regions, M-flares and X-flares, as done in Bobra and Couvidat (2015) and Sun et al. (2022), might produce better results. Another improvement could be to discard non-flaring active regions and only consider flaring data. When both non-flaring and flaring data is taken into account, regions with complex and intense magnetic fields are compared against completely quiet regions. This gives the impression that all flaring active regions have similar properties. They will appear more distinct when only compared against each other. However, for flare prediction, in real-time data, the C-flares or non-flaring active regions can not be eliminated and need to be classified correctly as well.

Integrating more information into the analysis could also provide a clearer distinction. For example, the maximal difference in magnitude of the magnetic field over the active region could provide valuable information. In future research, the magnetic field parameters should be combined with other features, created through good feature engineering from the original images, for example through edge detection or with variational autoencoders. More data can be included by taking into account EUV observations, at multiple wavelenghts. Another extension to the use of the magnetic field parameters is to study their evolution, through time series. The data set we used is already in a time series format, but was not used like that for this research. The variation of the magnetic field in anticipation of the release of a flare will provide valuable information. The use of time series can also help to distinguish the natural variability of the solar magnetic field from a sudden change in the magnetic field due to flare formation.

9 Conclusion

Throughout this work detailed data cleaning and parameter transformation was conducted to enhance the quality of the Angryk data set and improve the classification results. Supervised clustering, with KNN, is able to distinguish the M- and X-flares, with respectively 99% and 87% correctly identified. However, only half of the C-flares are accurately classified. Unsupervised clustering, with K-means and GMM, identifies clusters with mainly non-flaring active regions and clusters with mainly flaring active regions. Moreover, K-means is able to distinguish X-flares from C-/M-flares. Unsupervised clustering is most useful for production purposes, since flare labels are not present. Therefore, we propose K-means as the most promising algorithm. For future projects, addi-

tional information should be included, like time series, different parameters - indicating e.g. the topology of active regions - or images of the active regions.

Open Research

This research uses the open source data set SWAN-SF of Angryk et al. (2020b). For more information we would like to refer the reader to the respective paper (Angryk et al., 2020a). The data is available for download through: `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EBCFKM`.

The code used to perform all data transformations and generate the clustering results is completely written in Python 3.10, and is accessible on Gitlab: `https://gitlab.com/hanneb/clustering_ar_sf_hbaeke.git` (Baeke, 2022).

Acknowledgments

This work was supported by the Research Foundation – Flanders (FWO), Frank De Winne PhD-aspirant grant (1SF8522N). This work has also received funding from the KULeuven Bijzonder Onderzoeksfonds (BOF) under the C1 project KULeuven Bijzonder Onderzoeksfonds (BOF), from the European Union's Horizon 2020 research and innovation program under grant agreement No. 955606 (DEEP-SEA). The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation – Flanders (FWO) and the Flemish Government. The data used in this project was constructed by Angryk et al. (2020b), which originates from the SHARP data series (Bobra et al., 2011), covering the period from 2010-05-01 until 2018-08-31, and the flare reports from GOES containing SSW and XRT flares. We would like to thank the assessors, Jasmina Magdalenic and Andrew Tkachenko, of the master thesis report for their feedback and suggestions. Their feedback led to an improved discussion section in this paper. We would also like to thank the reviewers for their constructive feedback. Finally, the authors declare that there are no conflicts of interest.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.* Retrieved from `https://www.tensorflow.org/` (Software available from tensorflow.org)

Abduallah, Y., Wang, J. T. L., Nie, Y., Liu, C., & Wang, H. (2020). DeepSun: Machine-Learning-as-a-Service for Solar Flare Prediction. *arXiv e-prints*, arXiv:2009.04238.

Ahmed, O. W., Qahwaji, R., Colak, T., Higgins, P. A., Gallagher, P. T., & Bloomfield, D. S. (2013, March). Solar Flare Prediction Using Advanced Feature Extraction, Machine Learning, and Feature Selection. , *283*(1), 157-175. doi: 10.1007/s11207-011-9896-1

Angryk, R., Martens, P., Aydin, B., Kempton, D., Mahajan, S., Basodi, S., ... Georgoulis, M. (2020a). Multivariate Time Series Dataset for Space Weather Data Analytics. *Scientific Data*, *7*, 227. doi: 10.1038/s41597-020-0548-x

Angryk, R., Martens, P., Aydin, B., Kempton, D., Mahajan, S., Basodi, S., ... Georgoulis, M. (2020b). *SWAN-SF [Dataset].* Retrieved from `https://doi.org/10.7910/DVN/EBCFKM` (Harvard Dataverse) doi: 10.7910/DVN/EBCFKM

Baeke, H. (2022). *Identifying Solar Flares by Clustering Active Regions [Code].* Retrieved from `https://gitlab.com/hanneb/clustering_ar_sf_hbaeke.git` (GitLab)

Barnes, G., Leka, K., Schumer, E., & Della-Rose, D. (2007). Probabilistic forecasting of solar flares from vector magnetogram data. *Space Weather*, *5*(9).

Barnes, G., Leka, K. D., Schrijver, C. J., Colak, T., Qahwaji, R., Ashamari, O. W., ... Wagner, E. L. (2016). A Comparison of Flare Forecasting Methods. I. Results from the All-Clear Workshop. *The Astrophysical Journal*, *829*(2), 89. Retrieved from `https://dx.doi.org/10.3847/0004-637X/829/2/89` doi: 10.3847/0004-637X/829/2/89

Benvenuto, F., Piana, M., Campi, C., & Massone, A. M. (2018, jan). A hybrid supervised/unsupervised machine learning approach to solar flare prediction. *The Astrophysical Journal*, *853*(1), 90. Retrieved from `https://doi.org/10.3847%2F1538-4357%2Faaa23c` doi: 10.3847/1538-4357/aaa23c

Biggs, J. (2019). *FactorAnalyzer release 0.3.2.* Retrieved from `https://github.com/EducationalTestingService/factor_analyzer`

Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. (2012, March). Toward Reliable Benchmarking of Solar Flare Forecasting Methods. , *747*(2), L41. doi: 10.1088/2041-8205/747/2/L41

Bobra, M., Hoeksema, J. T., Sun, X., & Turmon, M. (2011). SHARPs - A New Space Weather Data Product from SDO/HMI. In *Agu fall meeting abstracts* (Vol. 2011, p. SH51B-2006).

Bobra, M. G., & Couvidat, S. (2015). Solar Flare Prediction Using SDO/HMI Vector Magnetic Field Data with a Machine Learning Algorithm. *The Astrophysical Journal*, *798*(2), 135. Retrieved from `https://doi.org/10.1088/0004-637x/798/2/135` doi: 10.1088/0004-637x/798/2/135

Bobra, M. G., Sun, X., Hoeksema, J. T., Turmon, M., Liu, Y., Hayashi, K., ... Leka, K. D. (2014, September). The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs - Space-Weather HMI Active Region Patches. , *289*(9), 3549-3578. doi: 10.1007/s11207-014-0529-3

Bobra, M. G., Wright, P. J., Sun, X., & Turmon, M. J. (2021). SMARPs and SHARPs: Two Solar Cycles of Active Region Data. *The Astrophysical Journal Supplement Series*, *256*(2), 26. Retrieved from `https://doi.org/10.3847%2F1538-4365%2Fac1f1d` doi: 10.3847/1538-4365/ac1f1d

Bornmann, P. L., & Shaw, D. (1994, March). Flare Rates and the Mcintosh Active Region Classifications. , *150*(1-2), 127-146. doi: 10.1007/BF00712882

Brownlee, J. (2020). *Tour of Evaluation Metrics for Imbalanced Classification.* Retrieved from `https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/` (machinelearningmastery.com)

Caliński, T., & Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. *Communications in Statistics*, *3*(1), 1-27. Retrieved from `https://www.tandfonline.com/doi/abs/10.1080/03610927408827101` doi: 10.1080/03610927408827101

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in knowledge discovery and data mining* (pp. 160–172). Berlin, Heidelberg: Springer Berlin Heidelberg.

Campi, C., Benvenuto, F., Massone, A. M., Bloomfield, D. S., Georgoulis, M. K., & Piana, M. (2019, October). Feature Ranking of Active Region Source Properties in Solar Flare Forecasting and the Uncompromised Stochasticity of Flare Occurrence. , *883*(2), 150. doi: 10.3847/1538-4357/ab3c26

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, *1*(2), 245-276. Retrieved from `https://doi.org/10.1207/s15327906mbr0102_10` (PMID: 26828106) doi: 10.1207/s15327906mbr0102_10

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. Retrieved from `http://dx.doi.org/10.1613/jair.953` doi: 10.1613/jair.953

Chen, Y., Manchester, W. B., Hero, A. O., Toth, G., DuFumier, B., Zhou, T., ... Gombosi, T. I. (2019). Identifying Solar Flare Precursors Using Time Series of SDO/HMI Images and SHARP Parameters. *Space Weather*, *17*(10), 1404–1426. Retrieved from `http://dx.doi.org/10.1029/2019SW002214` doi:

776 10.1029/2019sw002214

777 Chollet, F., et al. (2015). *Keras.* `https://keras.io`.

778 Cinto, T., Gradvohl, A. L. S., Coelho, G. P., & da Silva, A. E. A. (2020). A
779 Framework for Designing and Evaluating Solar Flare Forecasting Systems.
780 *Monthly Notices of the Royal Astronomical Society*, *495*(3), 3332–3349.
781 Retrieved from `http://dx.doi.org/10.1093/mnras/staa1257` doi:
782 10.1093/mnras/staa1257

783 Colak, T., & Qahwaji, R. (2009, June). Automated Solar Activity Prediction:
784 A hybrid computer platform using machine learning and solar imaging for
785 automated prediction of solar flares. *Space Weather*, *7*(6), S06001. doi:
786 10.1029/2008SW000401

787 Cunningham, P., & Delany, S. (2007). k-Nearest Neighbour Classifiers. *Technical Re-*
788 *port UCD-CSI-2007-4*.

789 Curran-Everett, D. (2018, 06). Explorations in statistics: the log transfor-
790 mation. *Advances in physiology education*, *42*, 343-347. doi: 10.1152/
791 advan.00018.2018

792 Dang, T., Li, X., Luo, B., Li, R., Zhang, B., Pham, K., ... Wang, Y. (2022).
793 Unveiling the Space Weather During the Starlink Satellites Destruction
794 Event on 4 February 2022. *Space Weather*, *20*(8), e2022SW003152. doi:
795 10.1029/2022SW003152

796 Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from
797 Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Soci-*
798 *ety: Series B (Methodological)*, *39*(1), 1–22.

799 Eastwood, J., Biffis, E., Hapgood, M., Green, L., Bisi, M., Bentley, R., ... Bur-
800 nett, C. (2017, 02). The economic impact of space weather: Where do we
801 stand?: The economic impact of space weather. *Risk Analysis*, *37*. doi:
802 10.1111/risa.12765

803 Fawcett, T. (2006). Introduction to ROC analysis. *Pattern Recognition Letters*, *27*,
804 861-874. doi: 10.1016/j.patrec.2005.10.010

805 Feng, C., Hongyue, W., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. (2014, 04). Log-
806 transformation and its implications for data analysis. *Shanghai archives of psy-*
807 *chiatry*, *26*, 105-9. doi: 10.3969/j.issn.1002-0829.2014.02.009

808 Fisher, G. H., Bercik, D. J., Welsch, B. T., & Hudson, H. S. (2012). Global forces in
809 eruptive solar flares: The lorentz force acting on the solar atmosphere and the
810 solar interior. *Solar Physics*, *277*, 59–76.

811 Florios, K., Kontogiannis, I., Park, S.-H., Guerra, J. A., Benvenuto, F., Bloomfield,
812 D. S., & Georgoulis, M. K. (2018, February). Forecasting Solar Flares Us-

ing Magnetogram-based Predictors and Machine Learning. , *293*(2), 28. doi: 10.1007/s11207-018-1250-4

Gallagher, P. T., Moon, Y. J., & Wang, H. (2002, September). Active-Region Monitoring and Flare Forecasting I. Data Processing and First Results. , *209*(1), 171-183. doi: 10.1023/A:1020950221179

Guerra, J. A., Murray, S. A., Bloomfield, D. S., & Gallagher, P. T. (2020, July). Ensemble forecasting of major solar flares: methods for combining models. *Journal of Space Weather and Space Climate*, *10*, 38. doi: 10.1051/swsc/2020042

Guerra, J. A., Pulkkinen, A., & Uritsky, V. M. (2015, October). Ensemble forecasting of major solar flares: First results. *Space Weather*, *13*(10), 626-642. doi: 10.1002/2015SW001195

Hale, G. E., Ellerman, F., Nicholson, S. B., & Joy, A. H. (1919). The Magnetic Polarity of Sun-Spots. *The Astrophysical Journal*, *49*, 153. doi: 10.1086/142452

Hamdi, S. M., Kempton, D., Ma, R., Boubrahimi, S. F., & Angryk, R. A. (2017). A time series classification-based approach for solar flare prediction. In *2017 ieee international conference on big data (big data)* (p. 2543-2551). doi: 10.1109/BigData.2017.8258213

Hanssen, A., & Kuipers, W. (1965). *On the Relationship Between the Frequency of Rain and Various Meteorological Parameters: (with Reference to the Problem Ob Objective Forecasting)*. Staatsdrukkerij- en Uitgeverijbedrijf. Retrieved from `https://books.google.be/books?id=nTZ8OgAACAAJ`

Higgins, P. A., Gallagher, P. T., McAteer, R. T. J., & Bloomfield, D. S. (2011, June). Solar magnetic feature detection and tracking for space weather monitoring. *Advances in Space Research*, *47*(12), 2105-2117. doi: 10.1016/j.asr.2010.06.024

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, *313*(5786), 504–507. Retrieved from `https://science.sciencemag.org/content/313/5786/504` doi: 10.1126/science.1127647

Hoeksema, T., Scherrer, P., Bush, R., Title, A., & Tarbell, T. (1992, 12). The solar oscillations investigation: Michelson doppler imager.

Horn, J. (1965). A rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika*, *30*, 179-185.

Huang, X., Wang, H., Xu, L., Liu, J., Li, R., & Dai, X. (2018, March). Deep Learning Based Solar Flare Forecasting Model. I. Results for Line-of-sight Magnetograms. , *856*(1), 7. doi: 10.3847/1538-4357/aaae00

Huang, X., Yu, D., Hu, Q., Wang, H., & Cui, Y. (2010, 05). Short-term solar flare

prediction using predictor teams. *Solar Physics*, *263*, 175-184. doi: 10.1007/s11207-010-9542-3

Huang, X., Zhang, L., Wang, H., & Li, L. (2013, January). Improving the performance of solar flare prediction using active longitudes information. , *549*, A127. doi: 10.1051/0004-6361/201219742

Ilonidis, S., Bobra, M. G., & Couvidat, S. (2015). Solar Flare Forecasting Using Time Series of SDO/HMI Vector Magnetic Field Data and Machine Learning Methods. In *AAS/AGU triennial earth-sun summit* (Vol. 1, p. 311.03).

Jiang, N., Rong, W., Peng, B., Nie, Y., & Xiong, Z. (2015). An Empirical Analysis of Different Sparse Penalties for Autoencoder in Unsupervised Feature Learning. In *Ijcnn* (p. 1-8). IEEE. Retrieved from `http://dblp.uni-trier.de/db/conf/ijcnn/ijcnn2015.html#JiangRPNX15`

Jiao, Z., Sun, H., Wang, X., IV, W., Gombosi, T., Hero, A., & Chen, Y. (2020). Solar Flare Intensity Prediction With Machine Learning Models. *Space Weather*, *18*. doi: 10.1029/2020SW002440

Jing, J., Song, H., Abramenko, V., Tan, C., & Wang, H. (2006, June). The Statistical Relationship between the Photospheric Magnetic Parameters and the Flare Productivity of Active Regions. , *644*(2), 1273-1277. doi: 10.1086/503895

Jonas, E., Bobra, M., Shankar, V., Todd Hoeksema, J., & Recht, B. (2018). Flare Prediction Using Photospheric and Coronal Image Data. *Solar Physics*, *293*(3), 48. doi: 10.1007/s11207-018-1258-9

Kawabata, Y., Iida, Y., Doi, T., Akiyama, S., Yashiro, S., & Shimizu, T. (2018, dec). Statistical Relation between Solar Flares and Coronal Mass Ejections with Respect to Sigmoidal Structures in Active Regions. *The Astrophysical Journal*, *869*(2), 99. Retrieved from `https://dx.doi.org/10.3847/1538-4357/aaebfc` doi: 10.3847/1538-4357/aaebfc

Keene, O. N. (1995). The log transformation is special. *Statistics in medicine*, *14 8*, 811-9.

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *CoRR*, *abs/1412.6980*.

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-Normalizing Neural Networks. In *Proceedings of the 31st international conference on neural information processing systems* (p. 972–981). Red Hook, NY, USA: Curran Associates Inc.

Knipp, D., Ramsay, A., Beard, E., Boright, A., Cade, T., Hewins, I., ... Smart, D. (2016). The May 1967 Great Storm and Radio Disruption Event: Extreme Space Weather and Extraordinary Responses: May 1967 Solar and Geomag-

netic Storm. *Space Weather*, *14*. doi: 10.1002/2016SW001423

Lavorini, V. (2018). *Gaussian Mixture Model clustering: how to select the number of components.* Retrieved from `https://towardsdatascience.com/gaussian-mixture-model-clusterization-how-to-select-the-number-of-components-clusters-553bef45f6e4` (Towards Data Science)

Lee, J.-Y., Moon, Y.-J., Kim, K.-S., Park, Y., & Fletcher, A. (2007, 12). Prediction of daily maximum x-ray flux using multilinear regression and autoregressive time-series methods. *Journal of The Korean Astronomical Society*, *40*, 99-106. doi: 10.5303/JKAS.2007.40.4.099

Leka, K. D., & Barnes, G. (2003, October). Photospheric Magnetic Field Properties of Flaring versus Flare-quiet Active Regions. I. Data, General Approach, and Sample Results. , *595*(2), 1277-1295. doi: 10.1086/377511

Leka, K. D., & Barnes, G. (2007, February). Photospheric Magnetic Field Properties of Flaring versus Flare-quiet Active Regions. IV. A Statistically Significant Sample. , *656*(2), 1173-1186. doi: 10.1086/510282

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, *18*(17), 1-5. Retrieved from `http://jmlr.org/papers/v18/16-365.html`

Li, R., Cui, Y., He, H., & Wang, H. (2008, November). Application of support vector machine combined with K-nearest neighbors in solar flare and solar proton events forecasting. *Advances in Space Research*, *42*(9), 1469-1474. doi: 10.1016/j.asr.2007.12.015

Li, R., & Zhu, J. (2013, September). Solar flare forecasting based on sequential sunspot data. *Research in Astronomy and Astrophysics*, *13*(9), 1118-1126. doi: 10.1088/1674-4527/13/9/010

Liu, C., Deng, N., Wang, J., & Wang, H. (2017). Predicting Solar Flares Using SDO/HMI Vector Magnetic Data Products and the Random Forest Algorithm. *The Astrophysical Journal*, *843*, 104. doi: 10.3847/1538-4357/aa789b

Liu, J., Wang, W., Qian, L., Lotko, W., Burns, A. G., Pham, K., ... Wilder, F. (2021). Solar Flare Effects in the Earth's Magnetosphere. *Nature Physics*, *17*(7), 807-812. doi: 10.1038/s41567-021-01203-5

Lloyd, S. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129-137. doi: 10.1109/TIT.1982.1056489

Ma, R., Boubrahimi, S. F., Hamdi, S. M., & Angryk, R. A. (2017). Solar flare prediction using multivariate time series decision trees. In *2017 ieee international conference on big data (big data)* (p. 2569-2578). doi: 10.1109/BigData.2017

.8258216

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In *Berkeley symposium on mathematical statistics and probability* (p. 281-297).

Makhzani, A., & Frey, B. J. (2014). k-Sparse Autoencoders. *CoRR*, *abs/1312.5663*.

Mason, J. P., & Hoeksema, J. T. (2010, oct). Testing automated solar flare forecasting with 13 years of michelson doppler imager magnetograms. *The Astrophysical Journal*, *723*(1), 634. Retrieved from `https://dx.doi.org/10.1088/0004-637X/723/1/634` doi: 10.1088/0004-637X/723/1/634

McIntosh, P. S. (1990). The Classification of Sunspot Groups. *Solar Physics*, *125*(2), 251-267. doi: 10.1007/BF00158405

Mikaelian, T. (2009). *Spacecraft Charging and Hazards to Electronics in Space.* arXiv. Retrieved from `https://arxiv.org/abs/0906.3884` doi: 10.48550/ARXIV.0906.3884

Narkhede, S. (2018). *Understanding AUC-ROC Curve.* Retrieved from `https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5` (TowardsDataScience)

Navlani, A. (2019). *Introduction to Factor Analysis in Python.* (DataCamp)

Ng, A., et al. (2011). Sparse Autoencoder. *CS294A Lecture notes*, *72*(2011), 1–19.

Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., & Ishii, M. (2018, May). Deep Flare Net (DeFN) Model for Solar Flare Prediction. , *858*(2), 113. doi: 10.3847/1538-4357/aab9a7

Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Watari, S., & Ishii, M. (2017, February). Solar Flare Prediction Model with Three Machine-learning Algorithms using Ultraviolet Brightening and Vector Magnetograms. , *835*(2), 156. doi: 10.3847/1538-4357/835/2/156

Park, E., Moon, Y.-J., Shin, S., Yi, K., Lim, D., Lee, H., & Shin, G. (2018, December). Application of the Deep Convolutional Neural Network to the Forecast of Solar Flare Occurrence Using Full-disk Solar Magnetograms. , *869*(2), 91. doi: 10.3847/1538-4357/aaed40

Priest, E. R., & Forbes, T. G. (2002). The Magnetic Nature of Solar Flares. *Astronomy and Astrophysics Reviews*, *10*(4), 313-377. doi: 10.1007/s001590100013

Pulkkinen, A., Lindahl, S., Viljanen, A., & Pirjola, R. (2005). Geomagnetic Storm of 29-31 October 2003: Geomagnetically Induced Currents and their Relation to Problems in the Swedish High-Voltage Power Transmission System. *Space Weather*, *3*. doi: 10.1029/2004SW000123

Qahwaji, R., & Colak, T. (2007, March). Automatic Short-Term Solar Flare Predic-

tion Using Machine Learning and Sunspot Associations. , *241*(1), 195-211. doi: 10.1007/s11207-006-0272-5

Raboonik, A., Safari, H., Alipour, N., & Wheatland, M. S. (2016, dec). Prediction of solar flares using unique signatures of magnetic field images. *The Astrophysical Journal*, *834*(1), 11. Retrieved from https://dx.doi.org/10.3847/1538-4357/834/1/11 doi: 10.3847/1538-4357/834/1/11

Ran, H., Liu, Y. D., Guo, Y., & Wang, R. (2022). *Relationship between Successive Flares in the Same Active Region and Space-Weather HMI Active Region Patch (SHARP) Parameters.* arXiv. Retrieved from https://arxiv.org/abs/2207.07254 doi: 10.48550/ARXIV.2207.07254

Redmon, R., Seaton, D., Steenburgh, R., He, J., & Rodriguez, J. (2018). September 2017's Geoeffective Space Weather and Impacts to Caribbean Radio Communications During Hurricane Response. *Space Weather*, *16*. doi: 10.1029/2018SW001897

Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. Retrieved from https://www.sciencedirect.com/science/article/pii/0377042787901257 doi: https://doi.org/10.1016/0377-0427(87)90125-7

Sadykov, V. M., & Kosovichev, A. G. (2017, nov). Relationships between characteristics of the line-of-sight magnetic field and solar flare forecasts. *The Astrophysical Journal*, *849*(2), 148. Retrieved from https://doi.org/10.3847%2F1538-4357%2Faa9119 doi: 10.3847/1538-4357/aa9119

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *2011 31st international conference on distributed computing systems workshops* (p. 166-171). doi: 10.1109/ICDCSW.2011.20

Schrijver, C. J. (2007). A characteristic magnetic field pattern associated with all major solar flares and its use in flare forecasting. *The Astrophysical Journal*, *655*(2), L117.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461 – 464. Retrieved from https://doi.org/10.1214/aos/1176344136 doi: 10.1214/aos/1176344136

Sheeley, N.R. (2020). Solar Magnetic Field. *AccessScience*.

Sinha, S., Gupta, O., Singh, V., Lekshmi, B., Nandy, D., Mitra, D., ... Pal, S. (2022). A Comparative Analysis of Machine-learning Models for Solar Flare Forecasting: Identifying High-performing Active Region Flare Indicators. *The*

Astrophysical Journal, *935*(1), 45. Retrieved from https://dx.doi.org/10.3847/1538-4357/ac7955 doi: 10.3847/1538-4357/ac7955

Smith, L. N. (2017). *Cyclical Learning Rates for Training Neural Networks.* arXiv. Retrieved from https://arxiv.org/abs/1506.01186 doi: 10.48550/ARXIV.1506.01186

Song, H., Tan, C., Jing, J., Wang, H., Yurchyshyn, V., & Abramenko, V. (2009, January). Statistical Assessment of Photospheric Magnetic Features in Imminent Solar Flare Predictions. , *254*(1), 101-125. doi: 10.1007/s11207-008-9288-3

Sorkin, A. (1982). Economic Aspects of Natural Hazards. In *Economic aspects of natural hazards.* Lexington Books; distributed Gower.

Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292. Retrieved from http://www.jstor.org/stable/1412107

Sun, Z., Bobra, M. G., Wang, X., Wang, Y., Sun, H., Gombosi, T., ... Hero, A. (2022). Predicting Solar Flares Using CNN and LSTM on Two Solar Cycles of Active Region Data. *The Astrophysical Journal*, *931*(2), 163. Retrieved from https://doi.org/10.3847\%2F1538-4357\%2Fac64a6 doi: 10.3847/1538-4357/ac64a6

Tang, R., Liao, W., Chen, Z., Zeng, X., song Wang, J., Luo, B., ... Wu, Z. (2021, dec). Solar flare prediction based on the fusion of multiple deep-learning models. *The Astrophysical Journal Supplement Series*, *257*(2), 50. Retrieved from https://dx.doi.org/10.3847/1538-4365/ac249e doi: 10.3847/1538-4365/ac249e

Wang, H. N., Cui, Y. M., Li, R., Zhang, L. Y., & Han, H. (2008, November). Solar flare forecasting model supported with artificial neural network techniques. *Advances in Space Research*, *42*(9), 1464-1468. doi: 10.1016/j.asr.2007.06.070

West, R. M. (2022). Best practice in statistics: The use of log transformation. *Annals of Clinical Biochemistry*, *59*(3), 162-165. Retrieved from https://doi.org/10.1177/00045632211050531 (PMID: 34666549) doi: 10.1177/00045632211050531

Wheatland, M. S. (2004, July). A Bayesian Approach to Solar Flare Prediction. , *609*(2), 1134-1139. doi: 10.1086/421261

Winter, L., & Balasubramaniam, K. (2015, 04). Using the maximum x-ray flux ratio and x-ray background to predict solar flare class. *Space Weather*, *13*. doi: 10.1002/2015SW001170

Yu, & Liu, H. (2004). Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research*, *5*, 1205–1224.

Yu, D., Huang, X., Hu, Q., Zhou, R., Wang, H., & Cui, Y. (2009, dec). Short-term solar flare prediction using multiresolution predictors. *The Astrophysical Journal*, *709*(1), 321. Retrieved from https://dx.doi.org/10.1088/0004-637X/709/1/321 doi: 10.1088/0004-637X/709/1/321

Yuan, Y., Shih, F. Y., Jing, J., & Wang, H.-M. (2010, August). Automated flare forecasting using a statistical learning technique. *Research in Astronomy and Astrophysics*, *10*(8), 785-796. doi: 10.1088/1674-4527/10/8/008

Zhang, D., & Tsai, J. J. P. (2005). *Machine Learning Applications in Software Engineering*. World Scientific. Retrieved from https://www.worldscientific.com/doi/abs/10.1142/5700 doi: 10.1142/5700

Zhang, H., Li, Q., Yang, Y., Jing, J., Wang, J. T. L., Wang, H., & Shang, Z. (2022). Solar Flare Index Prediction Using SDO/HMI Vector Magnetic Data Products with Statistical and Machine-learning Methods. *The Astrophysical Journal Supplement Series*, *263*(2), 28. Retrieved from https://doi.org/10.3847%2F1538-4365%2Fac9b17 doi: 10.3847/1538-4365/ac9b17

Zheng, Y., Li, X., & Wang, X. (2019, November). Solar Flare Prediction with the Hybrid Deep Convolutional Neural Network. , *885*(1), 73. doi: 10.3847/1538-4357/ab46bd

Zirin, H., & Wang, H. (1993, June). Narrow lanes of transverse magnetic field in sunspots. , *363*(6428), 426-428. doi: 10.1038/363426a0

Figure 1.

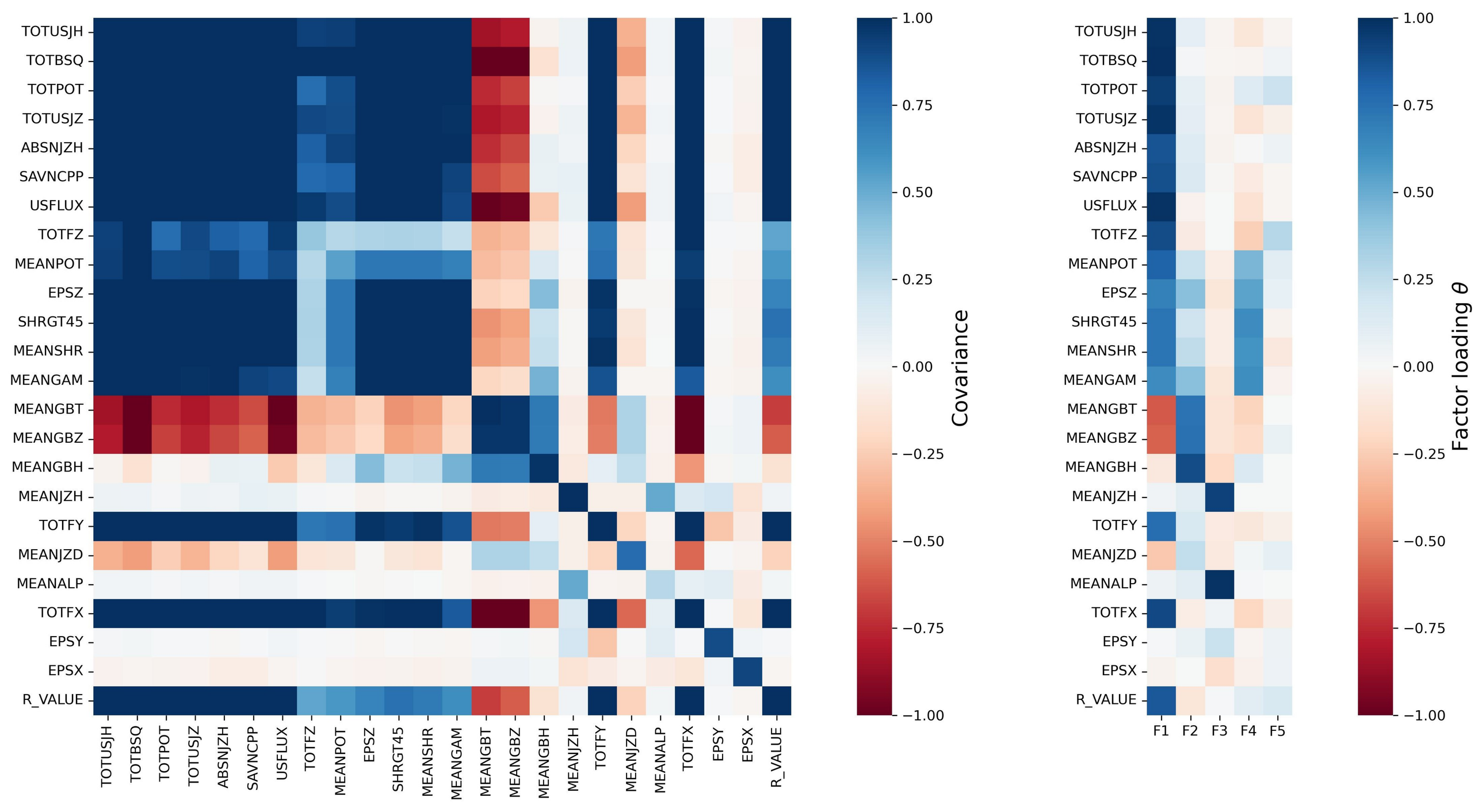

TOTUSJH
TOTBSQ
TOTPOT
TOTUSJZ
ABSNJZH
SAVNCPP
USFLUX
TOTFZ
MEANPOT
EPSZ
SHRGT45
MEANSHR
MEANGAM
MEANGBT
MEANGBZ
MEANGBH
MEANJZH
TOTFY
MEANJZD
MEANALP
TOTFX
EPSY
EPSX
R_VALUE
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Covariance
F1
F2
F3
F4
F5
Factor loading θ

Figure 2.

H1 H2 H3 H4 H5 H6 H7

H1 H2 H3 H4 H5 H6 H7

Figure 3.

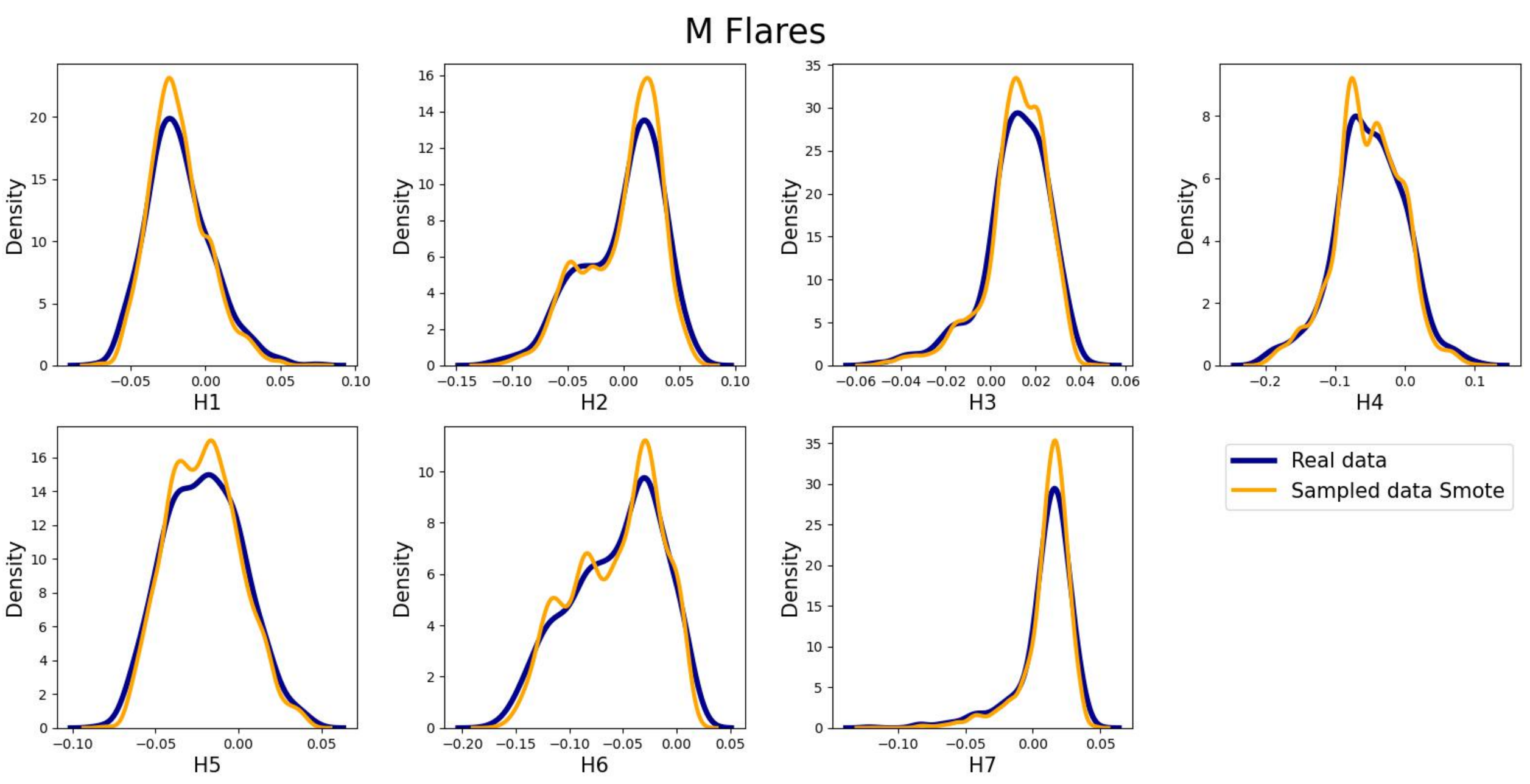

M Flares
Density
H1
H2
H3
H4
H5
H6
H7
Real data
Sampled data Smote

Figure 4.

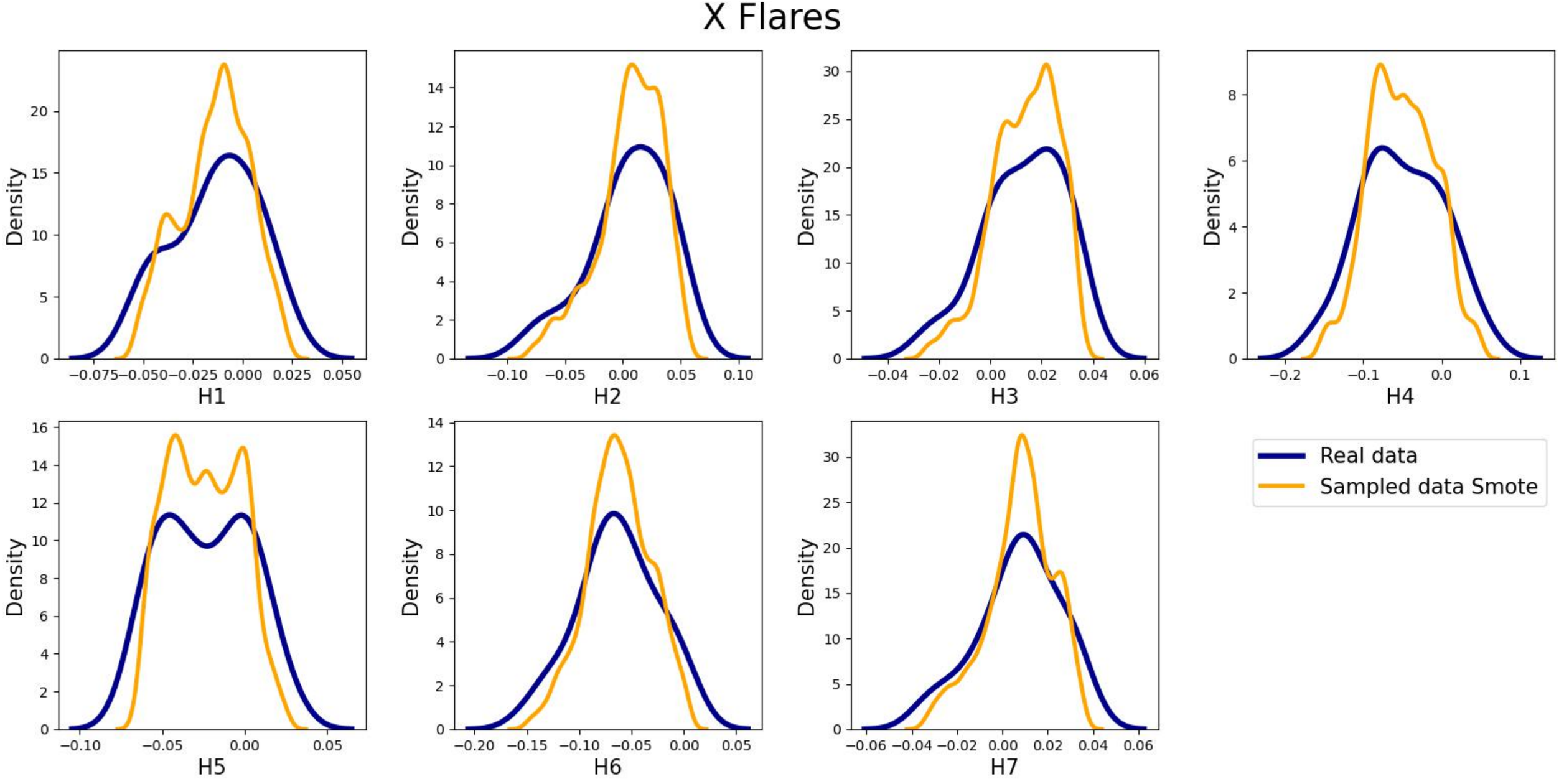

X Flares
Density
H1
H2
H3
H4
H5
H6
H7
Real data
Sampled data Smote

Figure 5.

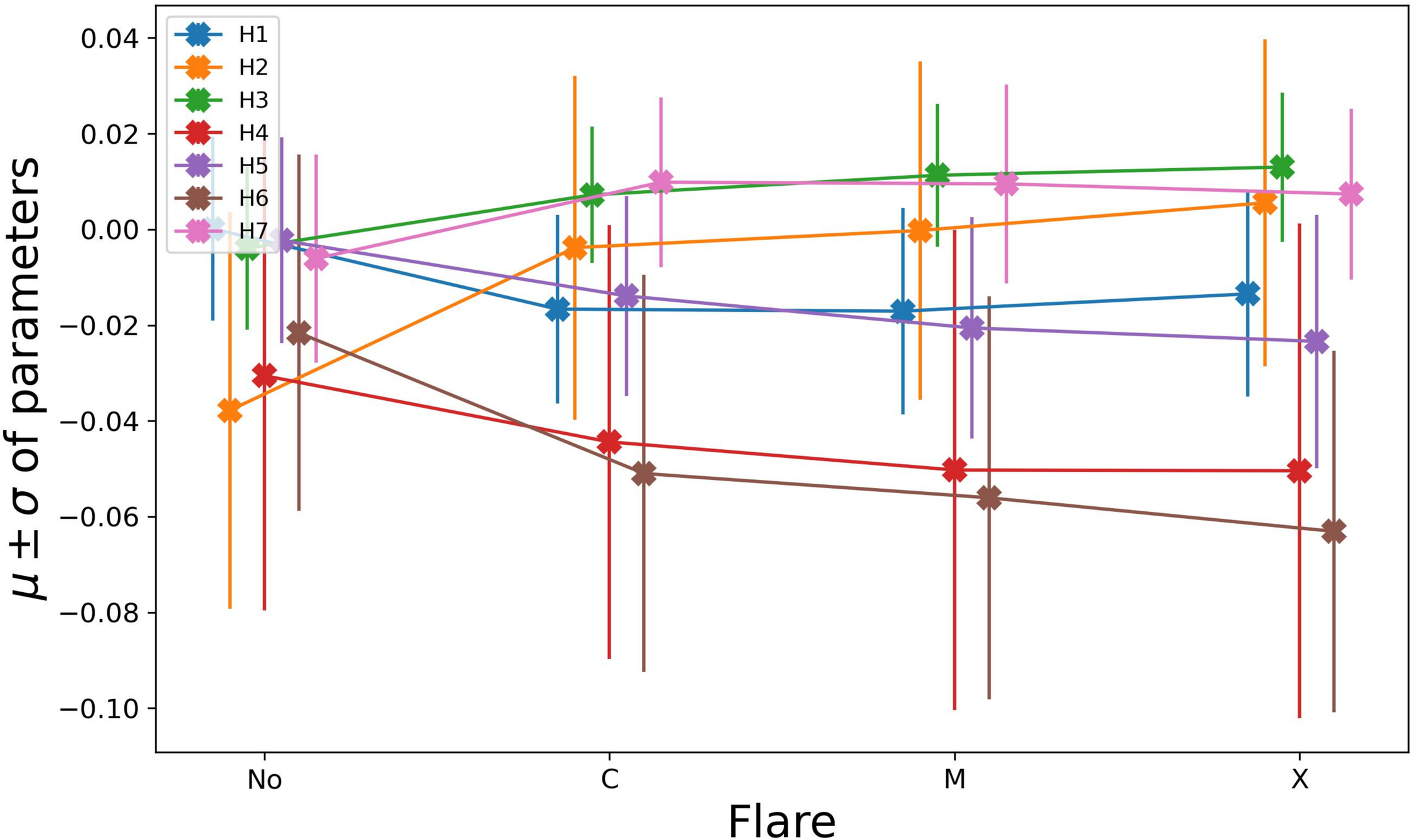

H1
H2
H3
H4
H5
H6
H7
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
$\mu \pm \sigma$ of parameters
No
C
M
X
Flare

Figure 6.

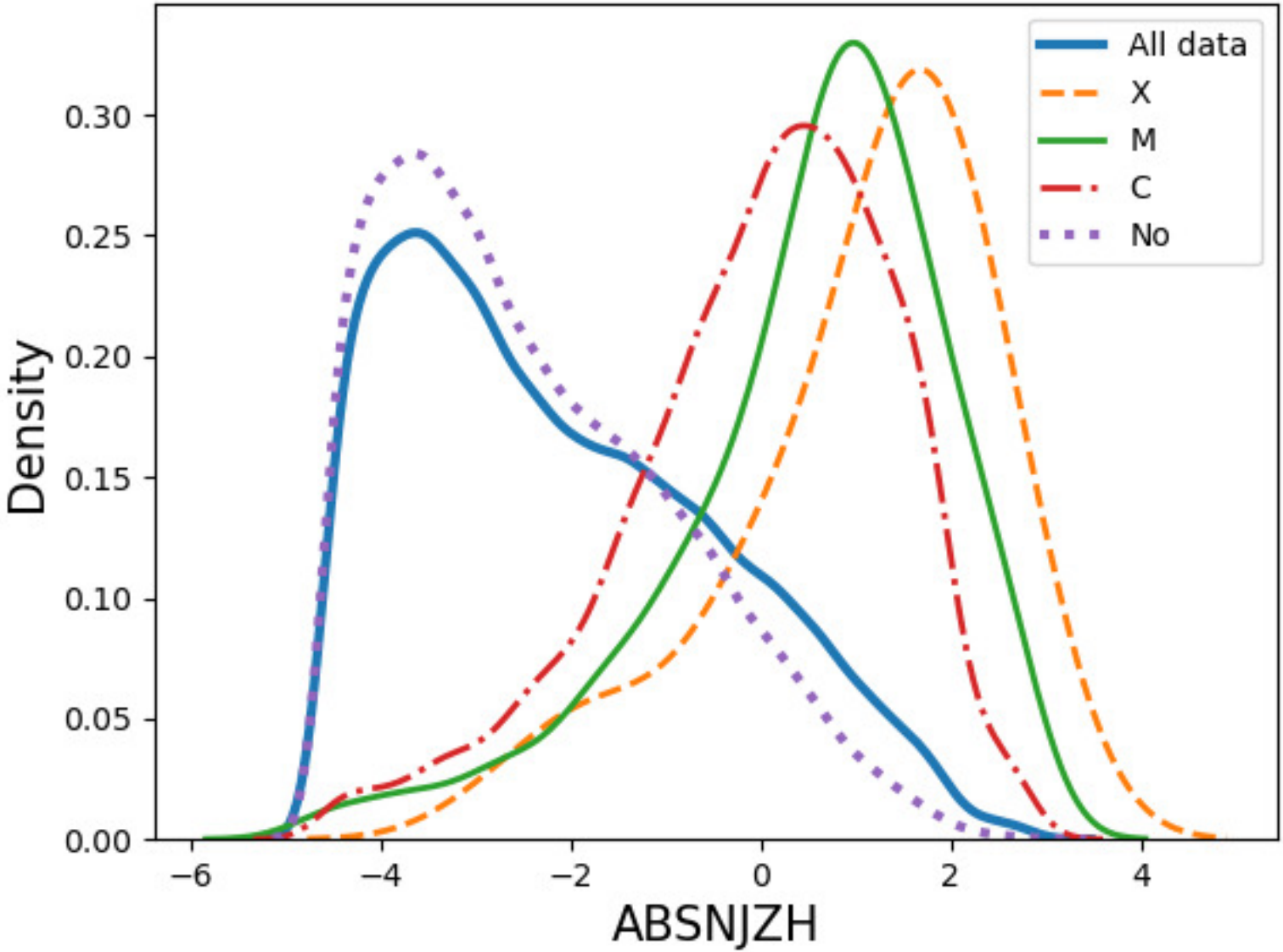
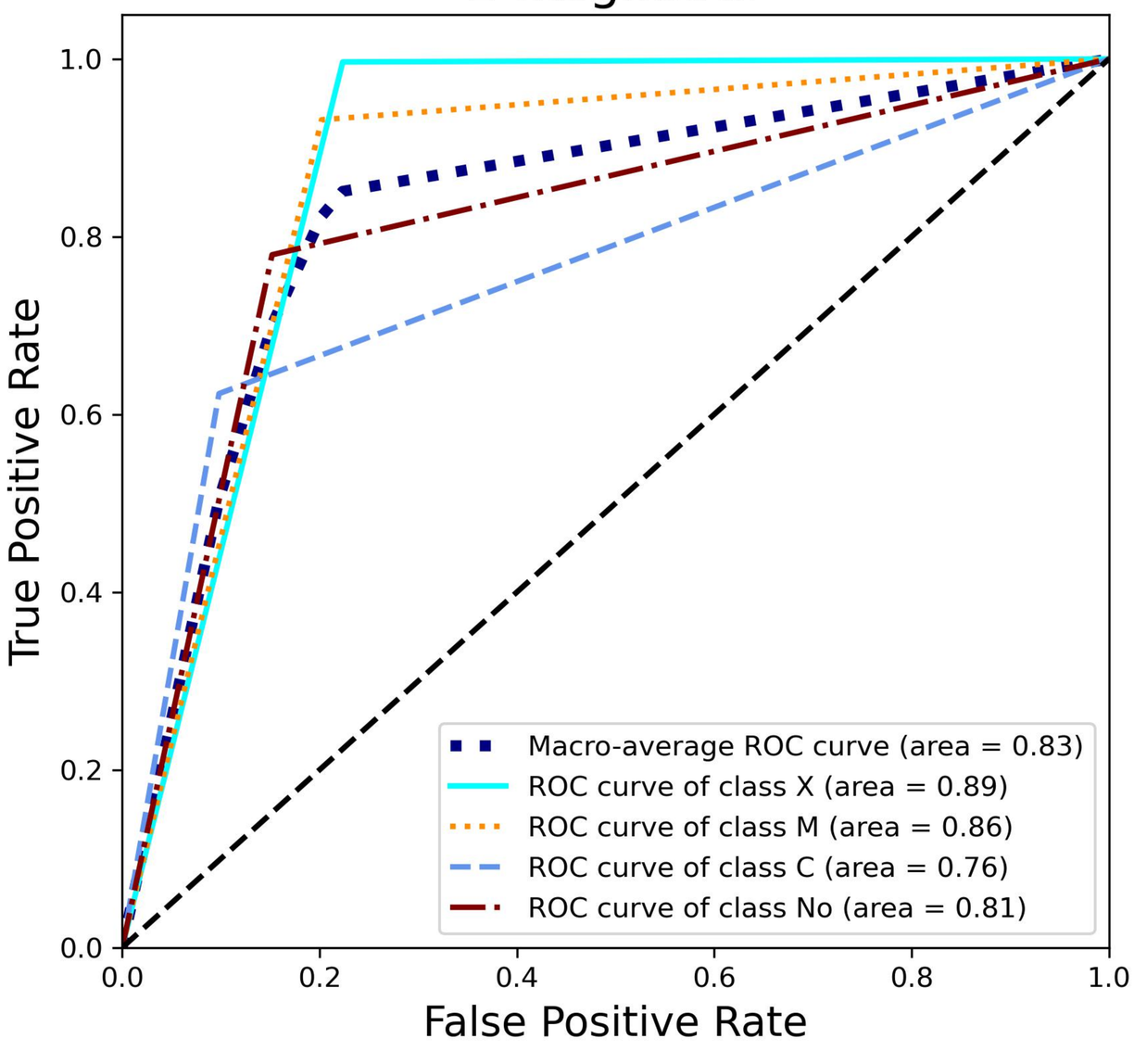
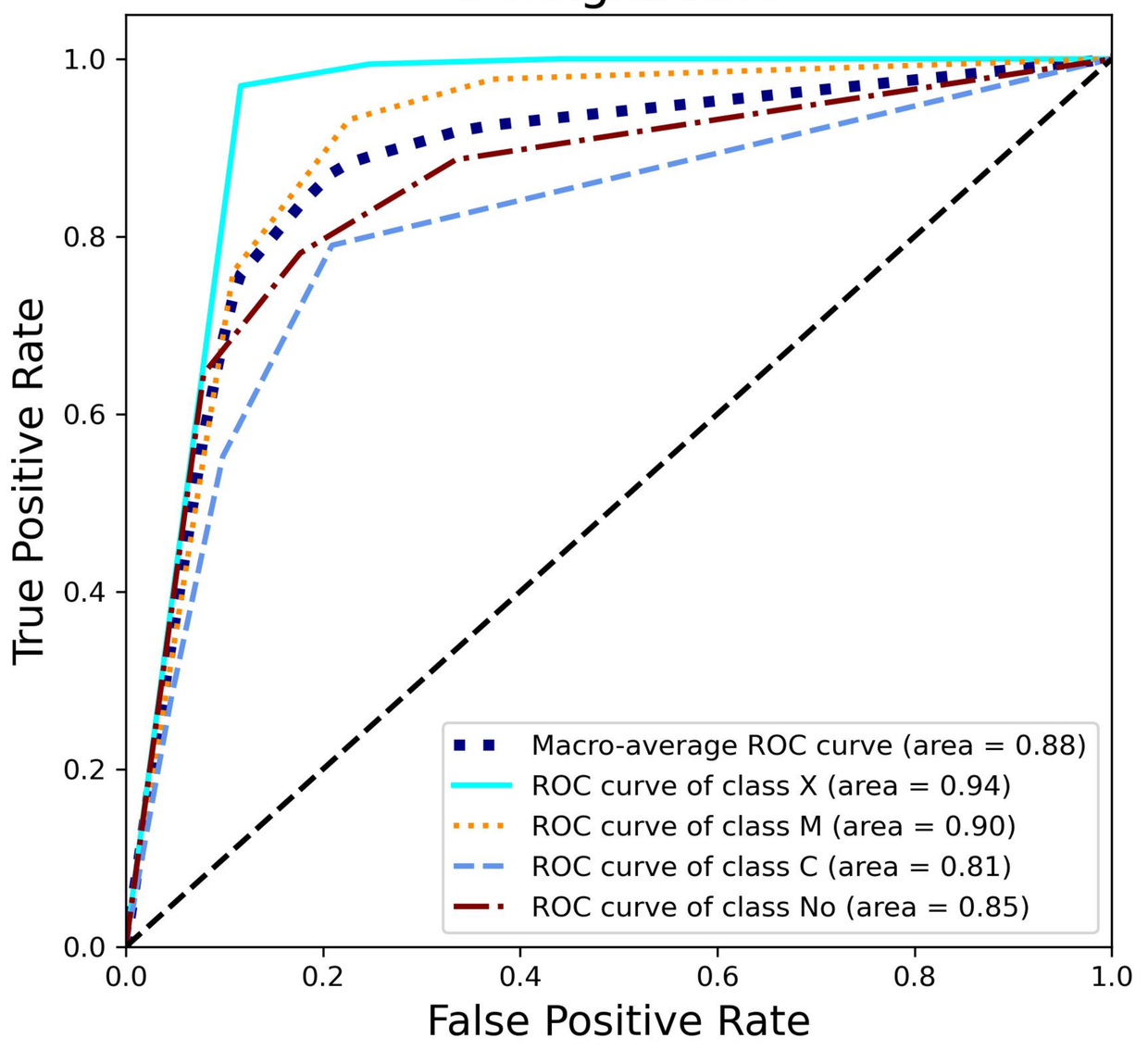
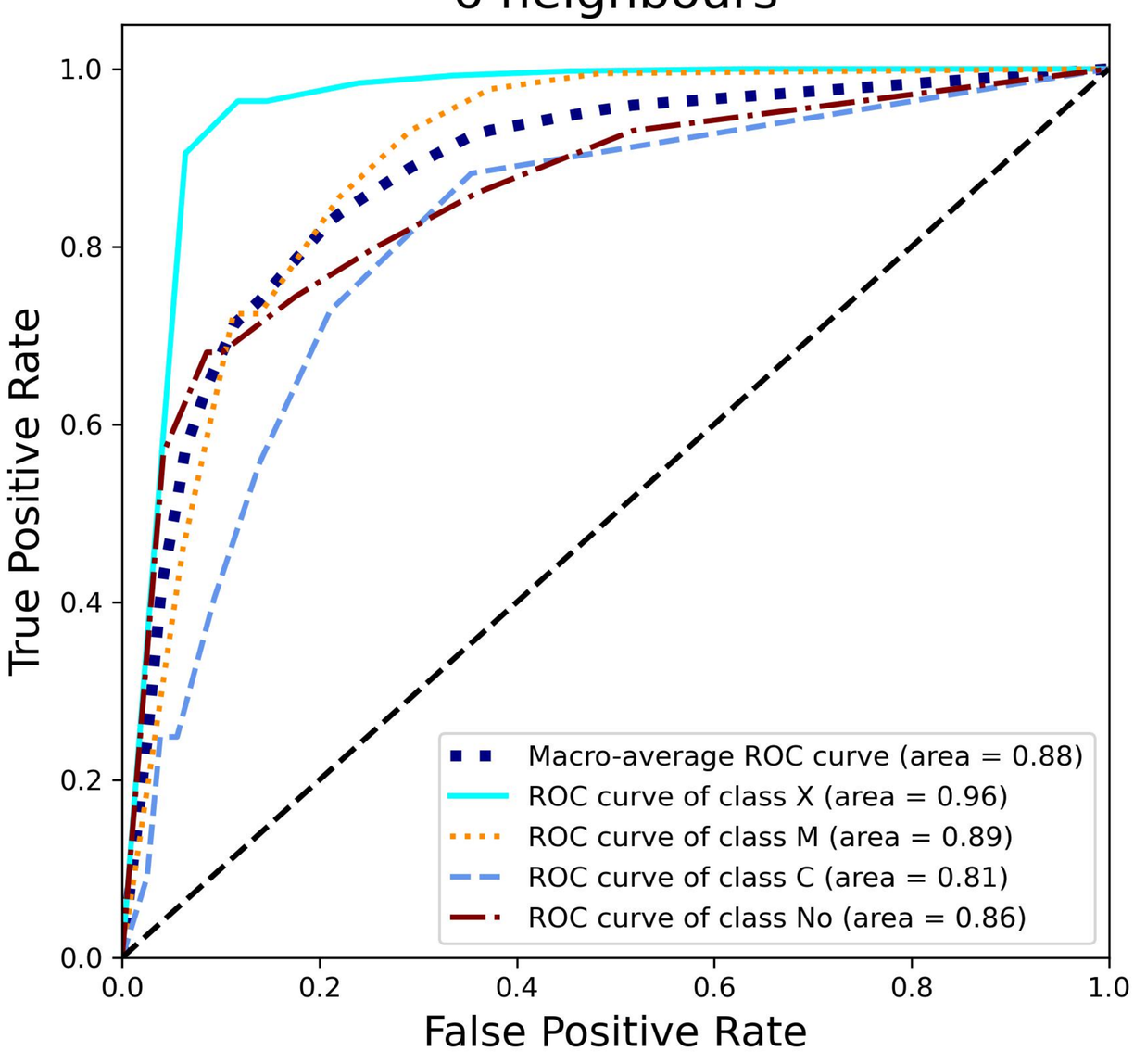

All data
X
M
C
No
Density
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
4
ABSNJZH

Figure 7.


1 neighbour
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Macro-average ROC curve (area = 0.83)
ROC curve of class X (area = 0.89)
ROC curve of class M (area = 0.86)
ROC curve of class C (area = 0.76)
ROC curve of class No (area = 0.81)


3 neighbours
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Macro-average ROC curve (area = 0.88)
ROC curve of class X (area = 0.94)
ROC curve of class M (area = 0.90)
ROC curve of class C (area = 0.81)
ROC curve of class No (area = 0.85)


6 neighbours
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Macro-average ROC curve (area = 0.88)
ROC curve of class X (area = 0.96)
ROC curve of class M (area = 0.89)
ROC curve of class C (area = 0.81)
ROC curve of class No (area = 0.86)

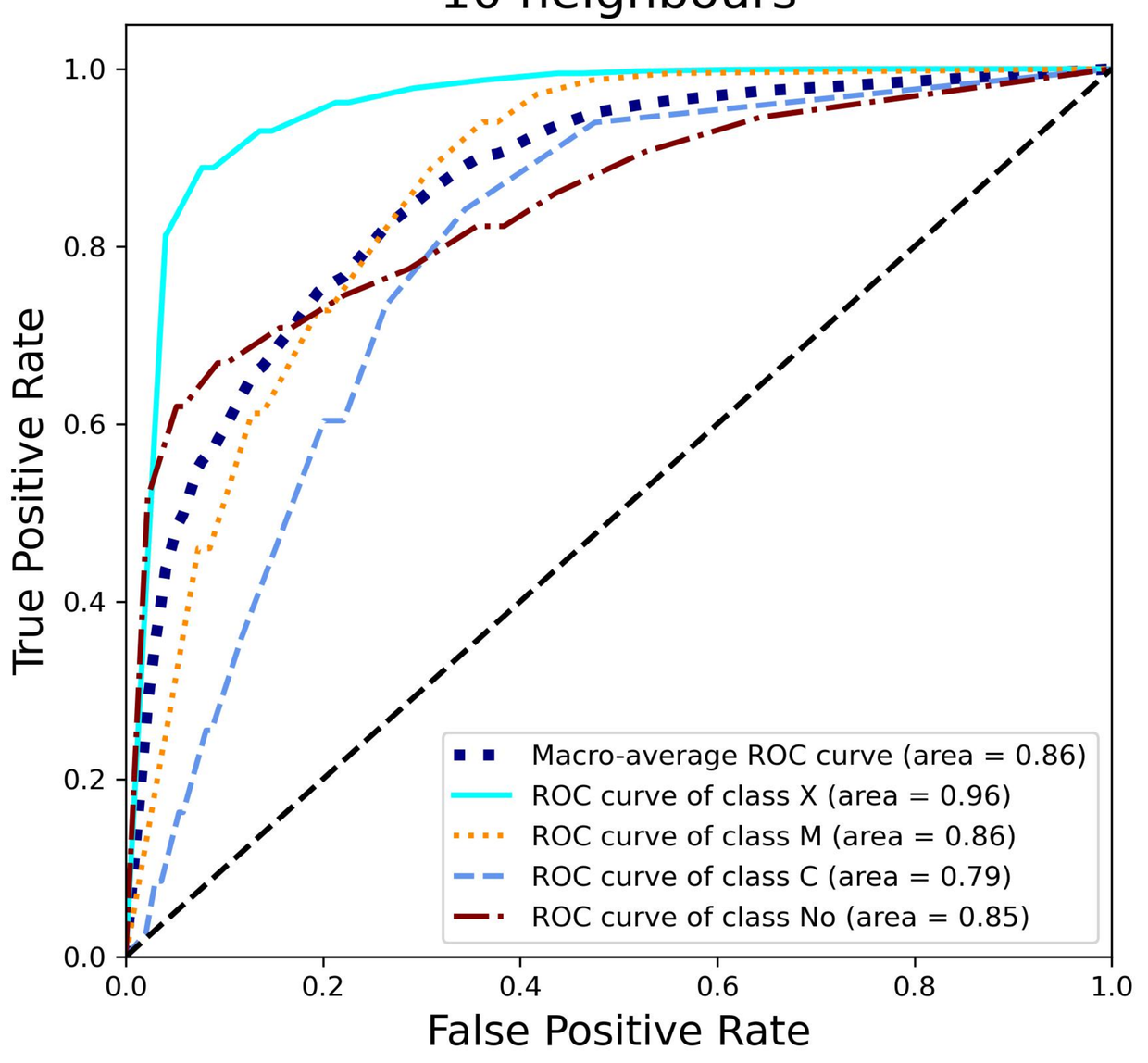

10 neighbours
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Macro-average ROC curve (area = 0.86)
ROC curve of class X (area = 0.96)
ROC curve of class M (area = 0.86)
ROC curve of class C (area = 0.79)
ROC curve of class No (area = 0.85)

Figure 8.

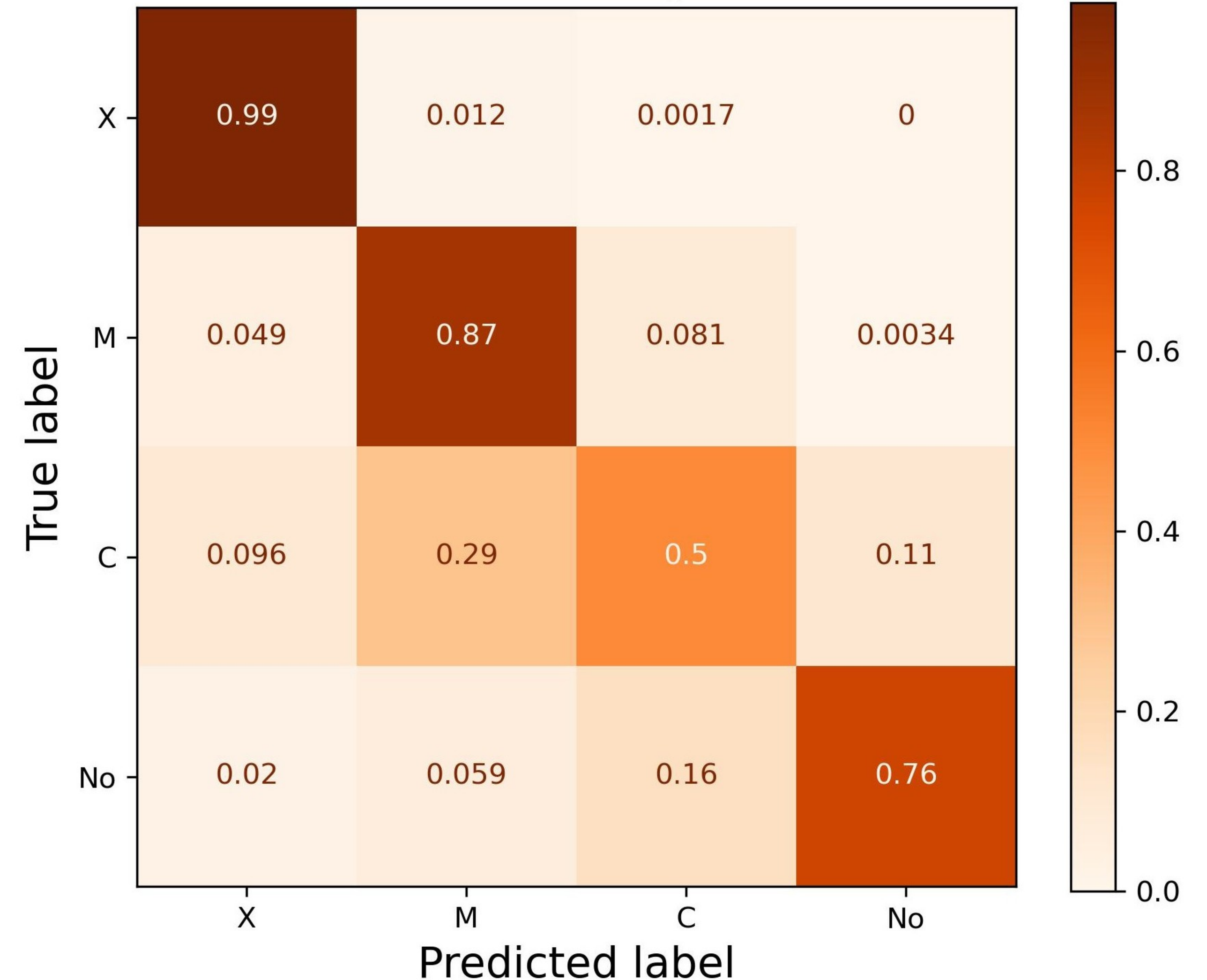

True label
X
M
C
No
0.99
0.012
0.0017
0
0.049
0.87
0.081
0.0034
0.096
0.29
0.5
0.11
0.02
0.059
0.16
0.76
X
M
C
No
Predicted label
0.8
0.6
0.4
0.2
0.0

Figure 9.

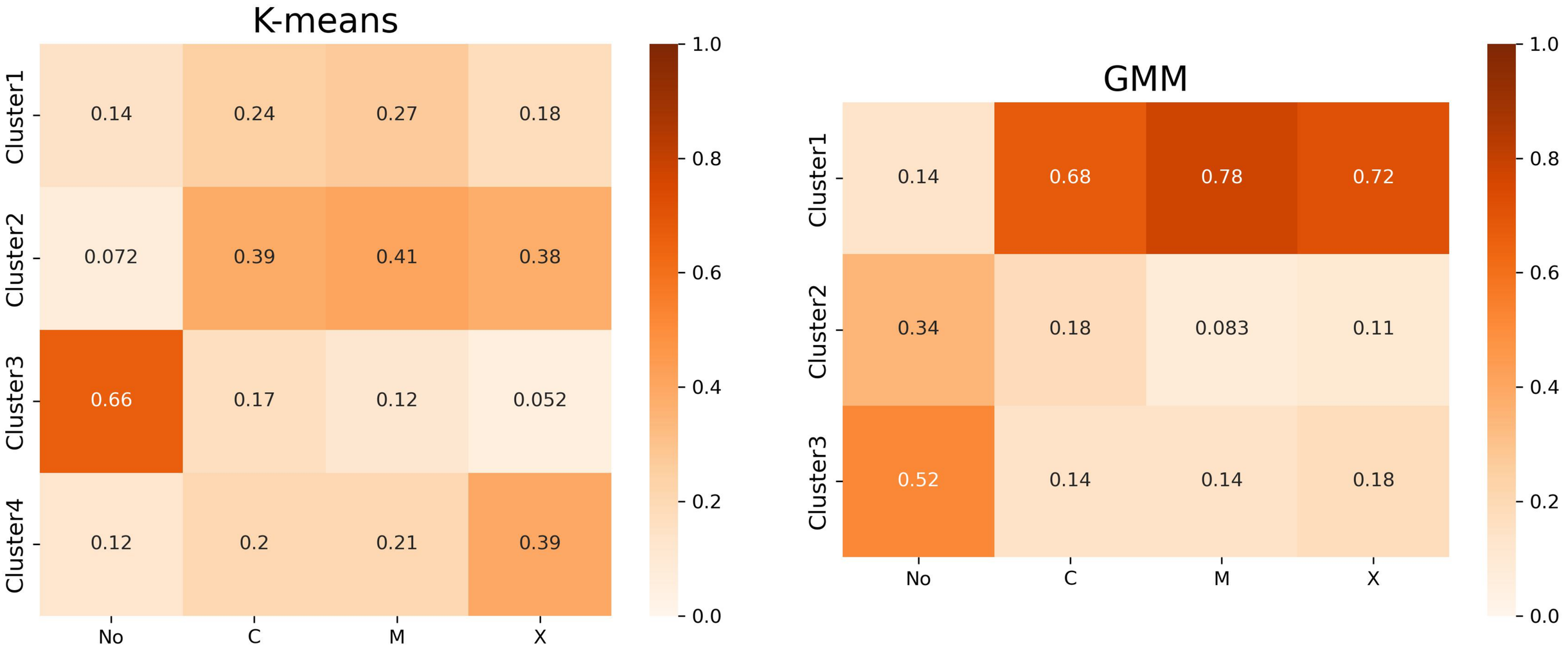
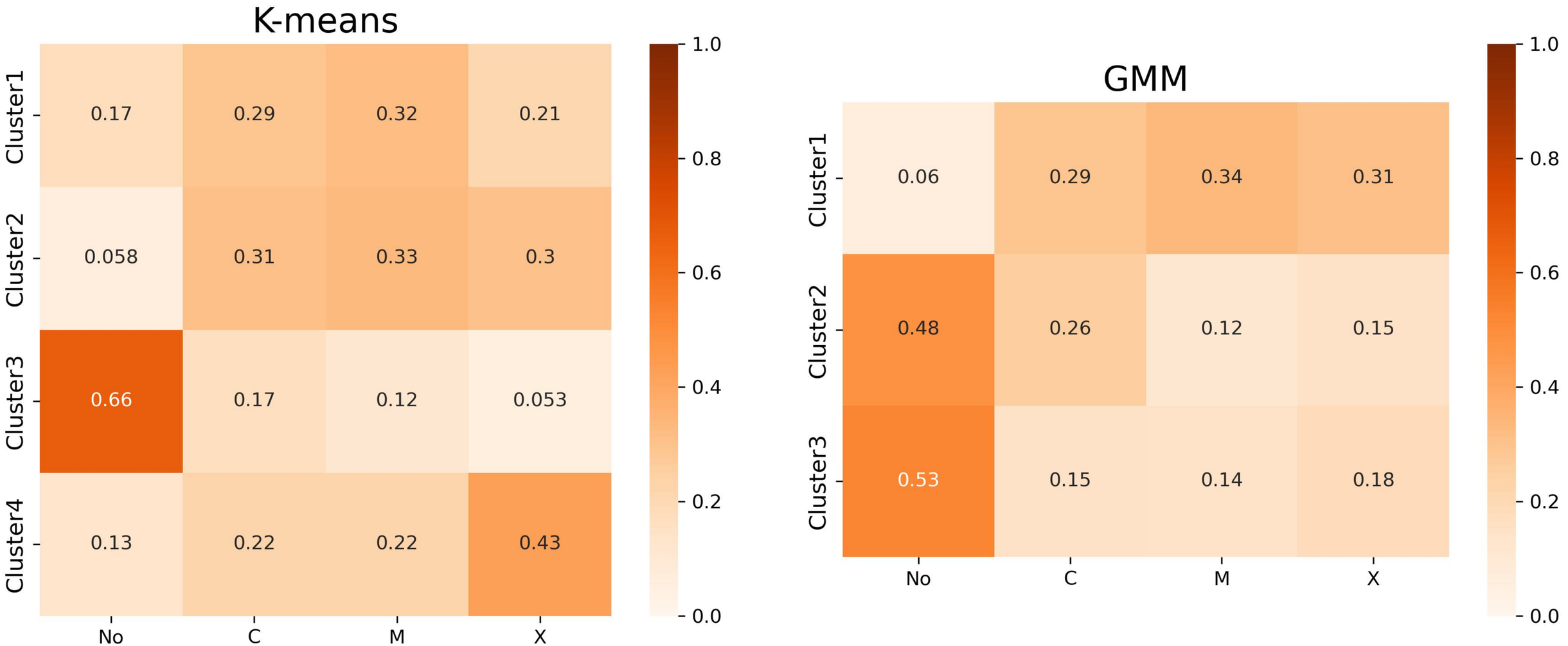

K-means
Cluster1
Cluster2
Cluster3
Cluster4
0.14
0.24
0.27
0.18
0.072
0.39
0.41
0.38
0.66
0.17
0.12
0.052
0.12
0.2
0.21
0.39
No
C
M
X
1.0
0.8
0.6
0.4
0.2
0.0
GMM
Cluster1
Cluster2
Cluster3
0.14
0.68
0.78
0.72
0.34
0.18
0.083
0.11
0.52
0.14
0.14
0.18
No
C
M
X
1.0
0.8
0.6
0.4
0.2
0.0

Figure 10.


K-means
Cluster1
Cluster2
Cluster3
Cluster4
0.17
0.29
0.32
0.21
0.058
0.31
0.33
0.3
0.66
0.17
0.12
0.053
0.13
0.22
0.22
0.43
No
C
M
X
1.0
0.8
0.6
0.4
0.2
0.0
GMM
Cluster1
Cluster2
Cluster3
0.06
0.29
0.34
0.31
0.48
0.26
0.12
0.15
0.53
0.15
0.14
0.18
No
C
M
X
1.0
0.8
0.6
0.4
0.2
0.0