



The Random Hivemind: An ensemble deep learning application to the solar energetic particle prediction problem

Patrick M. O'Keefe ^{a,*}, Viacheslav Sadykov ^b, Alexander Kosovichev ^{c,d}, Irina N. Kitiashvili ^d, Vincent Oria ^a, Gelu M. Nita ^c, Fraila Francis ^a, Chun-Jie Chong ^a, Paul Kosovich ^c, Aatiya Ali ^b, Russell D. Marroquin ^{e,b}

^a Computer Science Department, New Jersey Institute of Technology, 323 Dr Martin Luther King Jr Blvd, Newark 07102, NJ, USA

^b Physics & Astronomy Department, Georgia State University, 25 Park Place NE, Atlanta 30303, GA, USA

^c Physics Department, New Jersey Institute of Technology, 323 Dr Martin Luther King Jr Blvd, Newark 07102, NJ, USA

^d NASA Ames Research Center, NASA Research Park, Moffett Field 94035, CA, USA

^e Department of Physics, University of California San Diego, 9500 Gilman Dr, La Jolla 92093, CA, USA

Received 14 July 2023; received in revised form 19 April 2024; accepted 24 April 2024

Abstract

The application of machine learning and deep learning techniques, including the wide use of non-ensemble, conventional neural networks (CoNN), for predicting various phenomena has become very popular in recent years thanks to the efficiencies and the abilities of these techniques to find relationships in data without human intervention. However, certain CoNN setups may not work on some datasets, especially if the parameters passed to it, including model parameters and hyperparameters, are arguably arbitrary in nature and need to continuously be updated with the need to retrain the model, especially if the additions of new features render old models obsolete. This concern can be partially alleviated by employing committees of neural networks that are identical in terms of input features and initialized randomly and “vote” on the decisions made by the committees as a whole. Yet, members of the committee have similar architectures and features passed to them, making it possible for the committee members to “agree” on identical sets of weights and biases for all nodes and edges. Members of these committees also cannot be expanded to accommodate new features and entire committees must therefore be retrained in order to do so. We propose the Random Hivemind (RH) approach, which helps to alleviate this concern by having multiple neural network estimators make decisions based on random permutations of features and prescribing a method to determine the weight of the decision of each individual estimator. The effectiveness of RH is demonstrated through experimentation in the predictions of hazardous Solar Energetic Particle (SEP) events by comparing it to that of using both CoNNs and the aforementioned setup of committees identical in input features in this application. Our results demonstrate that RH, while having a comparable or better performance than the CoNN and a Committee-based approach, demonstrates a lesser score spread for the individual experiments, and shows promising results with respect to capturing almost every single flare instance leading to SEPs.

© 2024 COSPAR. Published by Elsevier B.V. All rights reserved.

Keywords: Sun: activity; Sun: particle emission; Solar-terrestrial relations; Computing methodologies: boosting; Machine learning; Neural networks

1. Introduction

The prediction of Solar Energetic Particle (SEP) events and the understanding of their precursors represent major challenges in heliophysics and space weather from both the operational and the research perspectives. Increased

* Corresponding author.

E-mail address: pmo6@njit.edu (P.M. O'Keefe).

fluxes of SPEs are of interest to various users, from governmental and private space weather agencies to airlines and power grid operators. Routine daily forecasting and short-term warning and alert systems for Solar Proton Events (SPEs) were implemented by the Space Weather Prediction Center (SWPC) at the National Oceanic and Atmospheric Administration (NOAA, [Balch, 1999; Balch, 2008](#)). The performance of the operational forecasting systems is still far from predicting every single SEP event ([Bain et al., 2021](#)).

SEP events are initiated by solar flares and coronal mass ejections (CMEs, [Reames, 2021](#)). Statistical relations between the flare soft X-ray properties (such as the peak ratios of the 1–8Å and 0.5–4Å fluxes, which is proportional to the flare temperature computed in a single-temperature approximation, [Ryan et al., 2012; Sadykov et al., 2019](#)) and the consequent CMEs and SPEs have been known for a long time. Solar flares are classified in terms of the maximum soft X-ray (SXR) flux observed by the NOAA Geostationary Operational Environmental Satellite (GOES) Network in the 1–8Å wavelength range. In this classification, the A-class flares have the maximum soft X-ray flux greater than 10^{-8} W/m²; for the B, C, M, and X-class flares, the SXR is greater than 10^{-7} , 10^{-6} , 10^{-5} , and 10^{-4} W/m² correspondingly. In particular, it was found that the lower the soft X-ray class, the greater the difference in the peak temperatures between the SEP-associated and non-SEP flares, with lower temperatures corresponding to the SEP-associated flares ([Garcia, 1994](#)). These relations were quantified and utilized for forecasting SPEs using a larger number of flare events ([Garcia, 2004](#)). The results were also reproduced later ([Kahler and Ling, 2018](#)), where the authors attempted to predict the SEP-associated flares using the k-nearest neighbors machine learning algorithm and neural networks separately for the Western and Eastern hemispheres of the Sun. Although the observed relationships are clear, the exact reason why the lower-temperature flares are more associated with SPEs remains largely unknown ([Kahler and Ling, 2018](#)). In addition, the durations and temperatures of the flares were found to be related statistically to the properties of CMEs ([Ling and Kahler, 2020; Kahler and Ling, 2022](#)). Thus, the flare duration and temperature can be used to constrain SEP parameters ([Kahler and Vourlidas, 2013](#)) or serve as a basis for establishing empirical models for SEP forecasting ([Richardson et al., 2018](#)). The Empirical model for Solar Proton Event Real-Time Alert (ESPERTA) forecasting tool ([Laurenza et al., 2009; Laurenza et al., 2018](#)) also utilizes the integrated SXR intensities and integrated radio intensities at 1 MHz to provide short-term predictions of SPEs with proton energy higher than 10 MeV and 100 MeV.

The extension of these works is the employment of machine learning (ML) and deep learning techniques, in particular, for forecasting SPEs based on the properties of the preceding (parental) solar flares. For example,

[Aminalragia-Giamini et al. \(2021\)](#) employed neural networks trained on the time series of the SXR fluxes observed during the solar flares directly. The authors found that the model can predict a large majority of SEP-associated flares (higher than 85%) during the period of 1988–2013 while maintaining a low false-positive rate. [Boubrahimi et al. \(2017\)](#) analyzed the correlations among the GOES soft X-ray and proton flux time series and employed the classification decision tree model for predicting the 100 MeV SPEs. [Lavasa et al. \(2021\)](#) analyzed a variety of ML algorithms such as random forest, neural networks, extremely randomized trees, and extreme gradient boosting. They concluded that, among the soft X-ray parameters, fluence is the most important for predicting SPEs.

The utilization of Conventional Neural Networks (hereafter CoNNs, defined in this work as a single, i.e., non-ensemble, neural network of any architecture) for space weather prediction purposes is very common in literature (e.g., [Torres et al., 2022; Nishizuka et al., 2020, etc.](#)). While CoNNs are very flexible and malleable in how they train on new data, the parameters they provide for the aspects, such as the size and shape of a given model, hyperparameters, and model selection, may need to be adjusted to be used on other data sets. This comes with the consequence of a requirement to continuously retrain models as data becomes increasingly outdated or if new features are to be added. A model's number of epochs and learning rate ideally need to be also adjusted based on the features fed into the model. In particular, the number of epochs low enough and the learning rate high enough can help to avoid overfitting the model to noise data ([Afaq and Rao, 2020; You et al., 2019](#)), yet one needs to make sure that the model is trained to a sufficient extent in general to learn the patterns in the data. This conflict may lead to a particular combination of an epoch count and learning rate for an entire model being sub-optimal, as a single combination may not work for all features, especially if some are more deterministic of the true labels of a given data set than others. Compromises could be made to adjust these parameters, but said compromises may, again, cause a given model to underfit based on some parameters and overfit based on others. One of the strategies to counter overfitting, in general, is to employ ensemble learning ([Cunningham, 2000](#)).

Another challenge related to the SEP prediction is that these events are rare and will represent minority-class events for the classification problem. For example, the ratio of the number of days with the enhanced proton flux with the energy greater than 10 MeV and the particle flux greater than 10 pfu (one particle flux unit, pfu, is equal to one particle per cm² per second per steradian) to the number of days with no enhanced flux is $\sim 1/23$ for Solar Cycles 22–24 ([Ali et al., 2023](#)). This ratio is even smaller ($\sim 1/34$) for Solar Cycle 24 alone ([Sadykov et al., 2021](#)). It was concluded for the ESPERTA model ([Stumpo et al., 2021](#)) that the performance of the algorithm (specifically, the False Alarm Rate, FAR) depends on the

class-imbalance ratio in the train data set. Various techniques can be implemented to deal with class-imbalanced data, such as oversampling, undersampling, and misclassification weights (Ahmadzadeh et al., 2021), and using synthetic data (Chen et al., 2021). In addition to the traditional data-centric approaches to dealing with class imbalance, ensemble classifiers can be employed in such problems (Galar et al., 2012). With respect to the problem of the prediction of SEPs, promising results were previously obtained employing neural network-based Committee ensembles (Aminalragia-Giamini et al., 2021) and random forest ensemble algorithms (Lavasa et al., 2021).

In our previous work, we presented the application of the random forest ML algorithm for predicting SEPs and tested various class-imbalance treatment techniques (O'Keefe et al., 2022). In this work, we expand our investigation to new types of ML algorithms, including Conventional Neural Networks (CoNN), an ensemble of CoNNs following a voting approach (Committee, Aminalragia-Giamini et al., 2021), and introduce a weighted consensus that we call a Random Hivemind (RH). Both considered ensemble approaches are so-called “bagging” ensemble classifiers when individual ensemble members do not depend on each other and deterministically contribute to the classification decision. *Investigation of the relative performance of the algorithms on the given data set of flares associated with SEPs is the primary focus of this paper.* The paper is structured as follows. Section 2 describes the data preparation employed in this paper, namely the processing of the soft X-ray data, the association of flares and SEPs, and the preparation of data sets ready for machine learning (ML) analyses. Section 3 describes the ML algorithms tested in this work. The results and discussion are presented in Section 4 and followed by the conclusion in Section 5.

2. Data preparation

The solar soft X-ray (SXR) emission observed by the GOES satellites in 0.5–4 Å and 1–8 Å wavelength channels can be represented under a single-temperature plasma approximation by two parameters, namely the plasma temperature (T) and its emission measure (EM). We utilize the legacy data set of the T and EM values estimated using the Temperature and Emission Measure Based Background Subtraction algorithm (TEBBS, Ryan et al., 2012; Sadykov et al., 2017) and collected in the Interactive Multi-Instrument Database of Solar Flares (IMIDS¹, Sadykov et al., 2017) for the 2002–2017 time period. In addition to peak values of the temperature and emission measure, T_{max} and EM_{max} , we utilize the background-subtracted flare classes (SXR_{max}), flare durations, the times of the peaks of T_{max} , EM_{max} , and SXR_{max} relative to the flare start and end times, and the observed disk X- and Y-

coordinates of the host flare. In total, this data processing gives 12 parameters for every solar flare. For some flares, due to the complexity of the X-ray emission variations and high background noise, this algorithm failed (producing unrealistic $T_{max} \geq 100$ MK, and negative time differences), and such flares were excluded from our data set. The total number of flares included in our analysis is 18311. Among these are: 5 A-class, 6919 B-class, 10074 C-class, 1207 M-class, and 106 X-class flares according to the SXR classification. These flare classes were calculated after subtracting the SXR background. As illustrated in Fig. 4c of Sadykov et al. (2019), the background subtraction mostly affects the flare class of the weak BC-class flares, yet leading to more reliable behaviors of T and EM curves.

To associate the flares with the SEP records, we utilize the list of the Solar Proton Events Affecting the Earth Environment² provided by the NOAA Space Environment Services Center. This data set represents the SEP events with the peak flux of >10 MeV protons higher than 10 pfu. A total of 64 flares from our list were associated with the SEP events and 18247 flares were without SEPs (non-SEP flares), providing an extreme class-imbalance ratio of 1/285. In terms of the SXR_{max} parameter, 8 of the SEPs correspond to the C-class flares, 36 — to M-class, and 20 — to X-class flares. The list of the studied flares is publicly available at the Solar Energetic Particle Prediction Portal (SEP³) website³. We note here, that the number of SEP events considered in some other works on the SEP prediction problem (e.g., Papaioannou et al., 2016; Lavasa et al., 2021) was significantly larger (314 and 257, correspondingly) than in the current work. Both of these studies consider a longer time span (events detected during 1984–2013 and 1988–2013) which led to larger statistics of SEP events. Also, in the current work, we omit from consideration the SEP events which (1) do not have an association with the flare event, and (2) do not have the coordinates of the host flare event identified according to the GOES flare catalog. Both restrictions are leading to the loss of 15 SEP event records, leaving us with 64 events in total. Fig. 1 illustrates the locations of the SEP and non-SEP flares on the solar disk. It is evident that the SEP flares have a preference to originate in the Western hemisphere, which has a more direct magnetic connectivity to Earth, although some SEP flares have originated close to the Eastern limb as well.

The distribution of the SEP-associated and non-SEP solar flares in the diagrams of the flare temperature vs. the soft X-ray peak flux, the flare duration vs. the emission measure, and the flare rise time vs. the flare decay time are presented in Fig. 2. Here, the flare rise time is determined as the time of SXR_{max} minus the flare start time, and the flare decay time is determined as the flare end time minus

¹ <https://data.nas.nasa.gov/helio/portals/solarflares/>

² <https://umbra.nascom.nasa.gov/SEP/>

³ <https://sun.njit.edu/SEP3/datasets.html>

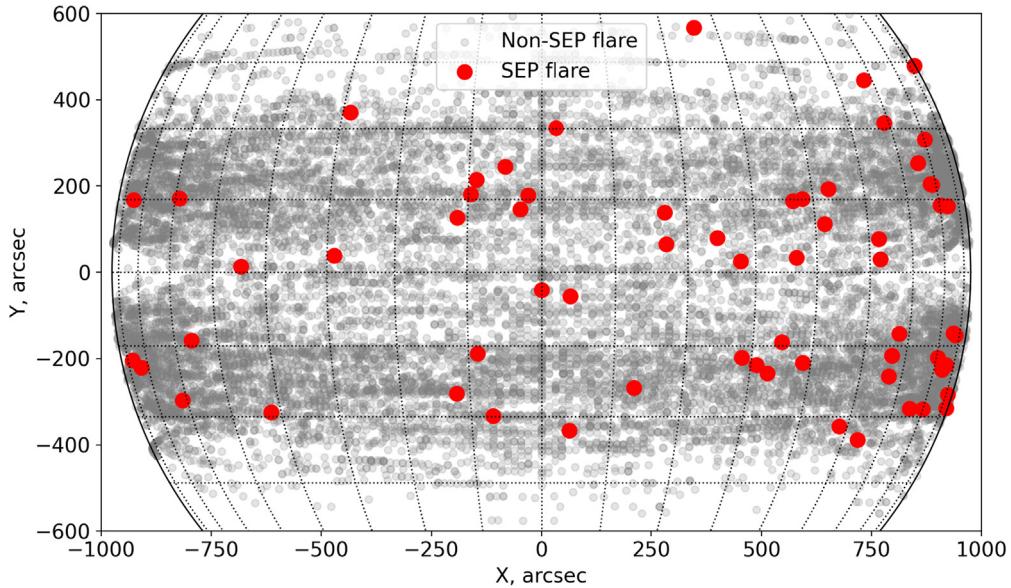


Fig. 1. The locations of the SEP and non-SEP flares in the data set on the solar disk. Gray dots mark the solar flares that did not produce SEP events, and red dots mark the flares that resulted in the SEP events. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

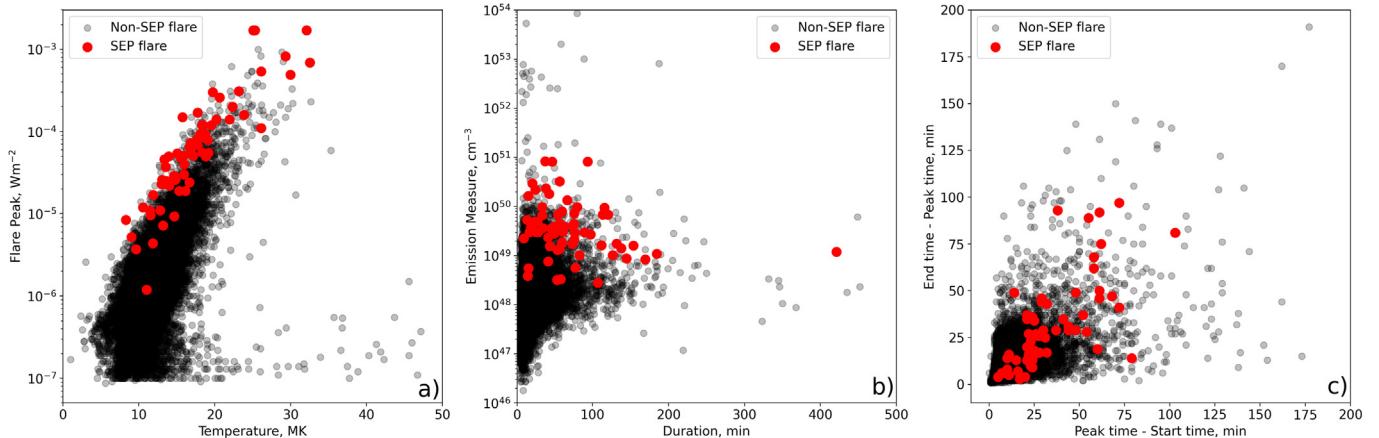


Fig. 2. (a) Distribution of the SEP and non-SEP flares in the soft X-ray flux at 1–8 Å and flare peak temperatures diagram, (b) the flare duration and the peak emission measure diagram, and (c) the flare rise time and flare decay time diagram. Black dots mark the solar flares that did not produce SEP events, and red dots mark the flares that resulted in the SEP events. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the time of SXR_{max} . One can see that the flares that resulted in SEPs are not distributed randomly, even among the flares of the same SXR peak fluxes. Specifically, Fig. 2a indicates that SEP-associated flares are colder on average among the flares with the same SXR peak flux (or the flare class). A similar dependence was found by Garcia (1994, 2004). While no obvious pattern is observed for the parameters presented in Figs. 2b and 2c, a more detailed investigation is required to understand their relations to SEP-associated flares. For the ML analysis, we subdivide the data set into training and testing subsets; the latter contains 30% of the full dataset. The train-test separation is ran-

domly repeated ten times for every machine-learning experiment presented in this paper.

3. Machine learning methodology

Three neural network-based approaches are considered in this paper for the problem of prediction of SEP events. The first is the conventional neural network (hereafter CoNN), which represents a single fully connected neural network architecture. For ensemble deep learners, two more neural network-based approaches are constructed. The first ensemble approach is the Committee scheme

employed by [Aminalragia-Giamini et al. \(2021\)](#), which involves several neural network estimators with the same input features and input layer shapes. The second ensemble approach is a Random Hivemind (RH) built using random down-selection of the flare characteristics from the training data as input features.

There are several motivations to consider the development of the RH approach. First, the RH allows for individual estimators to be grown individually to accommodate additional features without the need to retrain the entire ensemble. By requiring some, but not all, of the estimators to be retrained, this may reduce the amount of time it takes to retrain deep learning models. Second, the RH not only takes advantage of the ability of ensemble learning to counter overfitting in general ([Cunningham, 2000](#)), but also adjusts the training epoch counts, learning rates, and voting weights of individual estimators depending on the importance of the features used in these estimators. This allows RH to accommodate the risk of overfitting by reducing the chances of less important features being included or influencing the prediction.

For the series of tests presented in this paper, we consider two RH realizations that have several differences. First, the realizations are using a different number of input features. The first RH implementation (hereafter RH.v1) uses the square root of the total number of features, rounded up, as the number of input features for each neural network within an RH. The second implementation (hereafter RH v2) uses half of the features as input. The layout between estimators remains unchanged within each Committee. In some sense, the semi-random selection of features (the probability of selection is yet proportional to their weight; see below) is inspired by the Random Forest ensemble learning algorithm ([Breiman, 2001](#)). Correspondingly, one has 12 features entering the CoNN or each committee member, and 4 or 6 features selected using the procedure described below entering the RH classifier. Each ensemble setup (i.e. Committee, RH v1, and RH v2) has 10 neural network estimators. The architectures of the utilized ML methods are schematically illustrated in [Fig. 3](#).

The random permutations of features are chosen by first computing the χ^2 -squared and mutual information gain statistics between the features and the SEP presence to assign scores to each feature based on how significant each is in determining whether or not a given flare caused a SEP event. Each feature's score is calculated using this formula:

$$s_i = \sqrt{\chi_i^2} + \kappa_i \quad (1)$$

Here, s_i is a given feature's total score, χ_i^2 is the χ^2 -squared statistic between a given feature and SEP presence, and κ_i is the mutual information gain statistic between a given feature and SEP presence. After each feature is scored based on this formula, the scores are normalized so that the sum of these scores ("feature weights") equals one. The feature weights are then used as the probabilities that given

features with their respective weights will be chosen for a given RH estimator.

Each neural network, including CoNNs, Committee estimators, and RH estimators, has an input layer equal to the number of features being tested by the estimator, a dense layer with an input and output shape of 10, a dropout layer with a probability of 0.2, and an output layer with an output shape equal to the number of predicted features. The networks are implemented using the Skorch library ([Tietz et al., 2017](#)). The numbers of epochs and learning rates for all CoNN and Committee setups are $n_{\text{epochs}} = 500$ and $\alpha = 0.001$ following the hyperparameter optimization discussed later. In addition, RH boosts its epoch counts and learning rates using these formulae:

$$n_{\text{epochs}} = 500 \times (2 - x) \quad (2)$$

$$\alpha_{\text{RHv1}} = 0.001 \times \left[1 + \ln \left(\frac{e^x - x}{\eta} + 1 \right) \right] \quad (3)$$

$$\alpha_{\text{RHv2}} = 0.001 \times (0.5 + x) \quad (4)$$

Here, x is the total sum of feature weights for a given estimator, n_{epochs} is the number of epochs during the training process, α is the learning rate, and e is the base of the natural logarithm. The parameter $\eta = \frac{e^{\Sigma x}}{n_{\text{features}}}$, where n_{features} is the number of features selected for the given estimator, and Σx is the total weight of all features in all estimators within the ensemble. The RH classifier is tested twice, with α equal to α_1 in the first test and α_2 in the second. Outcomes are predicted by putting prediction data through each of the estimators constructed during the training phase and seeing what each estimator chooses as a predicted result. Each committee considers all results by all estimators as equal, using a simple plurality vote to determine which class a given datum belongs to. Each RH considers each estimator's value in a classification vote as equal to the sum of the feature weights said estimator's input features have. For all neural networks, including CoNNs, Committee estimators, and RH estimators, the Adam optimizer ([Kingma and Ba, 2015](#)) is used, the cross-entropy loss function is used with balanced class weights, and overfit prevention measures including dropout layers with probabilities of 0.2 and data shuffling are used.

Let us consider an example of feature weights in more detail. If a given flare's SXR peak flux had a feature weight of 0.25, its emission measure peak value had one of 0.1, its temperature peak value had one of 0.05, and its duration had one of 0.01, and the total sum of all the feature weights was 1, the peak SXR flux would have a probability of 0.25 of being chosen to be in an RH estimator, the emission measure peak flux would have one of 0.1, etc. A CoNN and a Committee, however, would consider all available features equally as input features. During training, an RH estimator that uses all four of these parameters would go through 15 epochs with learning rates of approximately $\alpha_{\text{RHv1}} = 0.00162$ in the first test and $\alpha_{\text{RHv2}} = 0.00186$ in the second. A CoNN and a Committee estimator in this example, however, would each only go through 10 epochs with a

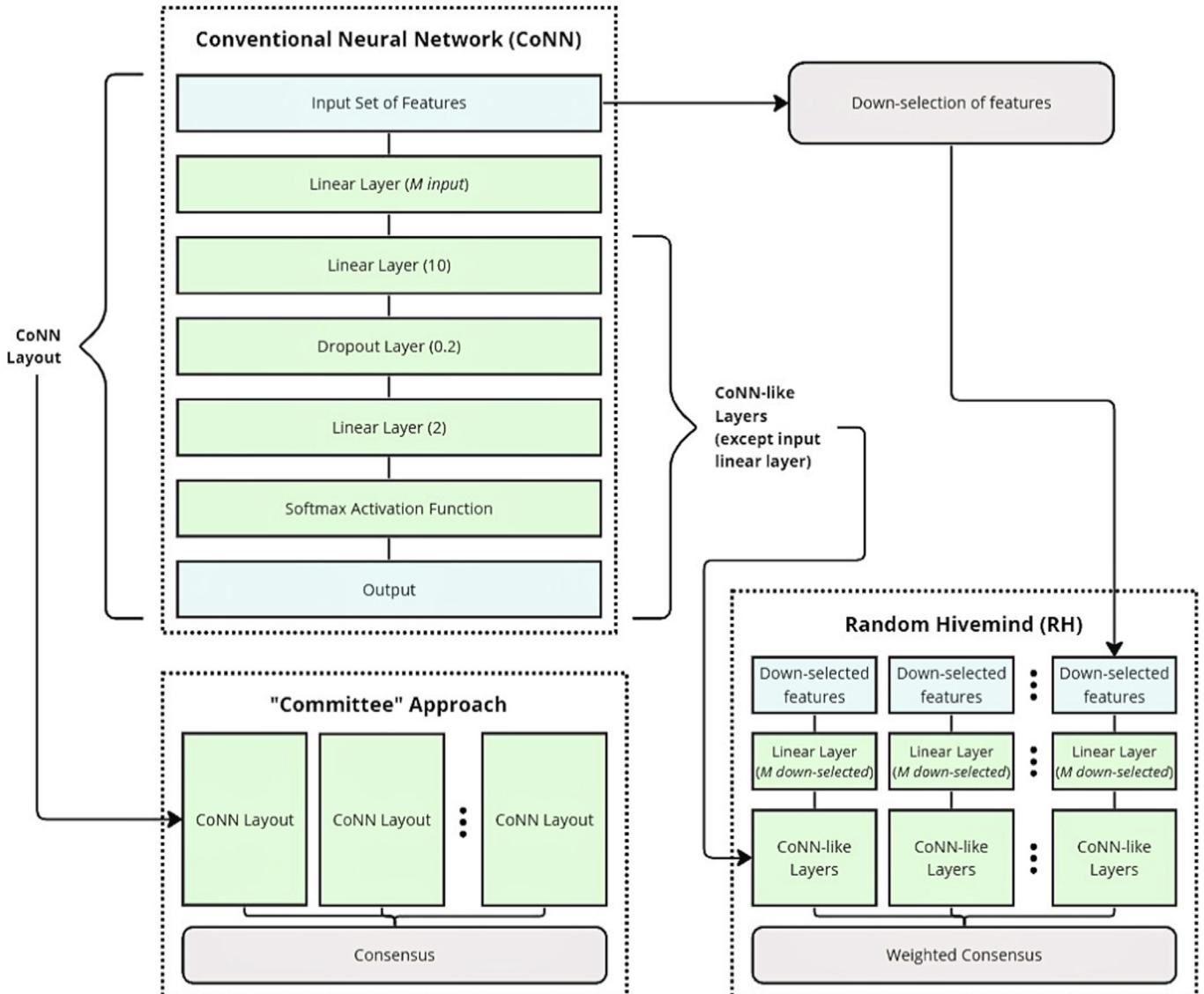


Fig. 3. The schematic representation of the layouts of the Conventional Neural Network (CoNN), Committee Network, and Random Hivemind Network (RH). The numbers in parentheses for the linear layers indicate the number of neurons in the layer, and the number in parentheses for the dropout layer indicates the probability of each connection/weight being dropped from the training procedure.

learning rate of $\alpha = 0.001$, since they cannot automatically calculate these parameters based on feature selection. When deciding, each RH estimator would use the sums of its feature weights as values, so an estimator with these four parameters would have a value of 0.5 when voting. Each Committee estimator would have a value of 1, since, again, no mechanism exists to determine how to calculate these figures based on feature selection.

Tuning of hyperparameters is of known importance for machine learning. For the neural network-based approaches, hyperparameters may include those related to the network architecture (number of hidden layers, neurons in each layer, activation functions, etc.) and the training process (number of epochs, learning rate, optimizer, and regularization parameters, etc.). The parameter space increases even further if considering ensemble approaches. Exploring the entire parameter space is very costly. There-

fore, in this work, we restrict the model architectures to those illustrated in Fig. 3, leaving only the training process-related parameters for optimization. We also notice that the ensemble approaches used in this work have the common CoNN structure as their basis; therefore, optimizing the training for CoNN should deliver the optimal training for the ensemble approaches as well. We explore a grid of the learning rates of $\alpha = \{0.01, 0.0025, 0.001, 0.0005\}$, and epoch counts up to $n_{\text{epochs}} = 2000$, with 10 experiments per each parameter pair, to find the optimal, yet not relatively costly, CoNN training procedure.

Our results for optimization of CoNN with the $\alpha = 0.001$ are illustrated in Fig. 4. We note here that the further decrease of the learning rate to $\alpha = 0.0005$ did not lead to an increase in the HSS and TSS scores. As one can see, the CoNN performance roughly increases with

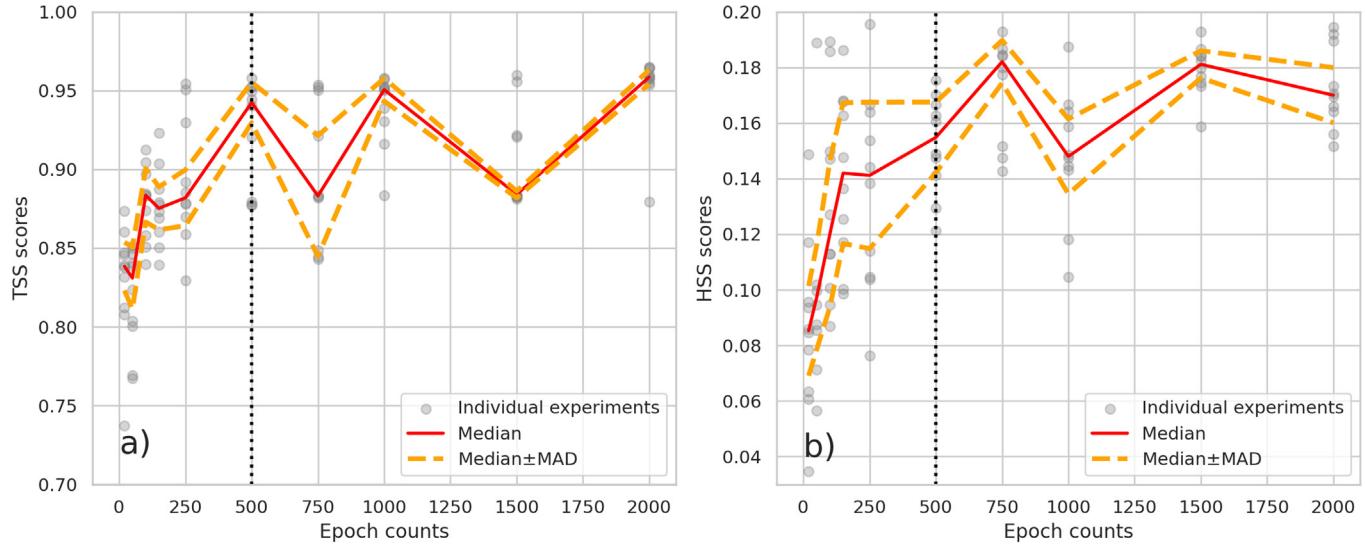


Fig. 4. Median (red) and median absolute deviation (orange) performance of the CoNN for the different number of the training epochs and the learning rate of $\alpha = 0.001$. Panel (a) corresponds to the TSS score, and panel (b) corresponds to the HSS score. Gray points indicate the results of individual experiments. black dotted line indicates the number of epochs considered as optimal and is used for the study.

the increase in the number of epochs. We decided to select $n_{\text{epochs}} = 500$ for the prediction experiments: as evident, at this number of epochs trained, the classifier's TSS score starts to decrease significantly for the first time. It is also a relatively low number of epoch counts which provides a possibility to perform a massive number of runs for CoNN and ensemble approaches while evaluating their performance. However, we note here that further fine-tuning of the hyperparameters is possible for this problem.

Overall, the hyperparameters of the RH algorithm include the architectures of the individual estimators, the implementation of the particular weighting scheme, feature importance estimation, the number of features to use in each estimator, and the adjustment of the RH training parameters, etc. The RH implementations considered here were only constructed to demonstrate a proof-of-concept for RH. In particular, we show further that RH v2 is performing better compared to RH v1 on the problem of SEP prediction. This illustrates that different realizations of the RH algorithm may work differently on the same data set and that a careful tuning of RH parameters and options is required to construct the most optimal model for the particular task.

The CoNN and each ensemble member have two output channels predicting the likelihood of the entry belonging to the positive (SEP-associated) or the negative (non-SEP) class. For the binary classification, the class of the larger likelihood is chosen as the prediction. To issue the ensemble prediction, an RH classifier chooses the result that receives the highest number of weighted votes. We demonstrate the individual elements of the confusion matrix (true positive predictions, TP , true negative predictions, TN , false positive predictions, FP , and false negative predictions, FN) for each approach averaged over 50 random train-test splits. We also use various metrics to compare

binary classification predictions including accuracy, true skill score (TSS), Heidke skill score 2 (HSS), precision, and recall. For a definition of these metrics, see, for example, [Bobra and Couvidat \(2015\)](#) and references therein. TSS and HSS scores are also discussed in more detail in Section 4. In order to mitigate the susceptibility of the accuracy metric to the high-class imbalance of the data set, we consider in addition the balanced accuracy, defined here as:

$$BA = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right). \quad (5)$$

We also construct the Receiver Operating Characteristic curves for each classifier and calculate the area under the curve (ROC_AUC) as a prediction metric. The probabilities of classes for the individual estimators are obtained from likelihoods using a softmax function. The positive class probabilities (which are to construct the ROCs) for the RH tests are calculated by taking weighted averages of the probability predictions of each estimator, with each weight being the total sum weight for each estimator following the normalizations of all weight sums.

4. Results and discussion

The results of the classification algorithms employed in this study in terms of confusion matrix elements and various prediction scores are presented in [Table 1](#) (summary results for all classifiers as the average scores and standard deviations) and [Table 2](#) (summary results for all classifiers as the median scores and median absolute deviations). The statistics of the results are also illustrated in the box-and-whisker plot presented in [Fig. 5](#) and the Receiver Operating Characteristic (ROC) curves averaged over 50 experiments in [Fig. 6](#). One of the indicators of the sufficiency of 50

Table 1

Average values and standard deviations of the performances of the classifiers considered in this paper.

Algorithm/ Metrics	CoNN	Committee	RH v1	RH v2
TN	5233.0 ± 74.4	5250.9 ± 18.9	5244.2 ± 25.5	5245.2 ± 22.8
FP	237.0 ± 74.4	219.0 ± 18.9	225.7 ± 25.5	224.8 ± 22.8
FN	1.2 ± 1.3	0.8 ± 0.9	1.1 ± 0.7	0.4 ± 0.6
TP	22.8 ± 1.3	23.2 ± 0.9	22.9 ± 0.7	23.6 ± 0.6
Precision	0.094 ± 0.022	0.096 ± 0.005	0.093 ± 0.009	0.096 ± 0.008
Recall	0.945 ± 0.053	0.966 ± 0.038	0.956 ± 0.031	0.985 ± 0.025
Accuracy	0.957 ± 0.013	0.960 ± 0.003	0.959 ± 0.005	0.959 ± 0.004
Balanced Accuracy	0.953 ± 0.021	0.963 ± 0.018	0.957 ± 0.014	0.972 ± 0.011
TSS	0.906 ± 0.043	0.926 ± 0.035	0.915 ± 0.029	0.944 ± 0.023
HSS	0.163 ± 0.036	0.168 ± 0.009	0.163 ± 0.014	0.168 ± 0.013
ROC_AUC	0.9903 ± 0.0005	0.9907 ± 0.0001	0.9901 ± 0.0005	0.9906 ± 0.0003

Table 2

Median values and median absolute deviations (computed as median values of the absolute deviations of the individual scores from the median) of the performances of the classifiers considered in this paper.

Algorithm/ Metrics	CoNN	Committee	RH v1	RH v2
TN	5233.0 ± 46.5	5250.9 ± 10.0	5244.2 ± 20.0	5245.2 ± 15.5
FP	237.0 ± 46.5	219.0 ± 10.0	225.7 ± 20.0	224.8 ± 15.5
FN	1.2 ± 1.5	0.8 ± 0.5	1.1 ± 0.0	0.4 ± 0.0
TP	22.8 ± 1.5	23.2 ± 0.5	22.9 ± 0.0	23.6 ± 0.0
Precision	0.094 ± 0.016	0.096 ± 0.003	0.093 ± 0.006	0.096 ± 0.005
Recall	0.949 ± 0.063	0.966 ± 0.021	0.956 ± 0.0	0.985 ± 0.0
Accuracy	0.957 ± 0.008	0.960 ± 0.002	0.959 ± 0.004	0.959 ± 0.003
Balanced Accuracy	0.953 ± 0.021	0.963 ± 0.011	0.957 ± 0.005	0.972 ± 0.002
TSS	0.906 ± 0.042	0.926 ± 0.023	0.915 ± 0.010	0.944 ± 0.005
HSS	0.163 ± 0.026	0.168 ± 0.005	0.163 ± 0.010	0.168 ± 0.008
ROC_AUC	0.9903 ± 0.0002	0.9907 ± 0.0001	0.9901 ± 0.0003	0.9906 ± 0.0001

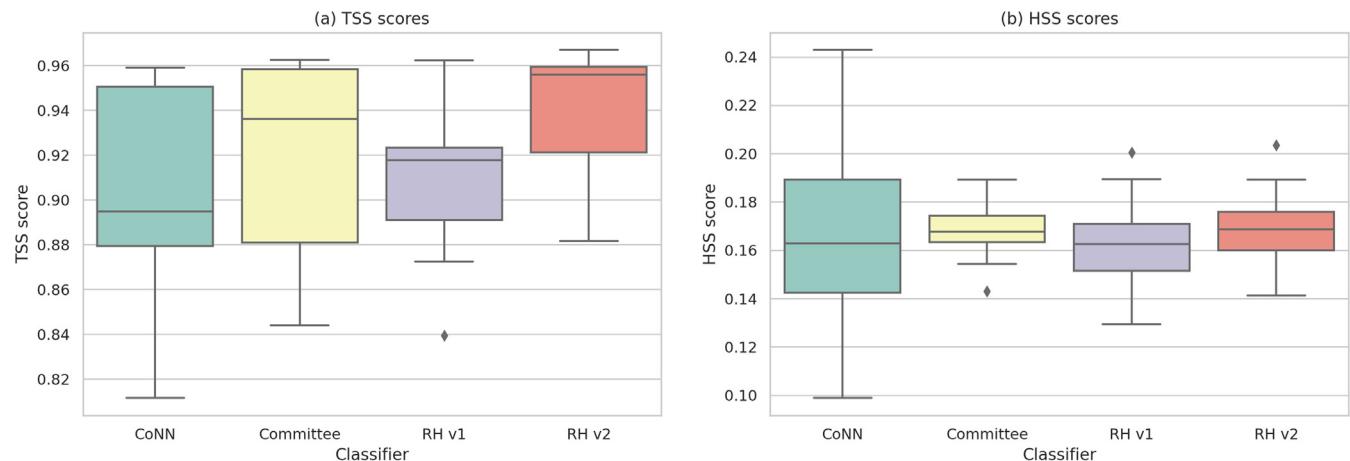


Fig. 5. Box-and-whiskers plots summarizing the performances of the CoNN, Committee, and two considered versions of RH, over the 50 experiments in terms of the TSS and HSS scores. Each colored rectangle spans through the second and third quartiles of the scores, with the horizontal bar marking the median. The whiskers indicate the locations of the last individual test results within the interquartile range from the boxes. The rhombus points mark outliers outside the interquartile range from the boxes.

experiments for the statistical robustness of the conclusions is that the mean and median values agree with each other to the last significant digit for almost every single metric and classifier considered.

Tables 1 and 2 indicate that the ensemble approaches are performing better, in general, than the CoNN classifiers with respect to the measures typically used in space weather

forecasting, HSS and TSS, both in terms of the mean and median values. For example, the TSS score had its median value of $TSS = 0.906 \pm 0.042$ for the CoNN classifier and increased to $TSS = 0.926 \pm 0.023$, $TSS = 0.915 \pm 0.010$, and $TSS = 0.944 \pm 0.005$ (RH v2) for the Committee and two RH ensemble classifiers. Although the HSS scores were relatively low, they still demonstrated either no drop

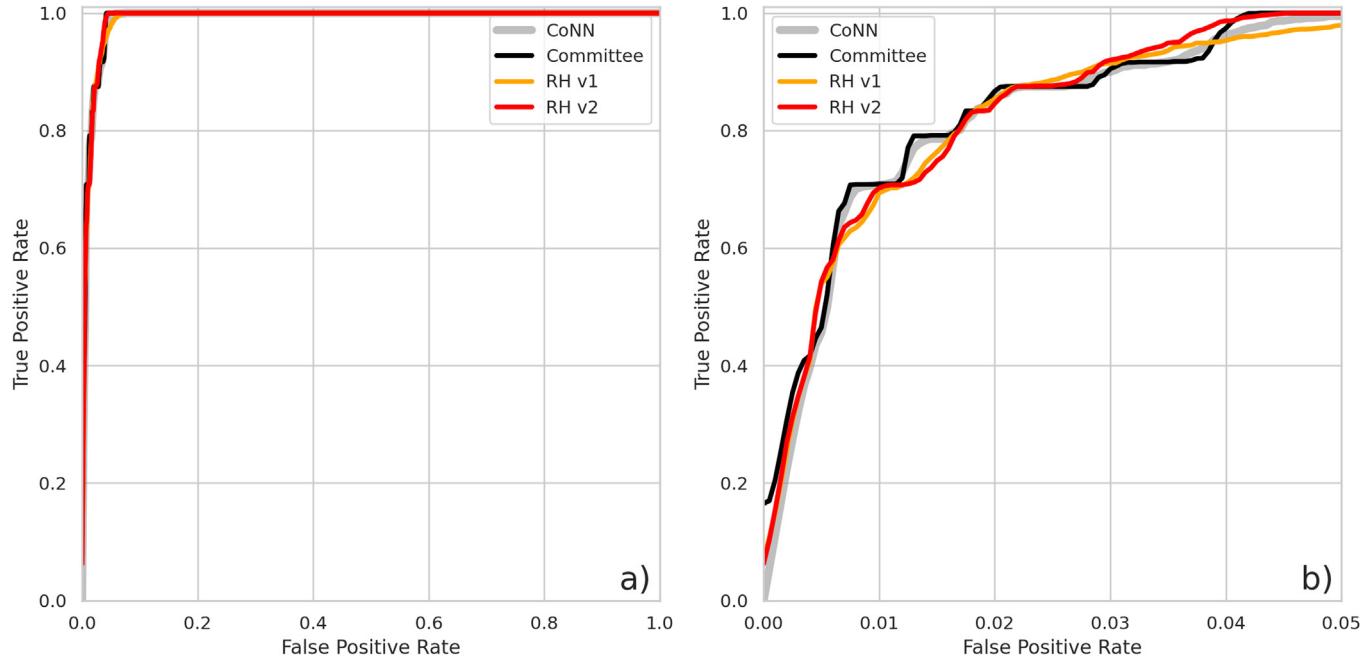


Fig. 6. (a) Receiver Operating Characteristic (ROC) curves averaged over 50 experiments for the classifiers considered in the current work. (b) The same curves zoomed in for the small False Positive Rates only.

or an increase from $HSS = 0.163 \pm 0.026$ to $HSS = 0.168 \pm 0.005$, $HSS = 0.163 \pm 0.010$, and $HSS = 0.168 \pm 0.008$ when transitioning from the CoNN to ensembles. This demonstrates that the ensemble approaches are performing better, on average, with respect to the CoNN classifier. On the other hand, we point out here that both the TSS and HSS scores for all classifiers almost always intersected within the uncertainties (either the standard deviation or the median absolute deviation) arising from the results of individual experiments. The overall closeness of the performance of the classifiers is evident as well from the ROC curves presented in Fig. 6 which experience a significant overlap. The good performance of the ensemble classifiers was previously noticed in the works of Aminalragia-Giamini et al. (2021) for the Committee approach and Lavasa et al. (2021) for the random forest classifier. Interestingly, the case performances of the CoNN classifier may even outperform the individual ensemble classifier tests (as evident from the upper boundary for interquartile ranges for CoNN in Fig. 5), which reveals the importance of evaluation of the methods on several train-test splits and demonstration of its robustness with respect to the random splitting.

Another pattern evident from Tables 1 and 2 is the noticeable differences between the standard deviations or median absolute deviations for the CoNN and ensemble classifiers. For example, the mean TSS score and its standard deviation for CoNN is $TSS = 0.906 \pm 0.043$ compared to the $TSS = 0.926 \pm 0.035$ (Committee), $TSS = 0.915 \pm 0.029$ (RH v1), and $TSS = 0.944 \pm 0.023$ (RH v2). The tendency is even sharper for the medians and median absolute deviations (compare

$TSS = 0.906 \pm 0.042$ for CoNN with $TSS = 0.926 \pm 0.023$, $TSS = 0.915 \pm 0.010$, and $TSS = 0.944 \pm 0.005$ for aforementioned ensemble approaches). The tendency of lower spread remains the same for all other scores considered in the tables. Fig. 5 also indicates that the interquartile range and the span of whiskers (indicating the spread of individual experiments outside of the interquartile range) is typically larger for CoNN, especially in the case of HSS skill score. Overall, such behavior indicates the relative robustness of the ensemble approaches with respect to the random train-test splits for the data set and the training process, while the training of the individual classifiers may fail. Therefore, the increase in the complexity of these ensemble algorithms is justified by their robust performance on the imbalanced data sets (Galar et al., 2012).

Tables 1 and 2 also demonstrate that the RH classifier (1) does not necessarily outperform the Committee approach, and (2) depends on the selection of its parameters. This is concluded from the fact that the performance of the RH v2 is, on average, the same or better, with respect to almost any metrics than that of the Committee or RH v1. Another noticeable difference between the Committee and RH classifiers is the step-wise behavior of the ROC curve for the Committee compared to a smoother curve for the RH, evident in Fig. 6. The key difference between the RH and the Committee classifier is in the selection of features used for each individual ensemble member. While the committee members use all features available, RH members use the down-selected number of features (either 4 or 6 out of 12 in our case) and use the deterministic algorithm of the contribution of each Committee member

to the final result. We may assume that the down-selection of features for each member increases the forecasting scores because it helps to filter out the attention of the individual committee member to the noisy or irrelevant features and prevents its members from “agreeing” on identical sets of model parameters (weights and biases). This confirms the importance of the feature selection process, which remains an active topic in predicting solar transient events (Bobra and Couvidat, 2015; Sadykov and Kosovichev, 2017; Yeolekar et al., 2021). Also, although the Committee approach (Aminalragia-Giamini et al., 2021) helps to reduce the “reliance on chance” in terms of the convergence of the network parameters (weights and biases) to the local or global minima, it still contains similarly-structured CoNNs as ensemble members. The RH introduces a more diverse population of ensemble members with the variable down-selected set of features as an input, which can be more beneficial than having full-scale but nearly identical learners.

As noted earlier, the results in [Tables 1 and 2](#) indicate relatively low values for the precision and HSS scores for all classifiers tested, including the RH classifiers. At the same time, the corresponding TSS scores are high. To understand the reason behind this behavior of the models, let us indicate some patterns in our SEP prediction. Looking at the median confusion matrix elements, one can notice the RH classifiers that, arranged by larger to smaller, $TN \approx 5244 \gg FP \approx 226 \gg TP \approx 22.9 \gg FN \approx 1.1$ and that $TN \approx 5245 \gg FP \approx 225 \gg TP \approx 23.6 \gg FN \approx 0.4$. Assuming that one can neglect the term of the next order of smallness, one can rewrite the metrics of interest as:

$$Precision = \frac{TP}{TP + FP} \approx \frac{TP}{FP} \quad (6)$$

$$HSS = \frac{2 \cdot (TP \cdot TN - FP \cdot FN)}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)} \approx \frac{2 \cdot TP \cdot TN}{TP \cdot TN + FP \cdot FN} \approx \frac{2 \cdot TP}{TP + FP} = 2 \cdot Precision \quad (7)$$

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN} = 1 - \frac{FN}{TP + FN} - \frac{FP}{FP + TN} \approx 1 - \frac{FN}{TP} - \frac{FP}{TN} \rightarrow 1 \quad (8)$$

Both the precision and HSS scores, under the conditions for the confusion matrix elements indicated above, are determined by the TP/FP ratio, which is of the order of $\sim 1/10$ for the models implemented in this study. At the same time, the TSS score is very close to 1 because both subtrahends are small: $FN/TP \sim 1/20$ or smaller, and $FP/TN \sim 1/23$. While capturing almost every single SEP flare in the test data set (which is 24 events on average), the models produce almost ten times larger number of false alarms than the number of SEP flares (with the median values of $FP = 225.7 \pm 20.0$ and $FP = 224.8 \pm 15.5$ for RH v1 and RH v2, correspondingly). The considered data set is also highly imbalanced given that the ratio of the positive to negative samples is $\sim 1/285$. The conditions above limit the HSS score to $HSS \approx 2 \times TP/FP \leq 2 \times P/FP$. They also make the HSS scores to be very susceptible to the change in

the number of false alarms. For example, if one manages to decrease the number of false alarms twice ($FP = 224.8 \rightarrow FP = 112.4$ and $TN = 5245.2 \rightarrow TN = 5357.6$ for RH v2), the corresponding HSS score would increase almost twice, to $HSS \approx 0.290$. At the same time, the TSS score would increase from $TSS \approx 0.944$ to $TSS \approx 0.963$, experiencing a mild relative change. One can conclude that for highly imbalanced data sets the small HSS scores are related to the large number of false alarms produced by the model (with respect to the number of positive samples in the test data set) and may still be accompanied by TSS scores very close to one.

Also, we note that while the Heidke Skill Score 2 (HSS) is often annotated as a measure of the performance with respect to a random chance forecast, the forecast presented here is definitely far from being random: with RH v1, the missed event rate $FN/(FN + TP) \approx 0.046$ is low (almost every SEP event is hit for each trial) and the false alarm rate $FP/(FP + TN) \approx 0.041$ is low as well, while with RH v2, the missed event rate is even more promising $FN/(FN + TP) \approx 0.017$ and the false alarm rate is at the still the same value of approximately 0.041. Nevertheless, the HSS scores are not so strongly deviating from 0 (on average, $HSS = 0.163 \pm 0.014$ with RH v1 and $HSS = 0.168 \pm 0.013$ with RH v2). Therefore, we argue that it is not correct to associate low HSS scores with the forecast being close to a random chance forecast. Moreover, the example in the previous paragraph demonstrates that it is very beneficial to consider the HSS scores complementary to the TSS scores for forecasting purposes. The HSS score is much more sensitive to the decrease of the false alarms, FP , if compared to the TSS . Therefore, it would allow one to better differentiate between the models with approximately the same number of missed events based on the number of false alarms they produce.

Although not tested for the all-clear forecasting explicitly, the classification approaches implemented here demonstrate usefulness with respect to the all-clear setting. We note here that the all-clear SEP forecasting is typically defined as the whole-disk endeavor which also has to be accompanied by the specification of the time window for which the forecast is issued. Here we discuss a related, but not identical, problem when one would like to predict very reliably every SEP-active flare (with the missed events being very undesirable) yet maintaining a low false alarm rate. Here the very low rate of missed events (the average rate of the missed events is $FN/(FN + TP) \approx 0.05$ even for CoNN) is what is typically desirable for such a type of forecast (Sadykov et al., 2021). For the RH v2 classifier, the virtually zero rate of missed events ($FN/(FN + TP) \approx 0.017$) renders the RH viable for this task. Although the median number of the FN was still relatively low for the CoNN classifier ($FN = 1.2 \pm 1.5$, see [Table 2](#)) and the Committee approach ($FN = 0.8 \pm 0.5$), the RH v2 has even lower $FN = 0.4 \pm 0.0$. The median absolute deviation of 0.0 indicates that more than half

(i.e., more than 25 out of 50) trials of RH v2 had zero false negative predictions as an outcome. Overall, we can also potentially expect, that certain configurations of the RH approach can be useful in delivering near-zero missed event forecasting with low false-alarm rates for other machine-learning problems, including those involving highly imbalanced classes.

5. Conclusion

In this work, we have introduced an ensemble algorithm — a Random Hivemind (RH) — and compared two selected implementations of it with respect to the Conventional Neural Network (CoNN) and a Committee ensemble approach for CoNNs. The comparison was done for the problem of the prediction of Solar Energetic Particle (SEP) events based on the properties of the host soft X-ray flares. The key outcomes of our work are as follows:

- The performance of the RH algorithm depends on the implementation and training parameters (which may include the number of input features, their weighting schemes, learning rate boosting strategies, etc.). The corresponding TSS and HSS scores for the RH v1 and RH v2 implementations are $TSS = 0.915 \pm 0.010$ and $TSS = 0.944 \pm 0.005$, and $HSS = 0.163 \pm 0.010$ and $HSS = 0.168 \pm 0.008$, correspondingly.
- Both ensemble approaches (Committee and RH) demonstrate the robustness of their performance with respect to the random train-test splits for the data set, which was reflected in the low standard deviations or median absolute deviations. Although often performing comparably to the committee approach in terms of the forecasting metrics, CoNN demonstrated much higher standard deviations (and often higher median absolute deviations).
- Both ensemble approaches demonstrated similar or better performance in terms of mean and median values compared to the metrics typically used in space weather forecasting, *HSS* and *TSS*. One can compare the median $TSS = 0.906 \pm 0.042$ for CoNN with $TSS = 0.926 \pm 0.023$, $TSS = 0.915 \pm 0.010$, and $TSS = 0.944 \pm 0.005$ for the committee, RH v1, and RH v2 correspondingly, and $HSS = 0.163 \pm 0.026$ with $HSS = 0.168 \pm 0.005$, $HSS = 0.163 \pm 0.010$, and $HSS = 0.168 \pm 0.008$.
- The RH v2 ensemble classifier performs better, on average, than the Committee, CoNN, and RH v1 approaches in terms of almost every metric and delivers consistent results over the ten random train-test split experiments.
- The performance of all classifiers, including RH, demonstrated relatively low precision and *HSS* scores for the SEP prediction problem. Nevertheless, it is very beneficial to consider the *HSS* as a complementary metric for the forecast as it is more susceptible to a decrease in the false alarm rate than the *TSS*. This leads to a bet-

ter differentiation between the models with approximately the same number of missed events based on the number of false alarms they produce.

- All classifiers had a very low number of false negative predictions. Median values of $FN = 1.2 \pm 1.5$, $FN = 0.8 \pm 0.5$, $FN = 1.1 \pm 0.0$, and $FN = 0.4 \pm 0.0$, were measured for the CoNN, Committee, RH v1, and RH v2 classifiers, respectively. However, the robustness of the RH classifiers noted previously, especially for the case of RH v2, makes it the viable candidate for employment in solving the “all-clear”-like forecasting problem for SEP-active flares.

From the results above, we can conclude that RH is a valid machine learning algorithm that can perform well despite class imbalance. RH is performing, on average, comparably or better to CoNNs and unweighted, identical CoNN committee machines. Further studies of the RH approach (including different implementations for the feature weights and handling, learning rate and epoch number adjustments, and the flare class boundaries considered for RH training) are required to understand its potential in general and specifically for space weather prediction purposes, including “all-clear” forecasting of SEPs.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank the anonymous referees for the suggestions which significantly improved the quality of the manuscript. This research was supported by NASA Early Stage Innovation program grant 80NSSC20K0302, NASA LWS grant 80NSSC19K0068, NSF EarthCube grants 1639683, 1743321, and 1927578, and NSF grant 1835958. VMS acknowledges the NSF FDSS grant 1936361 and NSF grant 1835958.

References

- Afaq, S., Rao, S., 2020. Significance of epochs on training a neural network. *Int. J. Scient. Technol. Res.* 9 (6), 485–488.
- Ahmadzadeh, A., Aydin, B., Georgoulis, M.K., et al., 2021. How to train your flare prediction model: revisiting robust sampling of rare events. *Astrophys. J. Suppl. Ser.* 254 (2), 23. <https://doi.org/10.3847/1538-4365/abec88>, arXiv:2103.07542.
- Ali, A., Sadykov, V., Kosovichev, A. et al., 2023. Predicting Solar Proton Events of Solar Cycles 22-24 using GOES Proton & Soft X-Ray Flux Statistics. *arXiv e-prints*, (p. arXiv:2303.05446). doi:10.48550/arXiv.2303.05446. arXiv:2303.05446.
- Aminalragia-Giamini, S., Raptis, S., Anastasiadis, A., et al., 2021. Solar energetic particle event occurrence prediction using solar flare soft x-ray measurements and machine learning. *J. Space Weather Space Clim.* 11, 59. <https://doi.org/10.1051/swsc/2021043>.

- Bain, H.M., Steenburgh, R.A., Onsager, T.G., et al., 2021. A summary of national oceanic and atmospheric administration space weather prediction center proton event forecast performance and skill. *Space Weather* 19 (7). <https://doi.org/10.1029/2020SW002670>.
- Balch, C.C., 1999. SEC proton prediction model: verification and analysis. *Radiat. Meas.* 30 (3), 231–250. [https://doi.org/10.1016/S1350-4487\(99\)00052-9](https://doi.org/10.1016/S1350-4487(99)00052-9).
- Balch, C.C., 2008. Updated verification of the Space Weather Prediction Center's solar energetic particle prediction model. *Space Weather* 6 (1), S01001. <https://doi.org/10.1029/2007SW000337>.
- Bobra, M.G., Couvidat, S., 2015. Solar Flare Prediction Using SDO/HMI vector magnetic field data with a machine-learning algorithm. *Astrophys. J.* 798 (2), 135. <https://doi.org/10.1088/0004-637X/798/2/135>, arXiv:1411.1405.
- Bouabrahimi, S.F., Aydin, B., Martens, P., et al., 2017. On the prediction of >100 mev solar energetic particle events using goes satellite data. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 2533–2542. <https://doi.org/10.1109/BigData.2017.8258212>.
- Breiman, L., 2001. Random Forests. *Mach. Learn.* 45, 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Chen, Y., Kempton, D.J., Ahmadzadeh, A., et al., 2021. Towards synthetic multivariate time series generation for flare forecasting. In: Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., Zurada, J.M. (Eds.), *Artificial Intelligence and Soft Computing*. Springer International Publishing, Cham, pp. 296–307.
- Cunningham, P., 2000. Overfitting and diversity in classification ensembles based on feature selection. *Trinity's Access to Research Archive*, URL: <http://hdl.handle.net/2262/13078>.
- Galar, M., Fernandez, A., Barrenechea, E., et al., 2012. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst., Man, Cybernet., Part C (Appl. Rev.)* 42 (4), 463–484. <https://doi.org/10.1109/TSMCC.2011.2161285>.
- Garcia, H.A., 1994. Temperature and hard x-ray signatures for energetic proton events. *Astrophys. J.* 420, 422. <https://doi.org/10.1086/173572>.
- Garcia, H.A., 2004. Forecasting methods for occurrence and magnitude of proton storms with solar soft X rays. *Space Weather* 2 (2), S02002. <https://doi.org/10.1029/2003SW000001>.
- Kahler, S.W., Ling, A.G., 2018. Forecasting Solar Energetic Particle (SEP) events with Flare X-ray peak ratios. *J. Space Weather Space Clim.* 8, A47. <https://doi.org/10.1051/swsc/2018033>.
- Kahler, S.W., Ling, A.G., 2022. A comparison of solar x-ray flare timescales and peak temperatures with associated coronal mass ejections. *Astrophys. J.* 934 (2), 175. <https://doi.org/10.3847/1538-4357/ac7e56>.
- Kahler, S.W., Vourlidas, A., 2013. A comparison of the intensities and energies of gradual solar energetic particle events with the dynamical properties of associated coronal mass ejections. *Astrophys. J.* 769 (2), 143. <https://doi.org/10.1088/0004-637X/769/2/143>.
- Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.
- Laurenza, M., Alberti, T., Cliver, E.W., 2018. A short-term esperta-based forecast tool for moderate-to-extreme solar proton events. *Astrophys. J.* 857 (2), 107. <https://doi.org/10.3847/1538-4357/aab712>.
- Laurenza, M., Cliver, E.W., Hewitt, J., et al., 2009. A technique for short-term warning of solar energetic particle events based on flare location, flare size, and evidence of particle escape. *Space Weather* 7 (4), S04008. <https://doi.org/10.1029/2007SW000379>.
- Lavasa, E., Giannopoulos, G., Papaioannou, A., et al., 2021. Assessing the predictability of solar energetic particles with the use of machine learning techniques. *Sol. Phys.* 296 (7), 107. <https://doi.org/10.1007/s11207-021-01837-x>.
- Ling, A.G., Kahler, S.W., 2020. Peak temperatures of large solar x-ray flares and associated CME Speeds and Widths. *Astrophys. J.* 891 (1), 54. <https://doi.org/10.3847/1538-4357/ab6f6c>.
- Nishizuka, N., Kubo, Y., Sugiura, K., et al., 2020. Reliable probability forecast of solar flares: deep flare net-reliable (DeFN-R). *Astrophys. J.* 899 (2), 150. <https://doi.org/10.3847/1538-4357/aba2f2>, arXiv:2007.02564.
- O'Keefe, P.M., Sadykov, V.M., Kosovichev, A.G., et al., 2022. Handling highly imbalanced data in machine learning applications. *Zenodo*. <https://doi.org/10.5281/zenodo.6780972>, URL: <https://doi.org/10.5281/zenodo.6780972>.
- Papaioannou, A., Sandberg, I., Anastasiadis, A., et al., 2016. Solar flares, coronal mass ejections and solar energetic particle event characteristics. *J. Space Weather Space Clim.* 6, A42. <https://doi.org/10.1051/swsc/2016035>.
- Reames, D.V., 2021. Solar energetic particles. A modern primer on understanding sources, *Accel. Propag.* 978. <https://doi.org/10.1007/978-3-030-66402-2>.
- Richardson, I.G., Mays, M.L., Thompson, B.J., 2018. Prediction of solar energetic particle event peak proton intensity using a simple algorithm based on CME speed and direction and observations of associated solar phenomena. *Space Weather* 16 (11), 1862–1881. <https://doi.org/10.1029/2018SW002032>.
- Ryan, D.F., Milligan, R.O., Gallagher, P.T., et al., 2012. The thermal properties of solar flares over three solar cycles using GOES X-ray observations. *Astrophys. J. Suppl. Ser.* 202 (2), 11. <https://doi.org/10.1088/0067-0049/202/2/11>, arXiv:1206.1005.
- Sadykov, V., Kosovichev, A., Kitiashvili, I., et al., 2021. Prediction of Solar Proton Events with Machine Learning: Comparison with Operational Forecasts and “All-Clear” Perspectives. *arXiv e-prints*, (p. arXiv:2107.03911). doi:10.48550/arXiv.2107.03911. arXiv:2107.03911.
- Sadykov, V.M., Kosovichev, A.G., 2017. Relationships between characteristics of the line-of-sight magnetic field and solar flare forecasts. *Astrophys. J.* 849 (2), 148. <https://doi.org/10.3847/1538-4357/aa9119>, arXiv:1704.03423.
- Sadykov, V.M., Kosovichev, A.G., Kitiashvili, I.N., et al., 2019. Statistical properties of soft x-ray emission of solar flares. *Astrophys. J.* 874 (1), 19. <https://doi.org/10.3847/1538-4357/ab06c3>, arXiv:1810.05610.
- Sadykov, V.M., Kosovichev, A.G., Oria, V., et al., 2017. An interactive multi-instrument database of solar flares. *Astrophys. J. Suppl. Ser.* 231 (1), 6. <https://doi.org/10.3847/1538-4365/aa79a9>, arXiv:1702.02991.
- Stumpo, M., Benella, S., Laurenza, M., et al., 2021. Open issues in statistical forecasting of solar proton events: a machine learning perspective. *Space Weather* 19 (10). <https://doi.org/10.1029/2021SW002794>, e2021SW002794.
- Tietz, M., Fan, T.J., Nouri, D., et al., 2017. skorch: A scikit-learn compatible neural network library that wraps PyTorch. URL: <https://skorch.readthedocs.io/en/stable/>.
- Torres, J., Zhao, L., Chan, P.K., et al., 2022. A machine learning approach to predicting sep events using properties of coronal mass ejections. *Space Weather* 20 (7). <https://doi.org/10.1029/2021SW002797>, e2021SW002797.
- Yeolekar, A., Patel, S., Talla, S., et al., 2021. Feature selection on a flare forecasting testbed: A comparative study of 24 methods. In: In 2021 International Conference on Data Mining Workshops (ICDMW). IEEE Computer Society, Los Alamitos, CA, USA, pp. 1067–1076. <https://doi.org/10.1109/ICDMW53433.2021.00138>, URL: <https://doi.ieeecomputersociety.org/10.1109/ICDMW53433.2021.00138>.
- You, K., Long, M., Wang, J., et al., 2019. How Does Learning Rate Decay Help Modern Neural Networks? *arXiv e-prints*, (p. arXiv:1908.01878). doi:10.48550/arXiv.1908.01878. arXiv:1908.01878.