Machine Learning to Classify Electrostatic Plasma Turbulence in Tokamaks

Hong-Jian Zhao ¹ , Zhe-Hua Guo ² , Xiang-Yu Wu ¹ and Yong Xiao¹

¹ Department of Physics, Zhejiang University, Hangzhou 310027, China ² Meta.com**

**This work is not supported by this author's affiliation.

Corresponding E-mail[: yxiao@zju.edu.cn](mailto:yxiao@zju.edu.cn)

Abstract

Artificial intelligence driven by large data is becoming increasingly important in magnetic fusion research. Here we scan the plasma gradient space for Cyclone Base case parameters using nonlinear gyrokinetic simulations to generate data for typical electrostatic drift wave turbulences. The main candidates, ion temperature gradient mode (ITG), and trapped electron mode (TEM), are then classified and labeled by conventional methods for the datasets in the linear stage. We then apply a classical machine learning algorithm, namely the support vector machine (SVM), and use plasma gradients or feature turbulence quantities as the input to classify the type of the drift wave turbulence. Simple distance formulae are obtained for fast classification of the turbulence type and justified for their effectiveness, which can be used for future theoretical analysis and real-time experiment.

Keywords: turbulence, support vector machine, ion temperature gradient mode, trapped electron mode

1. Introduction

Turbulence prevails in nature and laboratory with multiple temporal and spatial scales. In magnetic fusion plasmas, energy and particle transport caused by electromagnetic fluctuations or turbulence are much larger than those caused by collisional processes [1]. The confining capability of magnetic fusion device is critically determined by the turbulent transport largely caused by drift wave instabilities [2,3], which is usually excited by free energy stored in the radial temperature or density gradients of the fusion plasmas. Different temperature and density gradients corresponds to a variety of drift wave instabilities, typically including the ion temperature gradient mode (ITG) [4], trapped electron mode (TEM) [5], electron temperature mode (ETG) [6], etc. It is important to identify the type of these instabilities that drives the turbulence, which provides crucial information that could be used to predict the transport level for a given set of plasma parameters.

Artificial intelligence (AI) such as machine learning has been recently pursued to predict the occurrences of tokamak disruption [7] and proactively control the plasma profile evolution. For example, a number of surrogate transport models has been built to predict turbulent transport or kink mode based on the machine learning [8,9,10]. In this work, we try to use machine learning, specifically the support vector machine (SVM) method [11,12], to help classify different types of turbulence. This method has been used for effectively classifying experimental data in the low or high confinement mode of tokamaks [13], with a hyperplane generated for segmentation after the experimental data is properly transformed in the parameter space. The turbulence data for classification is obtained by first-principles gyrokinetic simulations using the Gyrokinetic Toroidal Code (GTC), which is widely used in the fusion community for simulate plasma turbulence in fusion devices such as tokamak and stellarator [14,15]. The GTC code has successfully demonstrated capabilities in simulating ITG, TEM, ETG, kinetic ballooning mode (KBM) and other important drift wave instabilities, as well as producing reliable turbulence data for the machine learning analysis [16].

In this paper, we first introduce the working framework in section 2. Then the details of dataset generation are given in section 3, including linear instability data and nonlinear turbulence data. In section 4, the feature selection for the turbulence data is carried out. Next, the SVM classification of turbulence is presented in section 5, using either instability drive or featured turbulent quantities as the classifiers. In the end, we conclude this paper with a short summary and some discussions.

2. Working Framework

In this work, we use the Gyrokinetic Toroidal Code (GTC) to generate a large amount of data for machine learning [17]. GTC is a three-dimensional global gyrokinetic particle simulation code, where the electron is simulated by a fluid-kinetic hybrid model and the ion is simulated by a gyrokinetic model [18]. Using the smallness of electron-ion mass ratio m_e/m_i , the electron response can be treated by a perturbative expansion. The zero order of the expansion reproduces adiabatic response for the electron, and the higher order expansion accounts for wave-particle resonance from parallel and perpendicular motion.

When generating the requisite datasets for machine learning, each GTC simulation can consume a significant amount of computing time. For example, it takes about 6.5 hours to run on the Tianhe-3F supercomputer with 1024 nodes for simulating 500 time steps for a typical TEM case. Therefore, it is rather expensive to generate a large number of sample datasets. Since no collisions are included in the simulation, the TEM mode in the works actually refers to the collisionless trapped electron mode or CTEM in the literature [19]. Moreover, in the nonlinear stage, the turbulence quantities fluctuate a lot with time. So it is generally not straightforward to train a large number of data from the saturated turbulence and nonlinearly fit these turbulence quantities using common AI tools such as machine learning or deep learning. However, in this work we focus on a relatively limited goal, i.e., to classify different types of turbulence by their nonlinear properties rather than the linear properties. Meanwhile, we find that in order to identify the type of instability with a certain number of parameters using the conventional method, a large number of linear simulations is required to find the dispersion relation for determining the instability type by comparing results with the linear theory. Figure 1 shows the dispersion relation from the linear simulation for ITG and CTEM, respectively. Each data point for the linear frequency or growth rate is computed from the linear simulation, which is approximately equivalent to several nonlinear simulations in terms of computing time due to the time-consuming numerical filtering procedure in the global particle simulation.

On the other hand, we note that the support vector machine (SVM) method from the machine learning has the advantage of using fewer training datasets to achieve better classification accuracy. Therefore, we here try a different approach and use the SVM method on the turbulent quantities from a single nonlinear gyrokinetic simulation to classify the type of the instability. Moreover, this approach can be used directly in experiments to fast determine turbulence type in real time without resorting to the plasma equilibrium reconstruction and conventional time-consuming linear gyrokinetic simulations.

Figure 1. Linear dispersion relation calculated by GTC with parameters: (a) $R/L_{Ti} = 6.92$, $R/L_{Te} = 6.92$, and $R/L_n = 2.22$ for ITG; and (b) $R/L_{Ti} = 2.22$, $R/L_{Te} = 6.92$, and $R/L_n = 2.22$ for TEM.

3. Dataset generation and selection

3.1. Generation of dataset

The prevailing drift wave turbulence in tokamaks is simulated by the GTC code in the collisionless electrostatic limit with the following cyclone base case parameters [20]: the density $n_0 = 0.79 \times 10^{20} / m^3$, the temperature $T_e = T_i = 2.22 \text{keV}$, $m_i/m_e = 1837$, safety factor $q = 1.4$ and magnetic shear $s = rdq/qdr = 0.78$ on the reference magnetic surface at $r = a/2$. In the simulation, we change the values for the normalized density gradient $R/L_n = R_0 dn/n dr$, normalized electron temperature gradient $R/L_{Te} = R_0 dT_e/T_e dr$ and normalized ion temperature $R/L_{Ti} = R_0 dT_i/T_i dr$ in the typical 3D parameter space with a scanning of $6\times6\times6$ matrix. The corresponding gradient values for each gradient dimension are shown in Table 1.

After skipping 13 sets of plasma gradients with which no instability is found, we used a set of total 203 plasma gradient combinations to carry out nonlinear GTC simulations and a large amount of data is produced from the simulated turbulence. Then we proceed to classify various electrostatic drift wave instabilities from these turbulence data. First we find that the dominant frequencies are away below the electron transit frequency, i.e., $\omega/k_{\parallel}v_{te}$ ~0.02 \ll 1. Thus, the contribution of the passing electrons is dominantly adiabatic and we shall focus on the kinetic response of the trapped electrons. This result is further verified by comparing the linear growth rates with and without the passing electrons, as is shown in Figure 2. We can see that for all cases, the relative difference between the growth rates is less than 20%. Therefore, with the selected parameters, the passing electrons respond adiabatically and contribute little to the instability. So when the electrons are the proactive species that drives the instability, the instability can only be considered as the TEM mode, with the ETG mode excluded from the candidate instabilities.

Figure 2. The difference of the growth rate between the cases with only trapped electrons and the case with both trapped electrons and passing electrons, where γ_{ot} refers to the growth rate from the cases with only trapped electrons.

In general, ITG and TEM can co-exist as two branches of the drift wave instability. It is relatively easier to distinguish these two instabilities in the parameter regions where only one branch of modes is excited and the other branch of modes is purely damping [21]. For example, in Case #161 where $\eta_i = L_n/L_{Ti} = 1.0$ and $\eta_e = L_n/L_{Te} = 4.4$, the ITG branch cannot be excited and this case can only be considered as TEM-dominant; in Case #116 where $\eta_i = 4.0$ and $\eta_e = 1.0$, the instability is ITG-dominant. Therefore, we can determine the type for the instability in these parameter regions with minimal efforts.

However, for other cases, it is difficult to distinguish between ion mode and electron mode since both of them can be unstable. The GTC code uses the δf algorithm to calculate linear instability and simulate nonlinear turbulence in a sense of perturbation theory [22]. With this δf algorithm, The entropy of the particle species α in the simulation is represented by $\iint d^3x d^3v \delta f_\alpha^2$. When the instability grows up, the entropy will keep increasing with time. By comparing the initial growths of the entropy for ions and electrons, we can determine which particle species dominates the instability. This unique feature can be used to judge whether the type of instability is driven by ions or electrons, i.e., ITG or TEM.

3.2. Sequence selection

The time history of turbulence amplitude or fluctuating electric potential amplitude ϕ^{rms} is shown in Figure 4 for the turbulence formation and evolution. After the linear exponential growth, the potential amplitude ϕ^{rms} saturates at a certain level after some transient bursts. In the experiment, we can only measure turbulent data via plasma diagnostics at the fully developed turbulent stage. The turbulent phase is often most interesting for analysis and modelling. Therefore, we select the simulation data in the well saturated stage to represent the typical turbulence quantities.

The SVM method from machine learning can be employed to help with the data selection. We choose a short sequence of data $x = \{x_1, x_2, \dots, x_s\}$ as an elementary dataset, with x the selected physical quantity from the simulation. The input features for this SVM procedure are $(\gamma(x), \overline{x}, \overline{(x - \overline{x}^2)^2})$, where $\gamma(x) = (x_s - x_1)/s$, \overline{x} is the mean value of the dataset, $\overline{(x - \overline{x}^2)^2}$ is the mean variance of the dataset. The output for this SVM procedure is the evolution stage of this dataset, namely linear stage, turbulent stage and the divergence stage. By using this SVM method, an example of automatic data selection is shown in Figure 4, with the data between the vertical lines selected as the target data sequence in the turbulent stage. Then we carry out a time average for the selected data sequence to obtain one data point. In this way, a total of 1411 data points are generated by varying the density and temperature gradients shown in Table 1 and comparing initial growth of the electron and ion entropy, among which 826 data points were identified as TEM and the remaining 586 data points were identified as ITG, with a ratio of approximately 6:4.

Figure 4. Time history of fluctuating electric potential amplitude ϕ^{rms} during the instability-turbulence evolution.

3.3. Feature selection

Next we ought to determine what physical quantities should be selected as the input features for the SVM classification algorithm. First, we calculate the correlations between some physically important candidate features: electric potential amplitude ϕ^{rms} , zonal flow potential ϕ^{rms}_{00} , ion particle diffusivity D_i , ion heat conductivity χ_i , and the corresponding electron transport quantities D_e , χ_e . The correlation coefficient $Corr(\alpha, \beta)$ for arbitrary physical quantities α and β is calculated by the following formula:

$$
Corr(\alpha, \beta) = \frac{\sum_{i} \alpha_{i} \beta_{i}}{\sqrt{\sum_{i} \alpha_{i}^{2} \sum_{i} \beta_{i}^{2}}}
$$
\n(1)

The resultant correlation coefficients between different physical quantities are shown in Table 2. We notice that $Corr(\phi^{rms}, \phi_{00}^{rms}) = 0.89$ and $Corr(D_i, D_e) = 1.00$, which shows that the physical quantities in these pairs can be interchanged for each other. This makes physics sense since the zonal flows are self-consistently generated by the turbulence, and the radial ambipolarity is well preserved for turbulent particle flux as it should be. Thus, we can remove ϕ_{00}^{rms} and D_e from the feature list to ensure the independence of each feature quantity. The rest feature quantities have no strong correlations or dependences between them and can be considered as independent variables.

		ϕ_{00}^{rms}	D_i χ_i D_e	χ_e
ϕ^{rms}	$1.00\,$		0.89 0.47 0.38 0.46 0.37	
ϕ^{rms}_{00}			0.89 1.00 0.40 0.28 0.40 0.28	
D_i	$\begin{vmatrix} 0.47 & 0.40 & 1.00 & -0.25 & 1.00 & 0.33 \end{vmatrix}$			
χ_i	$\begin{bmatrix} 0.38 & 0.28 & -0.25 & 1.00 & -0.25 & 0.10 \end{bmatrix}$			
D_e			0.46 0.40 1.00 -0.25 1.00 0.33	
χ_e	$\begin{bmatrix} 0.37 & 0.28 & 0.33 & 0.10 & 0.33 & 1.00 \end{bmatrix}$			

Table 2. The correlation coefficients between the selected alternative features.

Finally, we employ the linear SVM method with L1 regularization [23] to conduct the feature selection in the machine learning process. This method can reduce the model complexity and thus avoid overfitting. With the penalty coefficient C=0.0125, ϕ^{rms} is filtered out of the model. What is more, it is difficult for experimental diagnostics to obtain the electric potential fluctuations in the core plasmas due to its extreme high temperature. Thus, even from the experimental point of view, it is unfavourable to include ϕ^{rms} in the physics modelling. Next we employ the 5-fold cross-validation [11] to find out that the model with the aforementioned three physical quantities is accurate enough for the turbulence classification, as is shown in Figure 5. We note that including extra physical quantities would not introduce more accuracy for the classification. Based on these arguments, we finally choose three transport coefficients, D_i , χ_i , and χ_e , as the input features to classify the turbulence type. We normalize these features by the gyroBohm diffusivity $D_{gB} \equiv \rho_i e B_0 / aT_e$, where $\rho_i = \sqrt{m_i T_i / (e B_0)}$ is the ion gyroradius, $a = 0.36R_0$ with a the tokamak minor radius and R_0 the tokamak major radius and B_0 is the magnetic field at the magnetic axis.

Figure 5. The accuracy of linear SVM classification on the validation set versus the complexity of the physics models.

4. SVM classification results

In this section, we will use SVM to classify the dataset from two perspectives. First we will classify the instability type using the simulation parameters or the instability drives as the input quantities, which is similar to some previous literatures [24,25]. And then we use the selected feature quantities from the turbulent stage as the input quantities for classification. In this way, we are able to classify the instability type by the turbulence quantities, which caters more to experimental needs.

4.1. Parameter classification

We use the python library of Scikit-Learn to implement the SVM algorithm and perform the classification with the simulation input parameters first, i.e., the instability drives or the plasma temperature and density gradients. The normalized characteristic gradients(R/L_n , R/L_{Ti} , R/L_{Te}) are regarded as the input features and the type of instability is classified according to the aforementioned entropy-based criterion, as is shown in Figure 6, where the blue dots represent TEM-dominant cases and the red dots represents the ITG-dominant cases. The dividing plane is found by the linear SVM classification, with an accuracy as high as 98%. This demonstrates that the SVM classification can correctly determine the type of instability by using the plasma gradient combination. In addition, we can also obtain an analytic formula for the dividing plane with three alternative free parameters, R/L_n , η_e and η_i :

$$
\tilde{d} = 1 - \frac{2.1}{R/L_n} + 0.63\eta_e - 1.0\eta_i
$$
 (2)

It is found that the TEM mode dominates for $\tilde{d} > 0$ and the ITG mode dominates for $\tilde{d} < 0$. The conclusion is consistent with the existing theory that in a toroidal system, the ITG mode is driven by η_i , while larger values of R/L_n and η_e tend to excite the TEM mode[26]. This analytic formula can help us to fast determine what kind of instability for a given set of instability-driven plasma gradients, and also to locate the appropriate parameter region for exploring the ITG-TEM transition physics.

Figure 6. Turbulence classification for the simulation cases with varying gradient drives $(R/L_n, R/L_{Ti}, R/L_{Te})$ in the parameter space, with the dividing plane generated by the SVM algorithm.

4.2. Turbulence classification

In many cases, the feature physical quantities during turbulent stage can be measured by advanced diagnostics [27,28]. We can directly use these turbulent quantities to classify the type of instability. First, we run a large number of GTC nonlinear gyrokinetic simulations by varying the plasma gradients shown in Table 1 and select the desired feature turbulence quantities as the training dataset and test dataset by the procedure shown in section 3. In this way, we collected a total number of 1411 data points for each feature quantity from these nonlinear gyrokinetic simulations. These data are then randomly split into a training dataset and a testing dataset with a ratio of 8:2, leading to a total 283 testing data points. Then we use the grid search algorithm [29] to determine the proper parameters for the classifier.

At first the linear kernel is used for the data training. It is found that when the penalty coefficient $C=13.5$, the SVM algorithm on the testing dataset yields an accuracy of 97.17% and on the training dataset it yields an accuracy of 97.52%. We can then obtain the following distance formula by using the linear kernel in the Scikit-Learn library,

$$
d = 7.64D_i - 6.06\chi_i + 4.05\chi_e + 0.35\tag{3}
$$

where d represents the distance to the dividing hyper-plane.

For the training cases, when *d* > 0, it is found to be the TEM-dominant case, and when *d* < 0, it is found to be the ITGdominant case, as is shown in Figure 7(a). Moreover, a large positive d value suggests a turbulence deep in the TEM region, and a large negative value of d suggests a turbulence deep in the ITG region. We can even make more implications about the turbulent properties through the distance formula. For example, in the typical ITG turbulence, the electron particle diffusivity D_e and thermal conductivity χ_e ought to be both much smaller than the ion thermal conductivity χ_i . Thus, the sign before χ_i should be negative since the negative value of *d* is defined as ITG-like in our notation. While in the TEM turbulence, the three transport coefficients D_e , χ_e and χ_i are close to each other [28]. Therefore, the positive sign before χ_e and negative sign before χ_i suggests that TEM tends to produce more electron heat transport and ITG tends to produce more ion heat transport, which is consistent with the fundamental physics, i.e., the turbulent transport are more effectively driven by resonant particles. As a matter of fact, the trapped electrons are resonant with the unstable drift waves in TEM and the ions are resonant with the unstable waves in ITG.

Finally, we switch to the radial basis function (RBF) kernel to carry out the turbulence classification. With $C = 30$, $\gamma = 0.1$, we obtain an accuracy of 98.23% for the classification. The classification report on test set is shown in Table 3, which demonstrates that the classification method developed here has a considerable degree of accuracy. The distribution of the data

points and the dividing surface are shown in Figure 7(b). In the test set with 283 cases, all ITG data points and almost all TEM data points are correctly classified, and only five TEM data points are incorrectly classified as ITG. All of the five incorrectly classified data points were located in the parameter space of the ITG-TEM transition region, which may be the cause of the classification difficulties. The nonlinear classification only gives slightly more accurate results than the linear classification, which can be seen from the shape of the dividing surface that the linear classification approximation is accurate enough in most regions.

Table 3. The turbulence classification accuracy report for SVM with RBF kernel.

Figure 7. Data distribution of the training set is shown in the (D_i, χ_i, χ_e) space with the dividing plane from (a) linear classification for the training set; and (b) RBF-kernel classification for both training set and test set.

5. Conclusion and discussion

In this paper, the machine learning technique based on SVM has been used to classify the nature of turbulence and proved for its effectiveness. We use the gyrokinetic code GTC to carry out a large number of linear and nonlinear simulations with Cyclone Base Case parameters, and classify the instability type as either ITG or TEM for each linear simulation by inspecting the role of the trapped electron in the simulations and comparing the entropy growing of ions with that of electrons. With time history of the target physical quantities, the SVM method is employed to achieve automatic data selection. The SVM method has also been employed to find an analytic formula for instability classifier with the input as the plasma gradients. We carry out a feature selection via correlation analysis and identify the turbulence quantities including χ_i , χ_e , and D_e as the feature quantities. These feature turbulence quantities have been used to find a second SVM classifier, i.e., the distance formula in Eq. (3), which can accurately classify the electrostatic drift wave instability types with accuracies as high as 97.52% and 98.23% with the linear and RBF kernel, respectively. Moreover, the obtained distance formulae in Eq. (2) & (3) can also be used to measure the depth of the turbulence in the classified region, which is consistent with the current physics understanding for the electrostatic drift wave turbulence. In another word, for some given feature turbulence quantities, which could come from the experimental diagnostics, we can judge from the distance formula in Eq. (3) whether the turbulence is in the deep ITG region, the deep TEM

region or the ITG-TEM transition region. These quantitative judgements may assist experimentalists to regulate fusion plasma turbulence in real time.

There are still some unsolved issues about our discoveries. For example, the ITG-TEM interface in the parameter space is found to be a flat plane as is shown in Fig. 6, instead of a curved plane in the parameter space. We further note that in the gyrokinetic simulation, physical quantities usually show a high fluctuation level in the turbulent phase, which could lead to a high computing cost in building a surrogate model for the turbulent transport, i.e., using nonlinear regression methods such as the neural network may not even be computationally feasible. The future work would also include extending the parameter space to cover more critical physics, and building a predictive surrogate model for turbulent transport in tokamaks. Another ongoing effort is to train a sequential model with a classifier head to predict turbulence types directly from the time sequences of multiple measurable signals.

Acknowledgement

The authors are grateful to the fruitful discussions with Professor Zhihong Lin, Drs. Jian Bao and Ge Dong. This work is supported by National MCF Energy R&D Programme of China Nos. 2019YFE03060000 and by NSFC under Grant No. 11975201.

References

- [1] F. F. Chen, 1965, Introduction to plasma physics and controlled fusion, 3rd edn. (New York:Dordrecht London) pp. 178
- [2] Rameswer Singh and P. H. Diamond 2020, Phys. Plasmas, **27** 042308
- [3] Coppi B., M. N. Rosenbluth and R. Z. Sagdeev 1967, Phys. Fluids 10 582
- [4] J. P. Freidberg and J. A. Wesson 1970, The Physics of Fluids, **13** 1009, DOI: 10.11063/1.1693002
- [5] P. C. Liewer 1985, Nucl. Fusion, **25** 543
- [6] Y. C. Lee, *et al* 1987 The Physics of Fluids, **30** 1331
- [7] J. Kates-Harbeck, A. Svyatkovskiy, W. Tang, Nature 568 526-531
- [8] Hui Li et al. 2021, Plasma Sci. Technol. 23 115102
- [9] P. Rodriguez-Fernandez, N. T. Howard and J. Candy 2022, Nucl. Fusion 62 076036
- [10] G. Dong et al 2021 Nucl. Fusion 61 126061
- [11] Vapnik V. The Nature of Sataistical Learning Theory 2nd edn (Berlin:Springer)
- [12] Duda R. O., Hart P. E. and Stork D. G. 2001 Pattern Classification 2nd edn (New York: Wiley-Interscience)
- [13] J. Vega, A. Murari, *et al* 2009, Nucl. Fusion, **49** 085023
- [14] Z. Lin, T. S. Hahm, W. W. Lee, W. M. Tang and R. B. White 1998, Science **281**(5384) 1835
- [15] H. Y. Wang, I. Holod, Z. Lin, et al. 2020, Phys. Plasmas 27 082305
- [16] P. Tham and A. K. Sen 1994, Phys. Plasmas, **1** 3577
- [17] Z. Lin, T. S. Hahm, W. W. Lee, et al. 2000, Physics of Plasmas 7 1857
- [18] Z. Lin, L. Chen 2001, Phys. Plasmas, **8** 1447
- [19] J. C. Adam, W. M. Tang and P. H. Rutherford 1976, The Physics of Fluids 19 561
- [20] A. M. Dimits, *et al* 2000, Phys. Plasmas, **7** 969
- [21] G. Rewoldt, Z. Lin, Y. Idomura 2007, Comput. Phys. Commun., **177** 775
- [22] S. E. Parker and W. W. Lee 1993, Phys. Fluids B., **5**(1) 77
- [23] Ng, Andrew Y. 2004, Twenty-First International Conference on Machine Learning, ICML 2004
- [24] J. Citrin et al Nucl. Fusion 55 092001
- [25] D. R. Ernst, P. T. Bonoli, P. J. Catto 2004, Physics of plasmas 11,2637
- [26] F. Ryter et al 2001 Plasma Phys. Control. Fusion 43 A323
- [27] B. S. Schneider et al 2019 Plasma Phys. Control. Fusion 61 054004
- [28] P. M. Lerman 1980, Appl. Statist. 29(1) 77-84
- [29] G.Rewoldt and W. M. Tang 1990, Phys. Fluids B: Plasma Physics 2,318